

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции
«Диалог» (2023)

Выпуск 22
Дополнительный том

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2023)

Issue 22
Supplementary volume

Редакционная коллегия: *В. П. Селегей (главный редактор), В. И. Беликов, И. М. Богуславский, Б. В. Добров, Д. О. Добровольский, Л. Л. Иомдин, И. М. Кобозева, Е. Б. Козеренко, М. А. Кронгауз, Н. В. Лукашевич, Д. Маккарти, П. Наков, Й. Нивре, В. Раскин, Э. Хови, Т. О. Шаврина, С. А. Шаров, Т. Е. Янко*

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 22, дополнительный том. 2023. С. I–1162.

Сборник включает 17 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2023», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

22-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 29-й международной онлайн конференции «Диалог». В 2023 году для публикации в дополнительном томе сборника редколлегией были отобраны 17 докладов, поданных на конференцию.

Работы, представленные в сборнике, отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на Диалоге:

- **Интеллектуальный анализ документов (Intelligent Document Processing):** классификация, Name Entity & Relation Extraction, саммаризация, генерация, анализ тональности, Argumentation Mining, Propaganda & Fake News Detection, etc., мультимодальные подходы (совместное использование моделей NLP и Computer Vision);
- **Глубокое обучение в компьютерной лингвистике:** методики применения нейронных сетей в исследованиях, содержательная интерпретация;
- **Компьютерные лингвистические ресурсы:** новые датасеты и новые сценарии и типы разметки, Evaluation Benchmarks;
- Компьютерный анализ Social Media;
- **Корпусная лингвистика и корпусометрия:** методики создания, использования и оценки корпусов;
- **Компьютерная семантика:** аналитические и дистрибуционные модели, связь между ними;
- Лингвистические онтологии и автоматическое извлечение знаний;
- **Мультимодальная коммуникация:** аналитические и нейронные модели речевого акта;
- Модели общения и диалоговые агенты;
- Лингвистический анализ текста: морфология, синтаксис, семантика (модели анализа);
- Компьютерная лексикография;
- **Полевая компьютерная лингвистика:** применение методов NLP для малоресурсных языков.

В соответствии с традициями «Диалога», конференции по компьютерной лингвистике с почти полувековой историей, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом. Диалог является де-факто крупнейшим форумом по проблемам создания современных компьютерных ресурсов, моделей и технологий для русского языка, поэтому ключевым событием «Диалога» является подведение итогов технологических соревнований между разработчиками систем лингвистического анализа русскоязычных текстов — *Dialogue Evaluation*. В этом году состоялись 4 соревнования:

- **RuCoCo:** Соревнование по разрешению кореференции;
- **RuSentNE:** Соревнование по анализу тональности к именованным сущностям в новостных текстах;
- **RECEIPT-AVQA:** Соревнование по генерации ответов на вопросы к изображениям;
- **SEMarkup:** Соревнование по автоматической семантической разметке.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике подаются на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов;
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов, они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что с 2018 года Редсовет отказался от печати сборника на бумаге. Все сборники размещаются на сайте конференции. Дополнительный том индексируется РИНЦ.

Программный комитет конференции «Диалог»
Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Рецензенты

Азарова Ирина Владимировна
Андрианов Андрей Иванович
Антонова Александра Александровна
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бурцев Михаил Сергеевич
Васильев Виталий Геннадьевич
Гусев Илья Олегович
Добров Борис Викторович
Добровольский Владимир Андреевич
Добровольский Дмитрий Олегович
Жарков Андрей Александрович
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Золотухин Денис Денисович
Иванов Владимир Владимирович
Ивойлова Александра Михайловна
Ильвовский Дмитрий Алексеевич
Инденбом Евгений Михайлович
Инькова Ольга Юрьевна
Иомдин Леонид Лейбович
Киосе Мария Ивановна
Клышинский Эдуард Станиславович
Клячко Елена Леонидовна
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович

Куратов Юрий Михайлович
Кутузов Андрей Борисович
Лапошина Антонина Николаевна
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Логинов Василий Васильевич
Лукашевич Наталья Валентиновна
Малафеев Алексей Юрьевич
Митрофанова Ольга Александровна
Мичурина Мария Александровна
Недолужко Анна
Никишина Ирина Юрьевна
Орлов Евгений Анатольевич
Пазельская Анна Германовна
Переверзева Светлана Игоревна
Петрова Мария Владимировна
Подлеская Вера Исааковна
Рыгаев Иван Петрович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Смирнов Иван Валентинович
Смулов Иван Михайлович
Татевосов Сергей Георгиевич
Урысон Елена Владимировна
Федорова Ольга Викторовна
Феногенова Алена Сергеевна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шамардина Татьяна Вячеславовна
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Contents¹

Белова Д. Д. Синтаксис и просодия расщепленного скрэмблинга: опыт экспериментального исследования	1001
Biagetti E., Villa L. B., Zanchi C., Luraghi S. Enhancing the semantic and conceptual description of Ancient Greek verbs in WordNet with VerbNet and FrameNet: a treebank-based study	1009
Bianco A., Combei C. R., Zanchi C. Painting the Senate #Green: A Corpus Study of Twitter Sentiment Towards the Italian Environmentalist Blitz	1021
Bolshakov V., Kolobov R., Borisov E., Mikhaylovskiy N., Mukhtarova G. Scaled Down Lean BERT-like Language Models for Anaphora Resolution and Beyond	1032
Буденная Е., Бажуков М., Баркова Л., Харламова Д., Дугричилов А., Резникова Т., Яковлева А., Литвинцева К., Андреева А. Диахроникон: новый ресурс для изучения русских конструкций в микродиахронической перспективе	1041
Buyanov I., Yaskova D., Sochenkov I. Who is answering to whom? Modeling reply-to relationships in Russian asynchronous chats	1052
Чага А. В. Бинарный классификатор как инструмент для поиска предложений, содержащих конструкции микросинтаксиса	1061
Евдокимова А. А. Корпус акцентуированных византийских письменных памятников и методы его разметки	1071
Margolina A. V., Kolmogorova A. V. Exploring Evaluation Techniques in Controlled Text Generation: A Comparative Study of Semantics and Sentiment in ruGPT3large-Generated and Human-Written Movie Reviews	1082
Муравьев Н., Гордеев Н., Макачук И., Кукушкина М., Бузанов А. База данных «Типологический конструкторикон»	1091
Petukhova K., Smilga V., Zharikova D. Abstract User Goals in Open-Domain Dialog Systems	1097
Russo M. Representation of lexical polysemy in the database (semantic shift 'sun/day')	1108
Turganbay R., Surkov V., Evseev D., Drobyshevskiy M. Generative Question Answering Systems over Knowledge Graphs and Text	1112
Убалехт И. Разработка морфологического анализатора для сибирского ингерманландского идиома	1127

* The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

Voloshina E., Leonova P. The Universal Database for Lexical Typology	1133
Бердонос В. Д., Животова А. А. Стратегия предредактирования исходного текста на основании автоматической оценки сложности задачи перевода для повышения качества машинного перевода узкоспециальных текстов на английский язык	1141
Zykova V. I., Klyshinsky E. S. Remus, Lupin and Moony Walk in a Bar... Grouping of Proper Names Related to the Same Denotation in Large Literary Texts Collections	1150
Abstracts	1158
Авторский указатель	1161
Author Index	1161

Syntax and Prosody of Split Scrambling: an Experimental Approach

Daria Belova

Lomonosov Moscow State University /
Moscow, Russia
dd.belova@yandex.ru

Abstract

The work is devoted to the experimental study of structures with split scrambling. With the help of acceptability assessment methods with the use of the Likert scale, self-paced reading, and prosodic experiment, the possibility of separating the left element (determinator or possessor) from the noun head within DP or PP is compared with the separation of the nominal head from its complement (dependent infinitive or prepositional phrase). The results show that separating the head for Russian speakers is not only possible but also rated higher than separating the left element from the head. This pattern is explained by the requirements of the information structure: the left element that has been fronted cannot form the only topic of the clause. The low scores are consistent with existing experimental studies; however, the results of the reading time appear to be inconsistent with existing views about the cognitive load required to process split sentences.

Keywords: constituent splitting; split scrambling; information structure

DOI: 10.28995/2075-7182-2023-22-1001-1008

Синтаксис и просодия расщепленного скрэмблинга: опыт экспериментального исследования

Белова Д. Д.

МГУ им. М.В. Ломоносова /
Москва, Россия
dd.belova@yandex.ru

Аннотация

Работа посвящена экспериментальному исследованию конструкций с расщепленным скрэмблингом. С помощью методик оценки приемлемости по шкале Ликерта, чтения с саморегулирующей скоростью и просодического эксперимента сравнивается возможность отделения левого элемента (детерминатора или посессора) от именной вершины в составе DP или PP с отделением именной вершины от комплемента (зависимого инфинитива или предложной группы). Результаты показывают, что отделение вершины для носителей русского языка не просто возможно, но и оценивается выше, чем отделение левого элемента от вершины. Данный паттерн объясняется требованиями информационной структуры: вынесенный левый элемент не может формировать собой единственный топик клаузы. Низкие оценки согласуются с существующими экспериментальными исследованиями, однако результаты времени чтения оказываются противоречащими существующим представлениям о когнитивной нагрузке, требуемой для обработки предложений с расщеплением.

Ключевые слова: расщепление составляющей; расщепленный скрэмблинг; информационная структура

1 Введение

Расщепленный скрэмблинг, то есть такие конфигурации, при которых некоторая составляющая оказывается разделена другими элементами предложения, описаны для целого ряда генетически и типологически разнородных языков (см., среди прочих, [5] для украинского, [4] для немецкого, японского и хорватского, [6] для грузинского). Например, в предложении (1) из сербского языка прилагательное *novi* оказывается вынесено в начало клаузы от именной вершины *auto*.

- (1) Novi je on auto slupao.
 новый AUX он машина разбил
 ‘Он разбил новую машину’. (Сербский, [1])

Подобные конструкции в русском языке также имеют долгую историю изучения [7][8][9]. Первым экспериментальным исследованием конструкций с расщепленным скрэмблингом в русском языке является работа И. Секериной [17]. В одном из описанных в работе экспериментов с помощью чтения с саморегуляцией скорости сравнивались предложения типа (2). Сравнение времени чтения глагола (второго элемента) и имени либо адverbиала (последнего элемента) показало, что оба элемента при расщеплении читаются респондентами дольше, чем при скрэмблинге полной именной или предложной группы, из чего автор заключает, что расщепленный скрэмблинг требует больше когнитивных ресурсов для обработки, чем полный.

- (2) а. О красивой вспоминал мой двоюродный брат студентке.
 б. О студентке вспоминал мой двоюродный брат постоянно.

При этом И. Секерина отмечает, что ее многие ее респонденты расценивали предъявляемые стимулы как «неестественные» или даже «невозможные». Тем не менее, конфигурации с расщепленным скрэмблингом распространены в разговорной речи, о чем говорит исследование А. Перельца на корпусном материале [14].

В подавляющем большинстве исследований — как русского языка, так и других — рассматриваются выносы элементов левой периферии именной группы, в частности адъективных адъюнктов. Однако гипотетически можно представить такую конфигурацию, при которой вершина группы оказывается отделена от ее правого зависимого — комплемента или адъюнкта. Мы находим такие примеры в неформальной квази-разговорной речи в социальных сетях (3), однако по спорадическим примерам нельзя оценивать общий уровень приемлемости.

- (3) Мощное желание испытываю попросить людей из Беларуси перестать писать, что в Казахстане происходит Беларусь. (Twitter, 05.02.2022)

Сравнить вынесение премодификатора (т.н. ранний сплит) с вынесением вершины от комплемента (т.н. поздний сплит) интересно нам еще и с точки зрения актуального членения предложения, содержащего расщепленную конструкцию: при позднем сплите левая часть расщепленной составляющей может сама по себе формировать (информационный) топик, что, в условиях предъявления вне контекста, как у И. Секериной, может облегчать респондентам обработку стимулов. В том, чтобы мы могли понимать интерпретацию, которая возникает у респондентов, может просодический эксперимент.

Таким образом, результаты предыдущих экспериментальных и корпусных исследований расщепленного скрэмблинга в русском языке могут быть уточнены и расширены при помощи двух методик, которые ранее к данным конфигурациям не применялись: оценок приемлемости и наиболее естественной просодии произнесения.

2 Синтаксические эксперименты

2.1 Дизайн и процедура

Среди факторов, потенциально влияющих на приемлемость конструкций с расщеплением, мы выделили следующие. Во-первых, на левой периферии именных групп могут находиться разные по статусу и семантике элементы, в частности, посессоры и детерминаторы. Во-вторых, расщеплению могут подвергаться как именные, так и предложные группы. Третьим фактором, который обсуждался выше, является позиция сплита: отщепление левого элемента от вершины (ранний сплит) или вершины от правого зависимого (поздний сплит).

Экспериментальное изучение расщепленного скрэмблинга является одной из частей более крупного исследования, посвященного расщеплению составляющих при разных типах передвижения. В связи с этим именные и предложные группы были разделены на два эксперимента.

В каждом из экспериментов скрэмблинг выступает одним из уровней фактора передвижения. Двумя другими факторами являются тип левого элемента (посессор или детерминатор; в эксперименте с именными группами также было добавлено условие без левого элемента вообще, где отщеплению подвергается только вершина) и позиция сплита. Факторный дизайн DP-эксперимента, таким образом, состоит из пяти релевантных для скрэмблинга условий, а дизайн PP-эксперимента — из четырех.

В качестве правых зависимых расщепляемой группы мы выбрали предложную группу и инфинитив; в экспериментах они были сбалансированы, то есть половина стимулов имела один тип комплемента, половина — другой. В примерах (4) и (5) приведены стимулы экспериментов с именными и предложными группами соответственно. В двух экспериментах использовались одинаковые филлеры; в качестве грамматичных использовались предложения с релятивными зависимыми клаузами и частными вопросами, неграмматичные содержали ошибки в падежных формах относительного местоимения, глаголов матричной клаузы и относительного местоимения. Примеры филлеров приведены в (6).

- (4) a. Твой Соня оценила соус из белых грибов.
 b. Твой соус Соня оценила из белых грибов.
 c. Этот Соня оценила соус из белых грибов.
 d. Этот соус Соня оценила из белых грибов.
 e. Соус Соня оценила из белых грибов.
- (5) a. Об Анином Соня вспомнила совете варить мясо на кости.
 b. Об Анином совете Соня вспомнила варить мясо на кости.
 c. Об одном Соня вспомнила совете варить мясо на кости.
 d. Об одном совете Соня вспомнила варить мясо на кости.
- (6) a. Торт, который Вера заказала на свадьбу, оказался ужасном невкусным.
 b. Какую справку нужно принести, чтобы оформить налоговый вычет?
 c. *Жиrafa, который дети видела в зоопарке, перевезли в соседний город.

Для измерения времени чтения стимулы предъявлялись респондентам по фреймам. В стимулах PP-эксперимента предлог объединялся в один фрейм в последующим словом; кроме того, предложные правые зависимые предъявлялись в виде одного фрейма, который объединял в себе предлог, прилагательное и имя.

Эксперименты были реализованы на платформе PСIbex Farm [20] и распространены в социальных сетях и на краудсорсинговом ресурсе «Яндекс.Толока»¹. Перед началом эксперимента респондент заполнял короткую социолингвистическую анкету, после которой ему предъявлялась инструкция с описанием процедуры и три грамматичных и неграмматичных тренировочных предложения. Стимулы предъявлялись в порядке «грамматичный филлер — тестовый стимул — неграмматичный филлер — тестовый стимул».

2.2 Результаты: оценки

Статистическая обработка результатов проводилась с помощью языка R [15]. В первую очередь оценки по шкале Ликерта приводились к нормальной форме для нивелирования индивидуальных тенденций каждого респондента в использовании шкалы [16]. Следующим шагом в обработке был отсев аутлаеров: из рассмотрения убирались ответы информантов, чьи оценки грамматичных и неграмматичных филлеров отличались от средних по выборке больше, чем на два стандартных отклонения.

После отсева аутлаеров для анализа DP-экспериментов остались данные 94 респондентов (возраст от 15 до 66 лет, средний возраст составил 38.44 лет); девять респондентов указали наличие лингвистического образования. Для PP-эксперимента рассматривались данные 97 респондентов (возраст от 18 до 76, средний возраст — 39 лет).

¹<https://toloka.yandex.ru>

Нормализованные оценки обоих экспериментов представлены на графике взаимодействия 1 ниже:

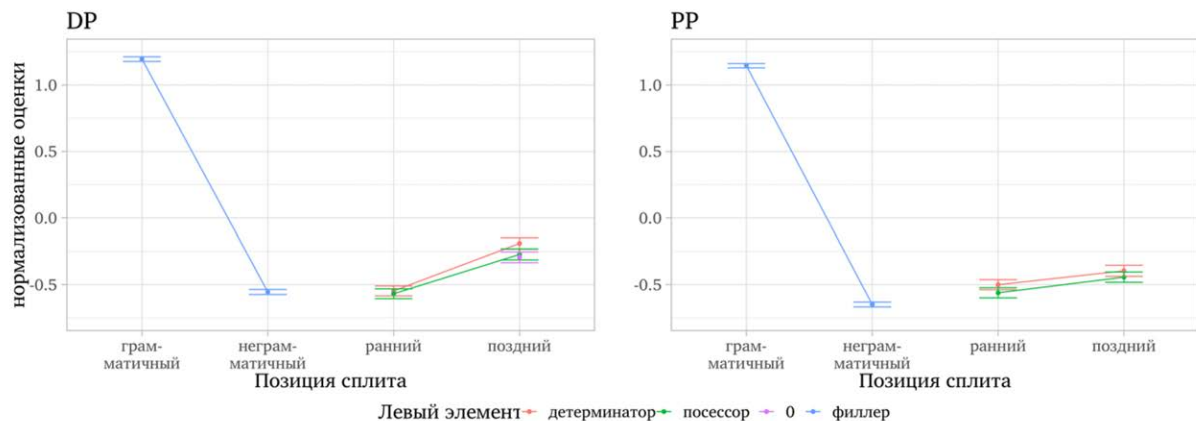


Рисунок 1: Нормализованные оценки экспериментов с именными и предложными группами

Для выявления статистически значимого влияния факторов использовались линейные смешанные модели. В качестве фиксированных эффектов проверялись независимые переменные: тип передвижения (movement), тип левого элемента (left) и позиция сплита (split_position). В качестве случайных эффектов были добавлены респондент (participant), предложение (sentence) и тип компонента (complement). Отбор формул модели проводился вручную на основании информационного критерия Акаике и Байесовского информационного критерия. Финальная формула для DP-эксперимента приведена в (7). В рамках данной модели влияние позднего сплита оказывается статистически значимым и положительным ($\beta = 0.37$, $SE = 0.05$, $p < 0.001$), при этом между двумя типами передвижений значимых различий нет ($\beta = -0.07$, $SE = 0.05$, $p = 0.19$) и взаимодействие между факторами позиции сплита и типа передвижения не значимо ($\beta = -0.06$, $SE = 0.05$, $p = 0.4$).

$$(7) \quad z\text{-scores} \sim 1 + \text{split_position} * \text{movement} + (1 | \text{sentence}) + (1 | \text{participant})$$

Общая модель для PP-эксперимента приведена в (8). При этом для скрэмблинга взаимодействие с позицией сплита положительно, но не значимо ($\beta = 0.08$, $SE = 0.08$, $p = 0.317$).

$$(8) \quad z\text{-scores} \sim 1 + \text{movement} * \text{split_position} + (1 | \text{sentence}) + (1 | \text{construction}) + (1 + \text{movement} + \text{split_position} | \text{participant})$$

Таким образом, ни для именных, ни для предложных групп тип левого элемента не оказал значимого влияния на возможность расщепления: этот параметр не вошел в финальные модели. Для именных групп, где мы также проверяли приемлемость отщепления вершины без левого элемента вообще, эти условия по уровню оценок не отличаются от условий с наличием левого элемента. Поздний сплит получает более высокие оценки, чем ранний, однако для именных групп это различие значимо, а для предложных — нет. Данная закономерность согласуется со свойствами расщепления двух типов групп при вопросительном передвижении, описанными в работе [2], где предложные группы стабильно оцениваются ниже, чем именные: учитывая, что в наших экспериментах стимулы с ранним сплитом оцениваются практически на уровне неграмматичных филлеров, даже более приемлемое условие для PP оказывается по уровню ниже, чем DP.

2.3 Результаты: время чтения

Относительно времени чтения для каждого эксперимента нас интересуют две точки сравнения, находящиеся на местах «разрыва» расщепленной конструкции: субъект — имя собственное, стоящий сразу после левой отщепленной части, и правое зависимое. Ниже представлены обобщенные схемы стимулов с ранним и поздним сплитом (9). Мы предполагаем, что различия во времени

чтения будет демонстрировать именно вторая точка: правое зависимое предположительно будет требовать больше времени на обработку при позднем сплите по сравнению с ранним, поскольку при позднем сплите читающий его не ожидает, что влечет необходимость в перестройке синтаксической структуры.

- (9) а. Ранний сплит: (P) Left Subj V Noun Comp
 б. Поздний сплит: (P) Left Noun Subj V Comp

Для каждой из двух точек создавалась своя линейная смешанная модель с тем же набором фиксированных и случайных эффектов, что и в моделях для нормализованных оценок. В связи с тем, что стимулы с инфинитивными и предложными правыми зависимыми имели разное количество фреймов и разный объем фрейма, представляющего комплемент, мы будем рассматривать их по отдельности. Содержательные результаты моделей суммированы в таблице:

Эксперимент	Субъект	Комплемент
DP: инфинитивная конструкция	нет различий	ниже при позднем сплите
DP: предложная конструкция	выше при позднем сплите	ниже при позднем сплите
PP: инфинитивная конструкция	выше при позднем сплите	ниже при позднем сплите
PP: предложная конструкция	нет различий	ниже при позднем сплите

Таблица 1. Среднее время чтения

Субъект требует больше либо столько же времени для обработки при позднем сплите по сравнению с ранним. По всей вероятности, это связано с тем, что при позднем сплите сочетание левого элемента с вершиной может интерпретироваться как вынесенный объект. Среди порядков слов, при которых объект располагается на первом месте, наиболее распространенным является OVS, а не OSV [12], поэтому после чтения вынесенного объекта респондент может ожидать глагол; кроме того, вынесение объекта в начало клаузы в любом случае предполагает специфическую информационную структуру, отличную от той, какую респондент может ожидать от стимула вне контекста. При раннем сплите такого конфликта между немаркированным ожидаемым и маркированным реальным порядками слов не возникает, так как вынесенный левый элемент не может быть полноценно встроен в синтаксическую структуру до момента появления вершины.

Комплемент, напротив, при позднем сплите стабильно требует для обработки меньше времени, чем при раннем. Эти результаты противоречат, во-первых, исследованию И. Секериной, а во-вторых, нашей гипотезе, но коррелируют с оценками приемлемости: меньшее время чтения сигнализирует о меньшей когнитивной нагрузке, что может приводить к более высоким оценкам. Однако в таком случае мы могли бы ожидать взаимодействия между временем чтения и другими факторами экспериментов, влияющими на оценки: типом группы и типом передвижения. Такого взаимодействия нет, поэтому мы можем заключить, что различия в приемлемости имеют не когнитивную, а грамматическую природу.

2.4 Синтаксические эксперименты: обсуждение

Таким образом, отщепление левого элемента от именной вершины оценивается носителями ниже, чем отщепление вершины от правого зависимого. Данные результаты кажутся неожиданными: как уже неоднократно отмечалось выше, существование в русском языке конструкций с вынесенным премодификатором подтверждено многочисленными корпусными примерами и не ставится под сомнение. Что, в таком случае, обуславливает настолько низкую приемлемость стимулов в условиях эксперимента и асимметрию между двумя типами расщеплений? Мы предполагаем, что здесь играет роль информационная структура.

Скрэмблинг как коммуникативно мотивированное передвижение может являться либо топикализацией, либо фокализацией выдвигаемой составляющей. В случае позднего сплита, существует три возможных интерпретации левой части расщепления: как информационный топик (10), как контрастный топик, требующий контрастного фокуса, например, предикатного (11), или как (контрастный) фокус (12).

- (10) {Что случилось с Петинными документами из архива?}
Петинны документа Денис испортил из архива.
- (11) Петинны документа Денис испортил из архива. {а Васины фотографии продал.}
- (12) Петинны документа Денис испортил из архива. {а не Васины фотографии.}

Для раннего сплита, однако, возможности ограничены контрастными интерпретациями. Отделенный левый элемент может быть (контрастным) фокусом либо контрастным топиком (см. также [19]). По уровню оценок, тем не менее, очевидно, что две эти интерпретации не являются для респондентов первыми приходящими на ум. Такая конфигурация, при которой отделенный премодификатор формирует неконтрастный топик, невозможна. Соответственно, мы предполагаем, что наиболее естественная для носителей русского языка стратегия интерпретации — это расценивать начало стимула как информационный топик, что приводит к неприемлемости предложений с ранним сплитом. Проверить эту гипотезу в рамках синтаксического эксперимента без контекста нельзя, поэтому мы провели просодический эксперимент.

3 Просодический эксперимент

3.1 Дизайн и результаты

Специфическая информационная структура и связанная с этим просодия конструкций с расщеплением в языках мира уже была предметом изучения [13][17]. Самым большим из является типологическое исследование К. Фэри и Дж. Фэнслоу [5] — однако русский язык в ней рассматривается мало, кроме того, не берется в расчет поздний сплит. Более подробно материал русского языка рассматривается в работах [9][10][14], авторы которых приходят к практически идентичным выводам. Расщепленный скрэмблинг и с топикализацией, и с фокализацией левого элемента демонстрирует такие же просодические свойства, как и скрэмблинг всей группы целиком. Контрастный фокус характеризуется ИК-2², то есть падением тона на контрастном фокальном элементе (цит. по [14]):

- (13) а. Малинового\ варенья она мне прислала! {а не клубничного}³
 б. Малинового\ она мне прислала варенья! {а не клубничного}

Контрастный топик, в свою очередь, произносится с ИК-5 и имеет два пика, между которыми наблюдается (относительно) высокое плато: подъем тона на контрастном топикальном элементе и падение на следующем за ним контрастном фокальном.

- (14) а. Малинового/ варенья она мне\ прислала! {а клубничного Саше}
 б. Малинового/ она мне\ прислала варенья! {а клубничного Саше}

Таким образом, мы можем сформулировать ожидания от просодического эксперимента: существует две характерные стратегии интонационного оформления расщепленных конструкций, при которых получаются интерпретируемые высказывания. Если респонденты будут предпочитать этим стратегиям немаркированный контур ИК-1 с тематическим подъемом на начале клаузы и падением в конце, значит, они воспринимают стимулы без контрастной интерпретации.

В качестве факторного дизайна мы использовали модифицированный дизайн DP-эксперимента, в котором при добавлении типа группы в качестве фактора получилось восемь условий. Для каждого из условий было отобрано по два стимула (один с предложным и один с инфинитивным компонентом) таких, чтобы каждая лексикализация встретилась в материале только один раз. В качестве филлеров было составлено 14 утвердительных предложений, максимально разнородных по структуре (16). Стимулы предъявлялись в порядке «тестовый стимул — филлер» с рандомизацией в каждой группе.

² В стандартной нотации [3]

³ В статье мы придерживаемся следующей просодической нотации: / — акцент-подъем (топик), \ — акцент-падение (информационный фокус, ИК-1), \\ — контрастное падение (контрастный фокус, ИК-2).

В эксперименте приняли участие 10 респондентов из Москвы и Санкт-Петербурга (возраст от 18 до 24, средний = 21.7), носители литературного произношения.

Результаты оказываются идиолектно устойчивыми: на нескольких первых тестовых стимулах респонденты настраиваются, после чего придерживаются стабильной просодической стратегии, на которую не оказывают влияния экспериментальные факторы. Контрастно-топикальная стратегия, оказалась регулярной только для одного респондента из десяти, причем она не зависит от позиции сплита (17a). Для двух респондентов частотной (но не единственной) оказывается фокусная стратегия с акцентом-падением на левом элементе (17b); у одного из них фокусное выделение также сопровождается эмфазой — удлинением ударного гласного и увеличением интервала падающего тона. Для оставшихся семи респондентов при раннем сплите самой частотной стратегией является контур с двумя подъемами: на левом элементе и на вершине (17c); второй может сопровождаться относительно несильным адаптивным падением на предыдущем слове — глаголе.

- (15) а. В Гошином/ Яна разобралась\| эксперименте для кандидатской диссертации\.
 б. В Гошином\| Яна разобралась эксперименте/ для кандидатской диссертации\.
 с. В Гошином/ Яна разобралась(\) эксперименте/ для кандидатской диссертации\.

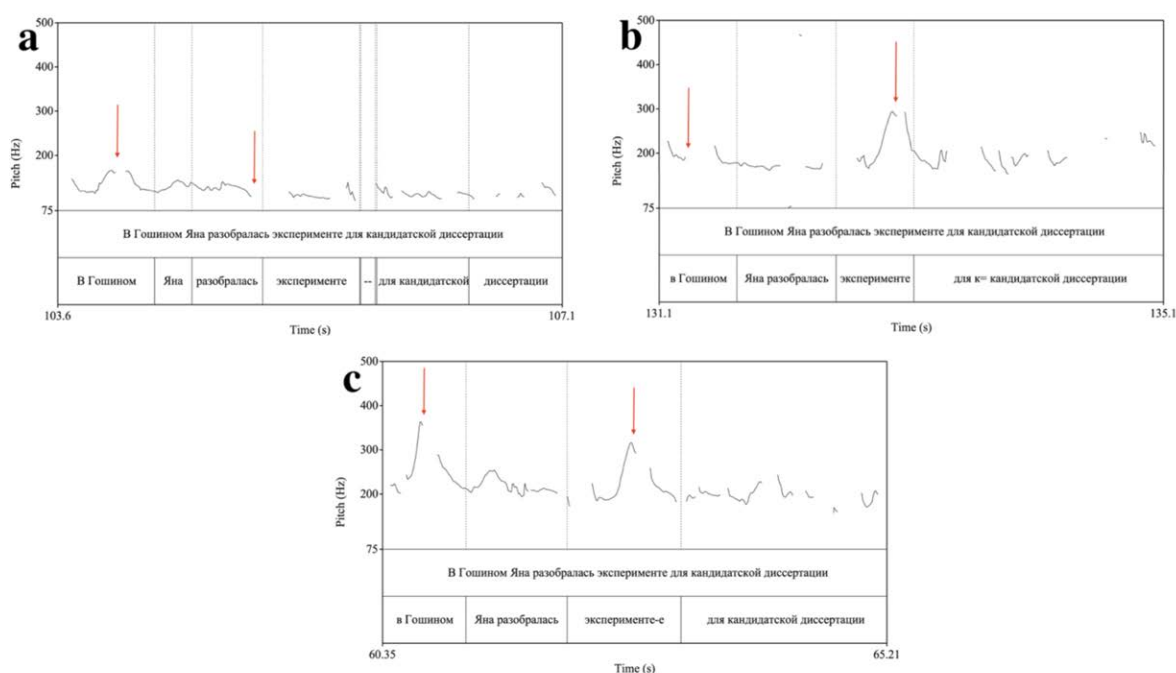


Рисунок 2. Интонационные стратегии раннего сплита

При позднем сплите характерной стратегией для последней группы из семи респондентов является подъем на вершине:

- (16) Эту полку/ Дима починил/ для фарфоровых фигурок\.

3.2 Просодический эксперимент: обсуждение

Результаты просодического эксперимента подтверждают нашу гипотезу относительно синтаксических экспериментов. Стимулы с ранним сплитом получили самые низкие оценки чуть выше или на уровне неграмматичных филлеров, несмотря на то, что существует две потенциальные коммуникативные структуры, которые позволяют их интерпретировать: (контрастный) фокус на отщепленном левом элементе либо контрастный топик на левом элементе и контрастный фокус на некотором другом компоненте предложения. Просодический эксперимент показывает, что, несмотря на возможность этих интерпретаций, большинство респондентов (семь из десяти) к ним не прибегает. Наиболее естественным для них оказывается начать произносить стимул с типичной просодией утвердительного предложения, то есть с подъемом в начале, что приводит к неинтерпретируемости высказывания.

4 Заключение

Комплекс из трех экспериментальных методик позволяет нам сделать следующие выводы о свойствах конфигураций с расщепленным скрэмлигом. Во-первых, отщепление вершины от компонента оказывается более приемлемым, чем отщепление элемента левой периферии (детерминатора или посессора) от вершины. Во-вторых, эта закономерность подтверждается данными по времени чтения фреймов в стимулах с двумя позициями сплита. В-третьих, паттерн приемлемости в каждом случае обусловлен особенностями информационной структуры клауз, что подтверждается просодическими данными: несмотря на две возможных контрастных интерпретации, участники синтаксического эксперимента тяготеют к тому, чтобы расценивать первый элемент предложения как топик, что в случае раннего сплита приводит к невозможности построения дальнейшего предложения.

Acknowledgements

Исследование выполнено за счет гранта Российского научного фонда № 22-18-00037, реализуемого в МГУ имени М.В. Ломоносова, <https://rscf.ru/project/22-18-00037/>.

References

- [1] Bašić Monika 2004. Nominal subextraction and the Structure of NPs in Serbian and English. — Universitetet i Tromsø, 2004. PhD dissertation.
- [2] Belova Daria. Discontinuous Spellout and Distributed Deletion in Russian: An Experimental Approach. // Proceedings of the 30th Annual Conference of Formal Approaches to Slavic Linguistics. — In press.
- [3] Bryzgunova E.A., Intonation [Intonacija], Russian Grammar [Russkaja grammatika], Vol. 1, pp. 96–120.
- [4] Fanselow Gisbert, Ćavar Damir. Distributed Deletion. // Theoretical Approaches to Universals. / A.Alexiadou (ed.). — Amsterdam: John Benjamins, 2002. — P. 65–107.
- [5] Féry Caroline, Paslawska Alla, Fanselow Gisbert. Nominal Split Constructions in Ukrainian. // Journal of Slavic Linguistics. — 2007. — Vol. 15, No. 1. — P. 3–48.
- [6] Fuchs Zuzanna. Movement vs. Base-generation in Georgian split DPs. — 91st annual meeting of the LSA in Austin, TX. [Talk handout]. — 2017.
- [7] van Gelderen Véronique. Scrambling Unscrambled. — Utrecht, Netherlands: LOT, 2003.
- [8] Goncharov Julie. P-doubling in split scrambling: A renaissance analysis. — Proceedings of the 2012 Annual Conference of the Canadian Linguistic Association. — 2012. — URL: homes.chass.utoronto.ca/~cla-acl/actes2012/Goncharov20 (retrieved: 10.05.2023).
- [9] Goncharov Julie 2015. P-doubling in split PPs and information structure. // Linguistic Inquiry. — 2015. — Vol. 46. No. 4. — P. 731–742.
- [10] Junghanns Uwe, and Zybatow Gerhild. Syntax and Information Structure of Russian Clauses. // The Fourth Annual Workshop on Formal Approaches to Slavic Linguistics, the Cornell Meeting / Browne W. et al. (eds.) — Ann Arbor: Michigan Slavic Publications, 1995. — P. 289–319.
- [11] Kallestinova E.D. Aspects of word order in Russian. — University of Iowa, 2007. PhD dissertation.
- [12] Mathieu Eric. Discontinuity and discourse structure: Stranded nominals as asserted background topics. // ZAS Papers in Linguistics. — 2004. — Vol. 35. No. 2. — P. 315–345.
- [13] Pereltsvaig Asya. Split phrases in Colloquial Russian. // Studia Linguistica. — 2008 — Vol. 62. No. 1. — P. 5–38.
- [14] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. — Vienna, 2020. — URL : <https://www.R-project.org/> (retrieved : 16.03.2023).
- [15] Schütze Carson T., Sprouse Jon. Chapter 3: Judgment Data // Research Methods in Linguistics. / Sharma D., Podesva R. (eds.). — Cambridge: Cambridge University Press, 2014. — P. 27–50.
- [16] Schultze-Berndt Eva, Simard Candide. Constraints on noun phrase discontinuity in an Australian language: The role of prosody and information structure. // Linguistics. — 2012. — Vol. 50. No. 5. — P. 1015–1058.
- [17] Sekerina Irina. The Syntax and Processing of Scrambling Constructions in Russian. — The City University of New York, 1997. PhD dissertation.
- [18] Sekerina Irina, Trueswell John C. Interactive processing of contrastive expressions by Russian children. // First Language. — 2012. — Vol. 32. No. 1–2. — P. 63–87.
- [19] Zehr Jérémy, Schwarz Florian. PennController for Internet Based Experiments (IBEX). — 2022. — URL: <https://osf.io/md832/> (retrieved: 02.03.2023).

Enhancing the semantic and conceptual description of Ancient Greek verbs in WordNet with VerbNet and FrameNet: a treebank-based study

Erica Biagetti

University of Pavia
Piazza del Lino 2, 27100 Pavia (PV), Italy
erica.biagetti@unipv.it

Chiara Zanchi

University of Pavia
Piazza del Lino 2, 27100 Pavia (PV), Italy
chiara.zanchi01@unipv.it

Luca Brigada Villa

University of Pavia/University of Bergamo
Piazza del Lino, 2, 27100 Pavia (PV), Italy
luca.brigadavilla@unibg.it

Silvia Luraghi

University of Pavia
Piazza del Lino 2, 27100 Pavia (PV), Italy
silvia.luraghi@unipv.it

Abstract

This paper presents a pilot study intended to enhance the semantic and conceptual description of Ancient Greek verbs in WordNet with information from two other resources, VerbNet and FrameNet, and to enrich a treebank of Ancient Greek texts with semantic information extracted from the three resources. We provided semantic annotation for verbs based on their morphosyntactic behavior, and performed a number of queries in order to extract occurrences from the Ancient Greek treebank that intended to match the different meanings of each verb. The manual check of the data extracted shows that, in spite of a limited number of mismatches, our queries yielded reliable results. The queries can be further refined in the future and complemented with a rule-based algorithm to map frame elements to dependency structure.

Keywords: WordNet; VerbNet; FrameNet; Ancient Greek verbs; dependency treebanks

DOI: 10.28995/2075-7182-2023-22-1009-1020

Улучшение семантического и концептуального описания древнегреческих глаголов в WordNet с помощью VerbNet и FrameNet: исследование на базе трибанка

Эрика Биаджетти

Университет Павии
Пьяцца дель Лино, 2,
27100 Павия (PV), Италия
erica.biagetti@unipv.it

Кьяра Занки

Университет Павии
Пьяцца дель Лино, 2,
27100 Павия (PV), Италия
chiara.zanchi01@unipv.it

Лука Бригада Вилья

Университет Бергамо /
Университет Павии
Пьяцца дель Лино, 2,
27100 Павия (PV), Италия
luca.brigadavilla@unibg.it

Сильвия Лураги

Университет Павии
Пьяцца дель Лино, 2,
27100 Павия (PV), Италия
silvia.luraghi@unipv.it

Аннотация

В статье предлагаются результаты предварительного исследования, целью которого является расширение семантического и концептуального описания древнегреческих глаголов в WordNet с использованием информации из FrameNet и VerbNet. Более того, настоящее исследование способствует пополнению трибанка древнегреческого языка семантической информацией из всех вышеупомянутых источников. Аннотирование глаголов было выполнено в соответствии с их морфосинтаксическим поведением; запросы, сделанные нами

в трибанке древнегреческого языка, предназначены для извлечения вхождений глаголов в соответствии с их различными значениями. Несмотря на незначительное число несоответствий результатов запросам, ручная проверка показывает, что полученные результаты являются надежными.

Ключевые слова: WordNet; VerbNet; FrameNet; древнегреческие глаголы; dependency treebanks

1 Introduction

WordNets (WN) are lexical databases that store meaning in a relational way. They comprise nodes for lemmas to which meanings are associated in the form of synsets, i.e. sets of synonymous words and phrases accompanied by brief definitions. Lemmas are connected to each other through lexical relations, whereas semantic relations establish connections among synsets.

The Princeton English WN, which serves as template for the creation of new WNs, contains limited morphosyntactic information for verb senses and semantic features of their participants [5], [6]. Attempts to enrich individual WNs with morphosyntactic and/or semantic information have been produced for several languages (e.g. [29], [27]). In fact, details on the morphosyntactic behavior of verbs and on the semantic properties of their dependents can be extracted from other resources, notably VerbNet and FrameNet (see Section 2). These resources can be jointly consulted, together with other ones, at the Unified Verb Index (<https://uvi.colorado.edu>). The reason for choosing these two resources over others was that VerbNet links syntactic and semantic patterns, which is ultimately the aim of this study, while FrameNet provides frames that, though being intended to describe English lexemes, are more easily adaptable to other languages. One effort towards integrating information from the three resources is described in Stayannova and Leseva [27]. The authors expanded the semantic and conceptual description of verbs in WordNet by combining information from FrameNet and VerbNet. They further supplemented the verbs' description with syntactic patterns extracted from the semantically annotated corpus BulSemCor ([15], [16]) in which all tokens are assigned a WordNet synset.

Building on the work by Stayannova and Leseva [27], in this paper, we present a pilot study for mapping semantic information onto Ancient Greek (AG) verbs contained in the Ancient Greek WordNet (AGWN) with morphosyntactic, semantic, and conceptual information provided by VerbNet and FrameNet. Notably, no semantically annotated corpus is available for AG. For this reason, we preliminarily enriched a selection of verbs in a morphosyntactically parsed corpus (treebank) of AG texts with semantic information from WordNet, VerbNet and FrameNet employing a semi-automatic methodology. The treebank contains various texts among which we selected two sub-corpora: the entire *Iliad* and *Odyssey* (ca. 11th century BCE) and a selection of texts from Attic orators (5th century BCE). This is a convenience sample which, though unbalanced (232.340 vs. 97.142 tokens respectively, see Table 2), shows the potential of our approach for studying verbal valency and verb meanings across different periods (Homeric vs. Classical) and genres (epic poetry vs. prose).

The paper is organized as follows. In Section 2, we introduce the language resources employed in this study. After providing the basics of Ancient Greek verbal morphosyntax and illustrating our methodology (Section 3), in Section 4 we discuss the results. Section 5 concludes the paper and presents future work.

2 Resources employed in this study

2.1 The Ancient Greek WordNet

The AGWN belongs to a family of WNs for ancient Indo-European languages, an ongoing project jointly developed at the University of Pavia, the University of Exeter, the Center for Hellenic Studies at Harvard University, the Alpheios Project, and the Catholic University of the Sacred Heart in Milan [2]. Besides Ancient Greek, it currently also comprises WNs for Sanskrit and Latin.

The Sanskrit, AG and Latin WNs were designed to be interoperable with each other and to enable cross-linguistic comparison of linguistic structures. To enhance compatibility, the developers maximize use of the synsets deriving from the Princeton English WN. As in other WNs, lemmas can be assigned multiple synsets, which indicates polysemy. In this family of WNs, however, polysemy is also framed within a cognitive linguistic approach (e.g. [17], [28]), which entails assuming that all senses of a word can be organized in a structured semantic network consisting of literal and non-literal senses.

For each synset, the three WNs provide information on periods and literary genres of attestation, and optionally *loci*, i.e. exemplifying occurrences referred to by author and work. In the AGWN, a word sense can be assigned to one of the following periods: Archaic (8th-6th c. BCE), Classical (5th c.-323 BCE), Hellenistic (323-31 BCE), Roman (31 BCE-290 CE). Labels for genres are Historiography, Liturgical, Miscellany, Novel, Oratory, Philosophy (epistles, treatise, dialogue), Poetry, Technical, Theater, Treatise. For several genres, additional subgenres can be identified; for example, Poetry can take epic, lyric, idyll, satire, epyllion or didactic as sub-label, whereas Oratory can be political, juridical or can take the form of a homily. Diachronic and stylistic metadata are meant to enable studies on semantic change over time and across literary genres and authors.

2.2 VerbNet

VerbNet (VN) [14] is a network of English verbs that links their syntactic and semantic patterns. The network is organized into verb classes extending Levin’s classes [18], with the understanding that the syntactic form of a verb and its arguments inform its semantics. Levin’s classes were then refined, and subclasses were added to achieve syntactic and semantic coherence among verb members of a class. Each verb class in VN is described by semantic roles and selectional preferences of its arguments, as well as by frames consisting of a syntactic description and a semantic representation with subevent structure based on the Dynamic Event Model [24], [25]. VN groups together verbs with identical sets of syntactic frames and semantic predicate structures. Each VN class contains a set of syntactic frames, depicting the possible surface realizations of the argument structure for constructions such as transitive, intransitive, resultative, and a large set of diathetic alternations. Semantic restrictions (such as animate, human, organization) are used to constrain the types of semantic roles taken by the arguments (Agent, Patient, Instrument, etc.). Each syntactic frame is then associated with explicit semantic information, expressed by semantic predicates such as ‘contact’, ‘manner’ or ‘cause’. Table 1 shows a complete entry for a frame in VerbNet class Hit-18.1:

Class Hit-18.1			
Semantic roles and restrictions: Agent[+int control], Patient[+concrete], Instrument[+concrete]			
Members: bang, bash, blast, bop, click, dash, knap, hit, smite ...			
Frames:			
Name	Example	Syntax	Semantics
NP V NP PP.instrument	<i>Paula hit the ball with a stick.</i>	Agent V Patient {with} Instrument	¬ CONTACT(e1 , Patient , Instrument) DO(e2 , Agent) UTILIZE(e2 , Agent , Instrument) MANNER(e2 , Directedmotion , Instrument) CONTACT(e3 , Patient , Instrument) MANNER(e3 , Forceful , Instrument) CAUSE(e2 , e3)

Table 1: Entry for a frame in VerbNet class Hit-18.1-1

2.3 FrameNet

FrameNet (FN) [26] is a lexical database of contemporary English which maps meaning to form through the theory of Frame Semantics [7]-[12]. The theory assumes that people understand the meaning of words largely by virtue of the larger conceptual structures, i.e. frames, which they evoke. FrameNet defines frames and their elements, and provides annotated sentences to show how frame elements fit syntactically around the lexical unit that evokes the frame. In the simplest case, the frame-evoking word is a verb, and the frame elements are its syntactic dependents. In (1), the Hit_target frame is evoked by the verb *hit*; the frame elements are the Agent instantiated by the pronoun *he*, the Target expressed by *the bull’s-eye*, and the Instrument instantiated by the prepositional phrase *with his first arrow*:

(1) It looks like [_{Agent} he] HIT [_{Target} the bull’s-eye] [_{Instrument} with his first arrow]!

Note that FN frame elements are frame-specific semantic roles. Consequently, they are more fine-grained than VN semantic roles. For example, the Telling frame evoked by speech verbs has Speaker, Addressee and Message as core elements, whereas VerbNet has the more generic Agent, Recipient and Theme as semantic roles of the roughly corresponding transfer_mesg-37.1.1 VN class; similarly, the Awareness frame evoked by verbs of cognition like *know*, *comprehend*, and *conceive* takes the core elements Cognizer and Content, instead of the more generic semantic roles Experiencer and Stimulus.

2.4 Dependency treebanks

Treebank	Tokens	Text	Author
aeschines-1-1-50-bu1.conllu	4353	Against Timarchus	Aeschines
aeschines-1-51-100-bu1.conllu	3879		
aeschines-1-101-150-bu1.conllu	3975		
aeschines-1-151-196-bu1.conllu	3710		
antiphon-1-bu2.conllu	2034	Against the Stepmother for Poisoning	Antiphon
antiphon-2-bu2.conllu	2908	First Tetralogy	
antiphon-5-bu2.conllu	7411	On the Murder of Herodes	
antiphon-6-bu2.conllu	3989	On the Chorus Boy	
dem-59-neaira-2019.conllu	10457	Apollodorus Against Neaera	Demosthe nes
demosthenes-18-1-50-bu2.conllu	4141	On the Crown	
demosthenes-18-51-100-bu1.conllu	4046		
demosthenes-18-101-150-bu2.conllu	4292		
demosthenes-18-151-200-bu2.conllu	4272		
demosthenes-18-201-275-bu1.conllu	5665		
demosthenes-18-276-324-bu1.conllu	3877		
demosthenes-1-bu1.conllu	2181	First Olynthiac	
demosthenes-4-phil1-bu1.conllu	3926	Philippic 1	
lysiias-1-bu1.conllu	2826	On the Murder of Eratosthenes	
lysiias-12-bu1.conllu	5614	Against Eratosthenes	
lysiias-13-bu1.conllu	5629	Against Agoratus	
lysiias-14-bu1.conllu	2786	Against Alcibiades 1	
lysiias-15.conllu	685	Against Alcibiades 2	
lysiias-19-bu1.conllu	3593	On the Property of Aristophanes	
lysiias-23-bu1.conllu	893	Against Panoleon	
tlg0012.tlg001.perseus-grc1.1-6.tb.conllu	31966	Iliad	Homer
tlg0012.tlg001.perseus-grc1.7-12.tb.conllu	30011		
tlg0012.tlg001.perseus-grc1.13-18.tb.conllu	35600		
tlg0012.tlg001.perseus-grc1.19-24.tb.conllu	31550		
tlg0012.tlg002.perseus-grc1.1-6.tb.conllu	26559	Odyssey	
tlg0012.tlg002.perseus-grc1.7-12.tb.conllu	27144		
tlg0012.tlg002.perseus-grc1.13-18.tb.conllu	26661		
tlg0012.tlg002.perseus-grc1.19-24.tb.conllu	24633		

Table 2: Treebanks used for our study

Dependency treebanks are collections of morphosyntactically annotated texts, representing sentences as trees where each word corresponds to a node. The treebanks¹ we considered for our work (Table 2) consist in the conversion made by Francesco Mambrini of a series of treebanks of AG texts to the Universal Dependencies (UD) format, which provides a standard set of annotation guidelines for part-of-speech tags, morphological features, and syntactic dependencies.²

3 Methods

The queries presented in Section 3.2 are in line with Levin’s [18] work and VN and rely on the assumption that certain morphosyntactic configurations in which specific Ancient Greek verbs occur (syntactic frames in VN terminology), which can be easily extracted from treebanks, are reliable predictors for verb senses in context. This means that, once the occurrences instantiating a syntactic frame are extracted, the annotation of the verbal nodes they contain can be easily enriched with semantic information from WN, VN and FN.

3.1 Basics of Ancient Greek verbal morphosyntax

In this paper, we take into account two aspects of verbal morphosyntax to which AG verb senses are sensitive: voice and argument structure. The verbal lemmas included in this case study (Table 3) are known to be good representatives of how voice and argument structure affect verb senses.

AG verbs distinguish three values for voice: active, middle, and passive (Act, Mid, Pass in Table 3). The middle and the passive voices are morphologically distinct only in some tenses (aorist and future), while others feature a mediopassive that instantiates both the middle and the passive diatheses (MidPass). The active/middle distinction can encode the (anti)causative alternation [13] in AG: *rhégnumi*(Act) ‘break(tr.)’ (as in *Rose broke the vase*) vs. *rhégnumai*(MidPass) ‘break(intr.)’ (as in *The vase broke*) [22] (queries 2a-d in Table 3 below). The same alternation is featured by some experiential verbs, such as *mimnḗskō*(Act) ‘remind’ vs. *mimnḗskomai*(MidPass) ‘remember’ and *kholólō*(Act) ‘anger someone’ and *kholóomai*(MidPass) ‘be/get angry’ [20] (queries 1a-c and 7a-b). Notably, not all experiential verbs behave in the same way in this respect, as shown by *lanthánō*(Act) ‘escape notice’ and *lanthánomai*(MidPass) ‘forget’ (queries 9a-b). In Homeric Greek (HG), morphologically passive verb forms sometimes do not express the passive diathesis (e.g. *The vase was broken [by Rose]*); rather, they are semantically equivalent to middles [1]. Voice distinctions often express meanings not related with either the (anti)causative alternation or with active/passive diatheses: this is the case of *phrázō*(Act) ‘show, tell’ vs. *phrázomai*(MidPass) ‘plan, devise, realize’ (queries 3a-c). In the case of *phobéō* the difference between the meanings of Pass and MidPass is not easy to tell given the verb’s semantics: *phobéō*(Act) HG ‘be set on flight’ CIG ‘be frightened’ vs. *phobéomai*(MidPass) HG ‘flee’ CIG ‘fear’ (queries 8a-d). In both corpora the Pass can occur in contexts in which an agent is overtly realized or implied but this does not always need to be the case; in addition, in HG the active can occur with passive agent PPs and function as lexical passive; see [21].

AG nouns feature five cases: nominative, genitive, dative, accusative, and vocative. Subjects of transitive and intransitive verbs are encoded in the nominative. Second arguments of two-place verbs (e.g. *hit*) most frequently (74% of two-place verbs in HG; see [22]) feature the accusative case (NomAcc construction), but can also occur in the genitive and dative cases (NomGen and NomDat constructions). The dative case is used for third arguments of three-place verbs of transfer (e.g. *give*; NomAccDat construction). Semantic distinctions brought about by voice may be paired with changes in argument structure: *phrázō*(Act) ‘show, tell’ is a knowledge transfer verb, which takes the NomAccDat construction (the dative third argument is optional); *phrázomai*(MidPass) ‘plan, devise, realize’ is a cognition verb, featuring the NomAcc construction. The behavior of *orégō/orégomai* is partially similar (queries 6a-e): *orégō*(Act) ‘give, assign’ takes the NomAccDat construction with an optional dative third argument. However, *orégomai*(MidPass) is more complex: in Homeric Greek, it features the NomAcc and the NomGen construction with changes in meaning (‘hit’ vs. ‘try to hit’; cf. [20]). In Post-Homeric Greek, only the NomGen construction remains with *orégomai*, but it displays a meaning shift, ‘desire,

¹<https://github.com/francescomambrini/katholou>.

² <https://universaldependencies.org/>.

wish’, hence moving to the experiential domain [20]. Thus, changes in argument structure can result in different senses of the same verb; this is also the case of *oída* ‘know, be skilled’, queries 4a-c, and *mnáomai* ‘have in mind, woo’, queries 5a-b. Another verb that shows a shift in argument structure is *punthánomai*, which can take both the NomAcc and the NomGen construction in HG, but is limited to the NomAcc construction in the CIG corpus we selected (queries 10a-b).

3.2 Extraction of the patterns from the treebanks

In order to infer synsets, verb classes and frames based on syntactic patterns, we chose ten AG lemmas (*mimnḗskō*, *rhḗgnumi*, *phrázō*, *oída*, *mnáomai*, *orégō*, *kholóō*, *phobéō*, *lanthánō* and *punthánomai*) and developed 30 queries (Table 3) to extract their corresponding syntactic patterns from the treebank.³ This process was facilitated by the tool UDeasy [3]. The ten verbs were chosen to exemplify a variety of different morphosyntactic behaviors (cf. Section 3.1). Regarding syntactic conditions, we imposed a criterion on all queries that required the direct object (obj) and indirect object (iobj) nodes, if present, to be directly dependent on the verb node.

We then provided semantic annotation with data from WN, VN and FN eliciting meanings from Liddel/Scott dictionary.⁴ Annotation was performed by adding the relevant synsets, verb classes and frames to the MISC field of the CoNNL-U file, the standard format followed by UD treebanks to store annotation at sentence and token level. In the case of verbs that could in principle feature three different voice values, i.e. *mimnḗskō*, *rhḗgnumi*, *phrázō*, *orégō*, *kholóō*, *phobéō*, and *lanthánō*, we predicted that Pass instantiated the passive diathesis and accordingly we used the same semantic annotation for the Pass and the Act.⁵ This prediction turned out to be correct in the case of *mimnḗskō*, but not in the case of *phrázō* and *kholóō*, which feature two and 11 Pass forms respectively with middle meaning (we corrected the annotation in Table 4 after the manual check). For *rhḗgnumi*, *orégō* and *lanthánō* we did not find any Pass forms. Concerning *phobéō*, as noted above, the verb’s semantics blurs the difference between MidPass ‘be afraid’ and Pass ‘be frightened’. Table 4 shows the semantic annotation aligning the three resources and the occurrences extracted through our queries from the two sub-corpora (HG and Classical Greek, CIG).

³ Note that the treebank contains texts in Greek script; to facilitate reading, in this article we provide transcriptions of AG words in Latin alphabet.

⁴ Henry George Liddell, Robert Scott, A Greek-English Lexicon (<https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0057>)

⁵ The verbs *mnáomai* and *oída* do not have Pass forms in AG.

id	verb lemma	node	conditions	texts	
1a	μιμνήσκω (mimnēskō)	<u>verb</u>	verb: Voice=Act	all	
1b		obj	verb: Voice=MidPass Mid		
1c		iobj	verb: Voice=Pass		
2a	ρήγνυμι (rhēgnumi)	<u>verb</u>	verb: Voice=Act	all	
		obj	obj: Case=Acc		
2b		<u>verb</u>	verb: Voice=Pass		
2c		obj	verb: Voice=MidPass Mid		
2d	iobj	verb: Voice=Act	obj: Case≠Acc		
3a	φράζω (phrázō)	<u>verb</u>	verb: Voice=Act	all	
3b		obj	verb: Voice=MidPass Mid		
3c		iobj	verb: Voice=Pass		
4a	οἶδα (oīda)	<u>verb</u>	obj: Case=Acc	all	
4b		obj	obj: Case=Gen		
4c		<u>verb</u>	no obj		
5a	μνάομαι (mnáomai)	<u>verb</u>	obj: Case=Gen	all	
5b		obj	obj: Case=Acc		
6a	ὀρέγω (orégō)	<u>verb</u>	verb: Voice=Act	all	
6b		obj	verb: Voice=MidPass Mid	hom	
6c		iobj	obj: Case=Acc	verb: Voice=MidPass Mid	
6d			obj: Case=Gen	verb: Voice=MidPass Mid	posthom
6e		<u>verb</u>	verb: Voice=Pass	all	
		obj			
		iobj			
7a	χολόω (kholóō)	<u>verb</u>	verb: Voice=Act	all	
7b		obj	verb: Voice=MidPass Mid Pass		
8a	φοβέω (phobéō)	<u>verb</u>	verb: Voice=Act	hom	
8b		obj	verb: Voice=MidPass Mid Pass		
8c		iobj	verb: Voice=Act	posthom	
8d			verb: Voice=MidPass Mid Pass		
9a	λανθάνω (lanthánō)	<u>verb</u>	verb: Voice=Act	all	
9b		obj	verb: Voice=MidPass Mid Pass		
10a	πυνθάνομαι (punthánomai)	<u>verb</u>	obj: Case=Acc Gen no obj	hom	
10b		obj	obj: Acc	posthom	
		iobj	obj: Gen		
		<u>verb</u>			
		obj			
		iobj			

Table 3: Conditions of the 30 queries
(underlined nodes are obligatory, non-underlined nodes are optional)

query	verb lemma	WordNet	VerbNet	FrameNet	HG	CIG
1a	μιμνήσκω (mimnēskō)	v#00410666 v#00412253 v#00413589 v#01609936 v#00414046 v#00491159	characterize-29.2-1-1 care-88.1	Evoking	10	
1b		v#00412961	characterize-29.2-1-1	Memory	100	30
1c		v#00412961	tell-37.2	Evoking		3
2a	ρήγνυμι (rhēgnumi)	v#00231588	break-45.1	Cause to fragment	29	
2b		v#00231588	break-45.1	Cause to fragment		
2c		v#00231062	break-45.1	Breaking apart	16	
2d		v#00231588	break-45.1	Breaking apart	4	
3a	φράζω (phrázō)	v#00639889	transfer msg-37.1.1-1-1	Telling	22	11
3b		v#01134235 v#00493867	intend-61.2-1-1 comprehend-87.2-1-1-1	Making arrangements Coming to believe	83	
3c		v#01134235 v#00493867	intend-61.2-1-1 comprehend-87.2-1-1-1	Making arrangements Coming to believe	2	
4a	οἶδα (oída)	v#00401762	comprehend-87.2-1-1-1	Awareness	151	80
4b		v#00402497	not found	Expertise	24	
4c		not found	not found	not found	119	125
5a	μνάομαι (mnáomai)	v#00491159	care-88.1	Cause emotion Awareness	3	
5b		v#01727931	compel-59.1-1	Forming relationships	14	
6a	ὀρέγω (orégō)	v#01583087	conduct-111.1	Giving	18	
6b		v#00841893	hit-18.1-1	Hit target	3	
6c		v#01153576	not found	Aiming	3	
6d		v#01245362	want-32.1	Desire		2
6e		v#01583087	conduct-111.1	Giving		
7a	χολώω (kholóō)	v#01789790	amuse-31.1	Stimulate emotion	4	
7b		v#01790925	marvel-31.3	Experiencer_focused_emotion	65	
8a	φοβέω (phobéō)	v#01111362	banish-10.2	Beat_opponent (rout) Removing (expel)	15	
8b		v#02079709	escape-51.1	Fleeing	17	
8c		v#01783082	amuse-31.1	Stimulate emotion		
8d		v#01784021	marvel-31.3	Experiencer_focused_emotion		17
9a	λανθάνω (lanthánō)	v#02078305 v#01419809	avoid-52	Elusive_goal	55	17
9b		v#00611721 v#00610654 v#00616520 v#00614532	not found	Abandonment Remembering_experience Remembering_information Remembering to do	31	
10a	πυνθάνομαι (punthánomai)	v#00899241 v#00592510 v#00600349	inquire-37.1.2 comprehend-87-2 learn-14-2-1	Questioning Coming_to_believe Hearsay	69	
10b		v#00899241	inquire-37.1.2	Questioning		38

Table 4: Semantic annotation and extracted occurrences

4 Results

After extracting the data, we manually checked them in order to see whether the meanings we selected actually matched the occurrences of each specific query: this turned out to be most frequently correct. In Table 5 we calculated Precision. Queries for which Precision is not relevant are those that turned out not to yield results after checking annotation errors. We could not calculate Recall, as the corpus we used was not semantically annotated.

Several interesting observations emerge from the data. In the first place, only two queries (9a and 10a) extracted occurrences that did not show the expected meaning, as Precision for query 2d is <1 due to an annotation error. Query 9a concerned the verb *lanthánō* ‘escape notice’. This verb instantiates the (anti)causative alternation through voice only exceptionally, the MidPass and in HG the Pass meaning ‘forget’. Hence, query 9a did not yield the expected results in a single occurrence (*Il.* 15.60), in which the active has the causative meaning ‘make forget’, which was not predicted. With query 10a we extracted forms of *punthánomai* in HG and found four occurrences which did not match the expected meaning but rather expressed the meaning ‘perceive (a sound)’.

With the verb *mimnēskō* the Act/MidPass distinction encodes the anticausative alternation. Interestingly, no Act forms occur in CIG, but Pass forms encode the passive diathesis for the meaning ‘remind’. Hence the same semantic annotation applies to the verb in both corpora for the Act (HG) and the Pass (CIG), in line with the preliminary choice we made. It needs to be remarked that the number of occurrences extracted may be higher than the actual number of occurrences of verbs: in the case of query 1a, for example, this is because *mimnēskō* (Act) is a three-place verb featuring the NomAccGen construction, but in some occurrences only the second or the third argument are overtly realized, without a change of meaning. For this reason, our query did not specify the case of the co-occurring argument(s) and how many arguments we were looking for; consequently in passages in which both arguments were overtly realized we extracted the same occurrence twice (e.g. *mēdé me toutōn mimnēsk’* ‘do not remind me (*me* Acc) of these things (*toutōn* Gen)’, *Od.* 14.168-169, yielded two results). Similarly, with *kholōō* the Act/MidPass distinction also encodes the anticausative alternation. This verb only occurs in HG, most frequently in non-Act voice meaning ‘be/become angry’. This includes Pass forms, as remarked in sec. 3.1 and 3.2 (see further below). It can occur with a dative second argument that indicates the stimulus, with a genitive indicating the reason or without a second argument, with no semantic differences, hence we performed a single query (7b). Act forms with causative meaning occur four times (7a).

For the verb *rhēgnumi* we performed two separate queries for active forms depending on whether the second argument was realized (2a) or not (2d). In this second case, we expected to find cases of P-lability, which are known from post-HG; however, all passages featured referential null objects (NO) easily recoverable from the preceding context. Due to the extent to which referential NOs occur under different syntactic and discourse conditions [19], this type of query does not seem to be suitable to automatically retrieve labile verb forms. In our sample, this verb only occurs in HG; however, according to reference works it also occurs in Attic prose writers, albeit infrequently, due to frequent occurrence of prefixed verbs such as *diarrēgnumi*, *ekrēgnumi*, *katarrēgnumi*, all featuring similar meanings. We also performed two different queries for *oīda* depending on whether the second argument was overtly realized or not (4a and 4b vs 4c). In this case, the motivation was that we wanted to see whether the verb could have some discourse function when occurring without a direct object. A survey of the occurrences shows that this is not the case, and that occurrences without an overt direct object may either feature referential NOs or subordinate clauses. In HG, occurrences of NOs may show the semantics associated to both query 4a and 4b (the latter query did not yield any result in CIG): this indicates that more queries should be added to separate occurrences of NOs from occurrences with subordinate clauses, as the latter feature the semantics associated to query 4a. This query is the only one which, in spite of the unbalance of the two sub-corpora, yielded a higher number of results for CIG. The reason might lie in the expansion of subordination from HG to CIG [4] or in differences between textual genres.

As anticipated in Sec. 3.2, we corrected the sense we had selected for Pass forms of *phrázō* and *kholōō* based on the results of the query. Indeed, as we remarked in Sec. 3.1, Pass forms frequently have middle meaning in HG. Remarkably, Pass forms of *mimnēskō* that only appear in post-HG actually encode the passive diathesis, in line with known developments of the voice system, whereby the passive diathesis is increasingly connected with Pass voice (and partly with MidPass) after HG [21]. This suggests that

the annotation should be better differentiated based on language stages and known facts about the diachrony of individual verbs.

Query	Extracted results	Results with expected meaning	Precision
1a	6	6	1
1b	99	99	1
1c	3	3	1
2a	24	24	1
2b	0	0	not relevant
2c	16	16	1
2d	5	4	0,800
3a	22	22	1
3b	83	83	1
3c	0	0	not relevant
4a	149	149	1
4b	22	22	1
4c	116	116	1
5a	3	3	1
5b	14	14	1
6a	18	18	1
6b	3	3	1
6c	3	3	1
6d	2	2	1
6e	0	0	not relevant
7a	4	4	1
7b	65	65	1
8a	15	15	1
8b	17	17	1
8c	17	17	1
9a	55	54	0,981
9b	31	30	1
10a	69	65	0,942
10b	44	44	1

Table 5: Precision index

With *mnáomai*, *orégō* and *phobéō* the queries yielded results that reflected the semantic annotation we provided in Table 4. In particular, *mnáomai* never occurs in Attic, but, outside poetry, is limited to Ionic and occurs in Herodotus (5th century BCE). It features the NomAcc construction and the meaning ‘sue for, solicit’, which results from semantic extension of the meaning ‘woo’ (query 5*b*) and is compatible with the VN class compel-59.1 that we selected for this query (it does not contain this meaning, but it contains the related meaning ‘persuade’). In the case of *orégō* it must be stressed that the queries we performed were more fine-grained than for other verbs, being based not only on morphosyntactic features but also on language stages. This is also the case for *phobéō*, a verb that underwent a semantic shift after HG from ‘flee/be put on flight’ to ‘frighten/fear’. Notably, while both Act and MidPass are equally frequent in HG (8*a*, 8*b*), in our sample of CIG only MidPass occurs (8*d*).

5 Conclusion

The manual check of the data extracted through the queries we designed based on the morphosyntactic behavior of a number of Ancient Greek verbs shows that the morphosyntactic features chosen were to a high extent reliable predictors of the associated meanings, as also shown by the Precision values. In future work some of the queries must be refined, e.g. specifying the form of subordinators for verbs likely to take complement clauses such as *oída* or extended with more fine-grained distinctions based on known developments of AG morphosyntax. Furthermore, we plan to implement the rule-based algorithm described in [23] to automatically add role-dependency links to event participants i.e. tags that indicate how a frame element is expressed syntactically with respect to the node that corresponds to the frame trigger.

Finally, we plan to connect resources by exploiting Linked Open Data principles. The Ancient Greek WordNet identifies each lemma through a dedicated URI; as a next step, we plan to apply the same URIs to the corresponding lemmas in treebank, so that each lemma is uniquely identifiable and findable.

Acknowledgements

We wish to express our gratitude toward the anonymous reviewers, who greatly contributed to improving the quality of this paper. This chapter results from joint work of the authors. For Italian academic purposes, Erica Biagetti is responsible for Section 2, Chiara Zanchi for Section 3.1, Luca Brigada Villa for Section 3.2, and Silvia Luraghi for Section 4. Sections 1 and 5 were written jointly by the authors.

References

- [1] Allan Rutger J. The middle voice in Ancient Greek: a study in polysemy. — Amsterdam: JC Gieben, 2003.
- [2] Biagetti Erica, Zanchi Chiara, Short William M. Toward the creation of WordNets for ancient Indo-European languages // Proceedings of the 11th Global Wordnet Conference. — Pretoria, South Africa, 2021. — P. 258–266.
- [3] Brigada Villa Luca. UDeasy: a Tool for Querying Treebanks in CoNLL-U Format // Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10). — Marseille, France, 2022. — P. 16–19.
- [4] Cristofaro Sonia. Aspetti sintattici e semantici delle frasi complete in greco antico. — Firenze: La Nuova Italia Ed., 1996.
- [5] Fellbaum Christiane (ed.). WordNet: An electronic lexical database. — Cambridge, MA: MIT Press, 1998.
- [6] Fellbaum Christiane. WordNet, The Encyclopedia of Applied Linguistics. — Hoboken, NJ: Wiley Online Library, 2012. <https://doi.org/10.1002/9781405198431.wbeal1285>
- [7] Fillmore Charles J. Frame semantics and the nature of language // Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech. — New York, NY, 1976. — P. 20-32.
- [8] Fillmore Charles J. Scenes-and-frames semantics, Linguistic Structures Processing. — Amsterdam: North Holland Publishing, 1977a.
- [9] Fillmore Charles J. The need for a frame semantics in linguistics, Statistical Methods in Linguistics. — Stockholm: Scriptor, 1977b.
- [10] Fillmore Charles J. Frame semantics, Linguistics in the morning calm. — Seoul: Hanshin, 1982. — P. 111-37.
- [11] Fillmore Charles J. Frames and the semantics of understanding, Quaderni di Semantica — Bologna, Italy: Il Mulino, 1985. — Vol. 6, P. 222-254.

- [12] Fillmore Charles J., Baker Collin F. *Frame Semantics for Text Understanding // Proceedings of WordNet and Other Lexical Resources Workshop*. — Pittsburgh, PA, 2001.
- [13] Haspelmath Martin. *More on the typology of inchoative/causative verb alternations, Causatives and transitivity*. — Amsterdam: John Benjamins, 1993. — P. 87–120.
- [14] Kipper-Schuler Karin. *VerbNet: A broad-coverage, comprehensive verb lexicon (Ph. D. thesis)*. — Philadelphia, PA: University of Pennsylvania, 2006.
- [15] Koeva Svetla, Leseva Svetlozara, Rizov Borislav, Tarpomanova Ekaterina, Dimitrova Tsvetana, Kukova Hristina, and Todorova Maria. 2011. *Design and development of the Bulgarian Sense-Annotated Corpus*. In *Proceedings of the Third International Corpus Linguistics Conference (CILC), 7-9 April 2011, Valencia, Spain*, pages 143–150. Universitat Politècnica de Valencia.
- [16] Koeva Svetla, Leseva Svetlozara, Tarpomanova Ekaterina, Rizov Borislav, Dimitrova Tsvetana, and Kukova Hristina. 2010. *Bulgarian Sense-Annotated Corpus – results and achievement*. In *Proceedings of the 7th International Conference of Formal Approaches to South Slavic and Balkan Languages (FASSBL-7)*, pages 41–49.
- [17] Lakoff George, Johnson Mark. *Conceptual metaphor in everyday language, The journal of Philosophy*. — New York, NY: The journal of Philosophy, 1980. — Vol. 8(77), P. 453-486.
- [18] Levin Beth. *English verb classes and alternations*. — Chicago: Chicago University Press, 1993.
- [19] Luraghi Silvia. *On the meaning of prepositions and cases*. — Amsterdam: John Benjamins, 2003.
- [20] Luraghi Silvia. *Experiential verbs in Homeric Greek*. — Leiden: Brill, 2020.
- [21] Luraghi Silvia. *Beyond affectedness – Partitive objects and degrees of agenthood in Ancient Greek, Linguistic Variation*. — Amsterdam: John Benjamins, 2023. — Vol. 23(1), P. 95–123
- [22] Luraghi Silvia, Mertyrus Dionysios. 2021. *Basic valency in diachrony: from Ancient to Modern Greek, Valency over time*. — Berlin: De Gruyter, 2021. — P. 169-208.
- [23] Minnema Gosse, Gemelli Sara, Zanchi Chiara, Caselli Tommaso, Nissim Malvina. *SocioFillmore: A Tool for Discovering Perspectives // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. — Dublin, Ireland: Association for Computational Linguistics, 2022. — P. 240–250.
- [24] Pustejovsky James, Moszkowicz Jessica. *The qualitative spatial dynamics of motion in language, Spatial Cognition & Computation*. — Milton Park, UK: Taylor & Francis, 2011 — Vol. 11(1), P. 15-44.
- [25] Pustejovsky James. *Dynamic event structure and habitat theory // Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*. — Pisa, Italy, 2013. — P. 1-10.
- [26] Ruppenhofer Josef, Ellsworth Michael, Petruck Miriam R. L., Johnson Christopher R., Baker Collin F., Scheffczyk. *FrameNet II: Extended Theory and Practice*. — https://framenet.icsi.berkeley.edu/fndrupal/the_book, 2016.
- [27] Stoyanova Ivelina, Leseva Svetlozara. *Expanding the Conceptual Description of Verbs in WordNet with Semantic and Syntactic Information // Proceedings of the 12th Global WordNet Conference*. — Donostia/San Sebastián, Spain, 2023.
- [28] Tyler Andrea, Evans Vyvyan. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. — Cambridge, UK: Cambridge University Press, 2003.
- [29] Zanchi Chiara, Luraghi Silvia, Biagetti Erica 2021. *Linking the Ancient Greek WordNet to the Homeric Dependency Lexicon. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2021” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2021”]*. — P. 729-737.

Painting the Senate #Green: A Corpus Study of Twitter Sentiment Towards the Italian Environmentalist Blitz¹

Antonio Bianco
University of
Bergamo/Pavia
Piazza del Lino, 2, 27100
Pavia (PV), Italy
antonio.bianco@unibg.it

Claudia Roberta Combei
University of Pavia
Piazza del Lino, 2, 27100
Pavia (PV), Italy
claudiaroberta.combei@unipv.it

Chiara Zanchi
University of Pavia
Piazza del Lino, 2, 27100
Pavia (PV), Italy
chiara.zanchi01@unipv.it

Abstract

This study analyzes the reactions of the Italian Twitter community to an environmental demonstration that occurred in Rome on January 2nd, 2023. We compiled a corpus of 368,531 tokens consisting of 11,780 tweets, collected during a 7-day period. We propose a mixed-method approach that combines automated and manual corpus analyses of sentiment, emotions, and implicit language. Our findings offer insights into how tweets reflected the users' attitudes toward a variety of subjects and entities. Although the sentiment of the overall debate was distributed rather evenly, the incident itself seems to have sparked negative sentiment and emotions among Twitter users. The results of our manual analyses revealed some issues with respect to the automatic classification of sentiment, due to the fact that some tweets contained irony, sarcasm, and slurs. Non-literal interpretations were ignored by the tools at hand that could not account for complex rhetorical-argumentative strategies.

Keywords: Twitter discourse; Italian; sentiment analysis; environmental demonstration

DOI: 10.28995/2075-7182-2023-22-1021-1031

Раскрашивание Сената в зелёный цвет. Корпусный анализ тональности в Твиттере по поводу акции итальянских экологов

Антонио Бьянко
Университет
Бергамо/Павии
Пьяцца дель Лино, 2,
27100 Павия (PV), Италия
antonio.bianco@unibg.it

Клаудия Роберта Комбей
Университет Павии
Пьяцца дель Лино, 2, 27100
Павия (PV), Италия
claudiaroberta.combei@unipv.it

Кьяра Занки
Университет Павии
Пьяцца дель Лино, 2, 27100
Павия (PV), Италия
chiara.zanchi01@unipv.it

Аннотация

В данном исследовании анализируется реакция итальянского сообщества в Твиттере на экологическую демонстрацию, которая прошла в Риме 2 января 2023 года. Для этой цели мы составили корпус из 368531 токена, включающий 11780 твитов, собранных в течение 7 дней. Нами был использован смешанный метод, сочетающий автоматический и ручной анализ тональности, эмоций и имплицитности языка на материале корпуса. Результаты показывают, что твиты отражают отношение пользователей к различным субъектам и организациям. Хотя тональность в ходе общих дебатов распределилась довольно равномерно, сам инцидент вызвал среди пользователей Твиттера негативные тональность и эмоции. Результаты ручного анализа выявили некоторые спорные случаи автоматической классификации тональности, связанные с тем, что некоторые твиты содержали иронию, сарказм и оскорбления. Поскольку использованные инструменты не учитывали сложные риторико-аргументативные стратегии, небуквальные интерпретации игнорировались.

Ключевые слова: дискурс в Твиттере; итальянский; анализ тональности текста; экологическая демонстрация

¹ This work represents a collaborative effort undertaken by all three authors, who have consistently contributed throughout its development. In accordance with the academic requirements in Italy, authorship attribution is specified as follows: Antonio Bianco authored section 4, Claudia Roberta Combei authored sections 2 and 3, and Chiara Zanchi authored sections 1 and 5.

1 Introduction

This paper analyzes the Italian Twitter users' reactions toward one of the latest "civil disobedience actions" by a group of environmental activists called *Ultima Generazione* 'last generation' (<https://ultima-generazione.com/chi-siamo/>). On January 2nd, 2023, five young activists splattered orange washable paint onto the façade of the Italian Senate to protest the immobilism of Italian politics at facing the current "ecoclimatic collapse".² This demonstration has been both minimized and strongly condemned; for sure, it raised a fierce debate in the Italian public opinion.

The paper unfolds as follows. Section 2 describes how our corpus of tweets was collected and preprocessed, as well as the tools used to analyze it. Section 3 contains the quantitative analyses. Section 4 qualitatively comments upon tweets that reveal to be challenging for sentiment analysis. Section 5 concludes the paper.

2 Data and methods

Over the last 20 years, there has been a significant change in how people interact. The development of social media platforms, such as Twitter or Facebook has played a crucial role in this transformation. Today, people use Twitter to stay informed, build networks, establish ingroup identities, and express opinions. Twitter has been, in fact, the subject of numerous studies that employed sentiment analysis tools to measure the attitudes of the public opinion on a variety of topics ([1], [2], [3]).

Our paper aims to measure the opinions of the Italian Twitter community regarding the environmental blitz described in section 1. For this goal, we collected a sample of 53,301 tweets using the Twitter API and the `academicwitteR` package [4] for R [5]. The complete dataset consisted of all Italian tweets, retweets, and quotes that matched our search criteria, based on a set of hashtags related to the event, shown below in Table 1:

Hashtag	English translation	Hashtag	English translation
#alessandrosulis	#alessandrosulis	#ecovandali	#ecovandals
#blitz	#blitz	#gas	#gas
#carabinieri	#carabinieri	#imbrattato	#stained
#clima	#climate	#inquinamento	#pollution
#climateemergency	#climateemergency	#larussa	#larussa
#crisiclimatica	#climatecrisis	#lauraparacini	#lauraparacini
#davidenensi	#davidenensi	#senato	#senate
#digos	#digos	#ultimagenerazione	#lastgeneration
#ecocretini	#ecocretins		

Table 1: The hashtags used for the queries

Even though the episode happened on January 2nd, 2023, at 08:00 AM, we extended our period of analysis to a full week, from January 2nd 5:00 AM through January 9th at 12:00 AM. This allowed us to also capture tweets written the days after the incident that might have contributed to framing the debate on this topic across Twitter. All duplicates in the corpus were automatically removed and the remaining data were manually examined to eliminate irrelevant tweets. The resulting corpus included 11,780 original tweets, consisting of 368,531 tokens, for a total of 23,733 lemmas. The tweets were processed and analyzed, both automatically and manually, to account for the perceptions and attitudes of the users.

The computerized corpus investigation consisted of several analyses. First, we explored the distribution of the most frequent lemmas in this debate. The corpus was lemmatized using the pre-trained model UD Italian PoSTWITA ([6], [7]) for tweets, available in the `udpipe` package [8] for R. The lemmatized corpus was manually adjusted to correct recurrent issues, such as the unnecessary lemmatization of proper names and the incorrect lemmatization of abbreviations, keywords, and

² The news can be read here: <https://video.corriere.it/cronaca/ultimo-blitz-ambientalisti-imbrattano-facciata-senato-fermati/21cb0ef6-8a71-11ed-8b19-cdc718310dd5>

misspelled text. The first 40 most frequent lemmas were plotted using the `ggplot` function from the `tidyverse` package [9] for R.

Given the main aim of this study, we employed a mixed-method approach to measure sentiment and emotions. We started from a lexicon-based technique that allowed us to map the lemmatized tweets to the Italian version of the `nrc` resource for sentiment and emotions proposed by [10] and made available through the packages `syuzhet` [11] and `tidytext` [12] for R. This readily available 13,875-word lexicon is based on [13] and [14], and identifies eight primary emotions: *trust*, *joy*, *anticipation*, *surprise*, *sadness*, *disgust*, *anger*, and *fear*. The approach assumes that a word may be associated with one or more emotions. If a lemma in the corpus is not matched to any lexicon entry, it will not be assigned sentiment and emotion scores. Before using this lexicon-based tool, we applied three stop-words lists for Italian (`snowball`, `stopwords-iso`, and `nlk`), available on the `tidytext` package, to remove lexically empty or uninformative words (e.g., prepositions, conjunctions, etc.). Numbers, punctuation, and one-character sequences were also stripped from the corpus. Sentiment is analyzed with the `get_sentiment`, `rescale_x_2`, and `get_dct_transform` functions in the `syuzhet` package that iterate over each tweet and return numeric values (from -1 for extremely negative to 1 for extremely positive) based on the mapping with the `nrc` lexicon. In addition, the `get_sentiment` function assigns numerical values to each primary emotion detected in the tweet.

The other technique proposed here regards the sentiment analysis of tweets with the tools of the `TextBlob` library [15] for Python, built upon the `NLTK` library [16]. Before analyzing the corpus, we automatically translated the tweets into English with the `Googletrans` library [17]. The potential errors generated by the translation process are a limitation of this study. We are aware that some subtleties of the original tweets were lost or altered in translation (e.g., the word *gretino*, a pun combining *cretino* ‘cretin’ and *Greta*, a reference to the environmental activist Greta Thunberg, was mistranslated as ‘Greek’). Nevertheless, we think that the solution proposed here, which combines machine translation and `TextBlob`, an established approach in the scholarly research on sentiment analysis in languages other than English [18], can be useful in assessing the Twitter users’ perceptions of this incident. The `TextBlob` library employs a Naïve Bayes classifier pre-trained on film reviews and considers the weight of intensifiers, emoticons, emojis, and exclamation marks to calculate the sentiment. The score is expressed on a scale between -1 (extremely negative) to 1 (extremely positive). Using average sentiment scores may help summarize and simplify sentiment analysis results, providing a concise representation of the overall sentiment in a corpus. Nevertheless, it is important to acknowledge that averages should not be regarded as anything more than what they are: measures of central tendency that fail to capture the intricacies or context-specific sentiment. Therefore, the average scores at the corpus level for both systems are calculated for comparison purposes only.

Parallely, a sample³ of 49 tweets was built for an experiment with ChatGPT, to test how this advanced language model interpreted the sentiment of some corpus data and to compare its performance to our NLP systems. The qualitative analyses examined the use of irony within this sample of tweets. This specific focus was chosen due to the inherent limitations of automated sentiment detection tools, particularly in relation to ironic tweets, as shown in sections 3 and 4. By conducting manual analyses, we also delved into the persuasive impact of slurs within the said set of tweets.

The results of the analyses were plotted with the `Matplotlib` visualization library for Python [19].

3 Quantitative analyses

Extracting the most frequent lemmas allowed us to gain preliminary insights into the subtopics discussed in the debate around the incident described in section 1. We intentionally retained words closely associated with the environmental demonstration under investigation, as our objective was to comprehend the significance of each subtopic within our corpus. As Figure 1 shows, *senato* ‘senate’ was the most frequent word in the corpus, holding a prominent position and appearing frequently both as a reference to the institution and the building. The second and the third most frequent lemmas were *imbrattare* ‘to stain’ and *vernice* ‘paint’; the choice of *imbrattare* may suggest that painting the senate walls (*muro* ‘wall’ and *facciata* ‘façade’, other frequent lemmas) was largely perceived as a vandalism act on the Twitter arena. The high frequency of the hashtag *#ultimagerazione* ‘last generation’ and of

³ The sample is here: https://osf.io/5jve9/?view_only=779b0287c774498ea85ea1b5d507ea27

the lemmas *ultimo* ‘last’, *generazione* ‘generation’, *attivista* ‘activist’, *ambientalista* ‘environmentalist’, *ragazzo* ‘boy’, and *ragazza* ‘girl’ indicates that Twitter users identified the environmentalist activists with the younger generation. The high frequency of lemmas such as *clima* ‘climate’, *climatico* ‘climatic’, *ambiente* ‘environment’, placed at the center of the discussion the environmental motivation of the protest. However, environmental issues were viewed as either a valid or an inadequate justification of the incident, depending on the users’ perspective. Lemmas such as *fascista* ‘fascist’ and *civile* ‘civil’ permeated the discourse, implying that Twitter users discussed the appropriate ways of protesting in a democratic society. This episode received, in fact, attention from politicians, especially from the president of the senate, Ignazio La Russa (his name is frequent in the corpus), imputing potential political implications.

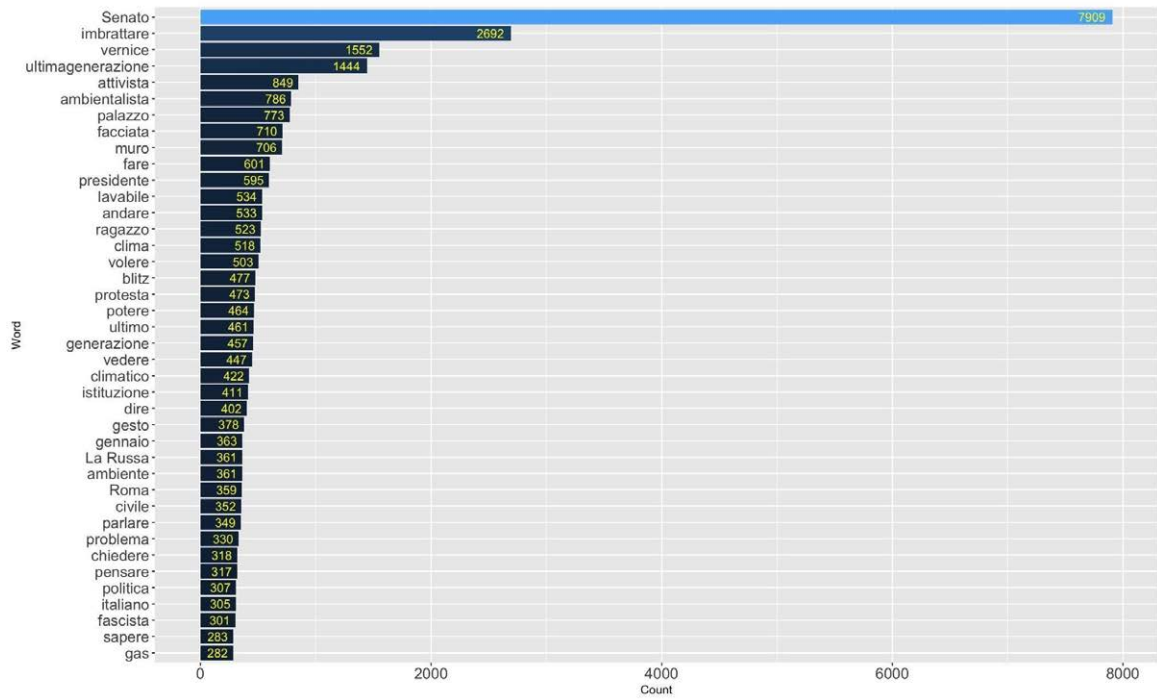


Figure 1: Top-40 most frequent lemmas

Next, we used the nrc lexicon to measure the sentiment and emotions conveyed in the debate. The overall average sentiment during the week considered was -0.059 (sd: 0.63), while the percentage of tweets labeled as negative, neutral, or positive is shown in Figure 2.

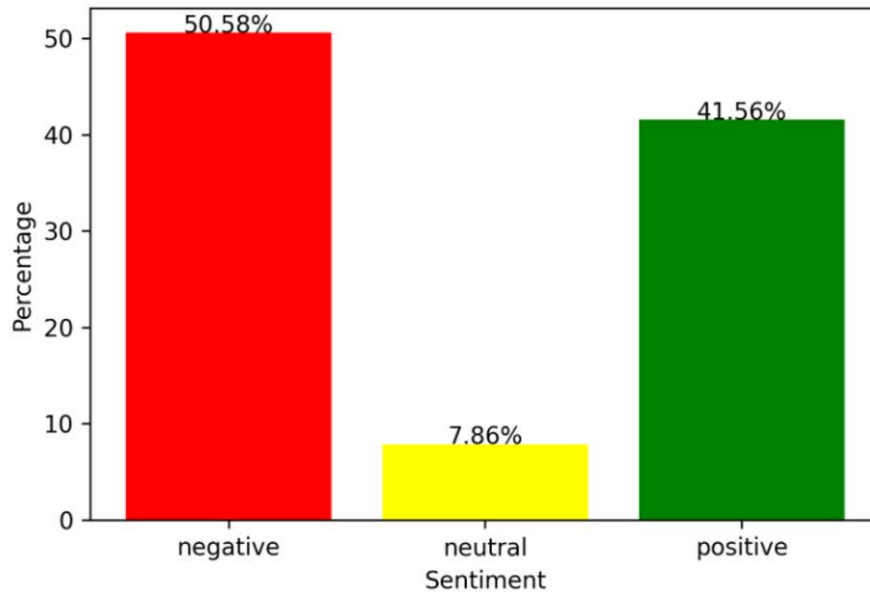


Figure 2: The distribution of sentiment based on nrc system

The average sentiment was only slightly negative (still, -0.059 is very close to a neutral score), but the fact that over 50% of the tweets were classified as negative may suggest that this environmental blitz was not debated in positive terms. However, the standard deviation of 0.63 indicates that tweets conveyed a wide range of sentiments: there was considerable variation in the users' attitudes with respect to this episode. For a better understanding of these results, we measured the emotions transmitted by the tweets. Figure 3 displays the percentage of words that were associated with each of Plutchik's primary emotions.

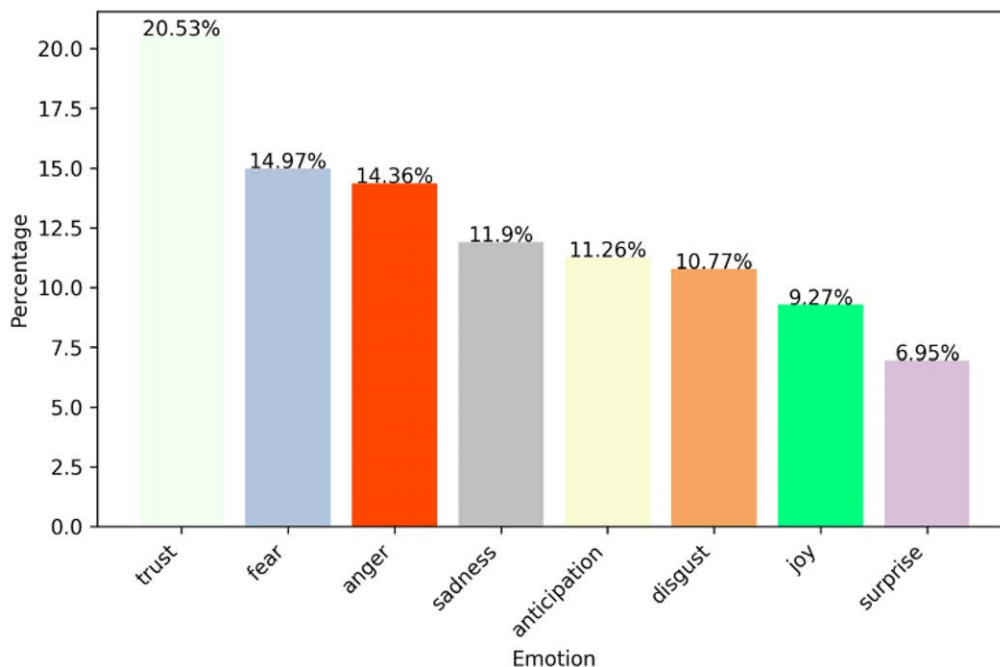


Figure 3: The distribution of emotions based on nrc system

Trust was the most frequent emotion, probably due to the frequent occurrence of political and institutional bodies and names (*senato* 'senate', *presidente* 'president', *politica* 'politics', etc.). Nevertheless, words related to trust did not always imply positive sentiment or support for politics and institutions; rather, they could transmit distrust or skepticism. In general, the incident appeared to have

elicited strong reactions among Twitter users, with a relatively higher occurrence of words associated with negative emotions, such as fear and anger. Specifically, words related to anger (*idiota* ‘idiot’, *delinquente* ‘delinquent’, *vergognoso* ‘shameful’, *condannare* ‘condemn’, etc.) seemed to be linked to a sense of breach of civil norms, while those evoking fear (*paura* ‘fear’, *attacco* ‘attack’, *pericoloso* ‘dangerous’, *danno* ‘damage’, etc.) tended to reflect worries about the possible effects of the incident. Finally, the episode did not appear to have surprised Twitter users, given the low percentage of words associated with this emotion.

To better capture the sentiment of the Twitter debate, we also measured it with TextBlob. The average sentiment score during the week considered was 0.035 (sd: 0.26), that appears aligned to the score we reported for the nrc system. However, at this point, it is important to note that when sentiment is extracted from a large corpus of tweets, its average value at corpus level tends to converge towards 0, since positive and negative values cancel one another out. This does not necessarily mean that there is a bigger percentage of neutral sentiment. Rather, it implies that the average score at the corpus level approaches 0 due to the counterbalancing effect of positive and negative sentiment expressed within the tweets.

For this reason, we also explored the distribution of sentiment as a function of time, as displayed in Figure 4.

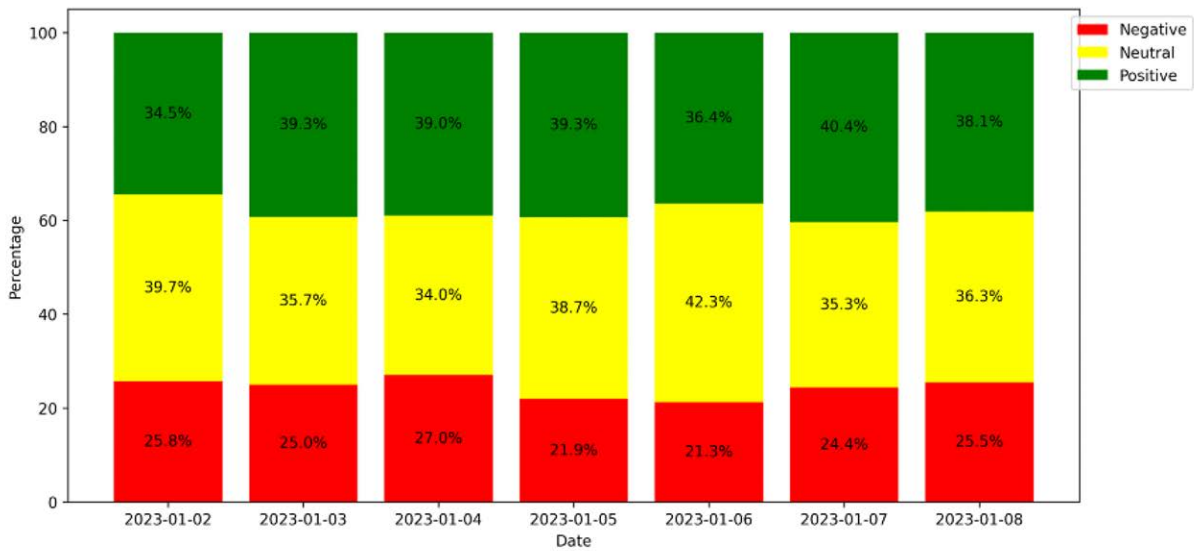


Figure 4: The percentage of sentiment as a function of time based on TextBlob system

In contrast to the findings of the nrc system, which exhibited a significant predominance of either negative (50.58%) or positive (41.56%) sentiment values, accompanied by a relatively low percentage of neutral tweets (7.86%), Figure 4 depicts a more balanced distribution across the three classes as regards the results obtained with the TextBlob system. In fact, 38.06% of the tweets were classified as positive, 36.68% as neutral, and 25.26% as negative. The largest proportion of tweets with a negative sentiment were published on January 4th - two days after the event - while most positive tweets occurred on January 7th. Instead, January 6th had the highest percentage of neutral tweets. Figure 4 suggests that the perspectives and responses of public opinion towards the incident changed only slightly during the week considered, with the results of the repeated measures ANOVA indicating that there were no statistically significant differences in the distribution of the three classes of sentiment over the seven days (F-value = 0.00187, p-value > 0.05, $\eta^2 < 0.001$). At the same time, the value of the standard deviation (0.26) at the corpus level suggests that the reactions of the Twitter users were not unanimous.

Due to the large number of tweets analyzed automatically, it was difficult to verify manually the accuracy of each label as well as to match it to the intended targets of the sentiment and emotions transmitted. In fact, tweets addressed several topics related to the environmentalist episode, and the sentiment reflected the users’ perception and reactions towards a variety of subjects and entities, including the environmental activists of Last Generation, the incident itself, the Senate, several politicians, and governmental organizations such as *Carabinieri*. Therefore, we purposely selected a

sample of 49 tweets that posed challenges to the sentiment analysis tools we used, because they contained irony, sarcasm, or slurs. In fact, these linguistic devices significantly impacted the sentiment of the tweets, and in many cases, the nrc system (and to a lesser extent, TextBlob) failed to accurately identify the correct sentiment score of tweets conveying irony and sarcasm. Before conducting the qualitative manual analysis of these data (see section 4), we also wanted to determine whether sophisticated language models, such as the one used for ChatGPT, could assign the proper sentiment and emotion labels to this type of challenging tweets. The said sample of 49 tweets was classified by ChatGPT into the following categories: 40 tweets as negative, 8 as neutral, and 1 as positive. In general, the labels reflected the sentiment we associated with these tweets, indicating rather correct identifications of instances of irony, sarcasm, amusement, and expressions evoking disgust and resentment. Nevertheless, 4 tweets with ironic content expressing negative sentiment towards the activists and the act of staining the walls were misclassified by ChatGPT as neutral instead of negative. Some of these tweets are commented on in section 4 below, specifically tweets (6)-(7).

4 Qualitative analyses

This section discusses irony and derogatory epithets or slurs, which are essential for understanding the Twitter debate on the environmental blitz, but which posed challenges to our NLP tools for sentiment analysis. Ignoring these phenomena may lead to a misinterpretation of what lies beneath the explicit content of tweets.

4.1 The interplay between irony and sentiment analysis

Figurative language (e.g., irony, metaphors, hyperboles) is one of the most complex aspects of human language, when it comes to NLP sentiment analysis ([20], [21]). Irony deserves particular attention, as it significantly characterizes the Twitter media ecosystem [20].

Irony is an implicit linguistic device: a content textually unavailable but nonetheless conveyed and present in the locutor's communicative intentions [22]. In fact, it is a conversational implicature generally induced by an intentional violation of the Maxim of Quality [23]. With an ironic statement, speakers usually want to communicate the opposite of what is explicitly said [24]. Thus, ironic utterances may reverse the polarity of the message in which they occur [25]. Since irony is contextually dependent and its decoding requires complex mental operations, NLP tools may fail in classifying the sentiment of ironic statements.

In fact, we encountered three main issues when using the nrc lexicon and the TextBlob system for sentiment analysis. First, some tweets, classified as having a negative sentiment toward the Last Generation's act, actually sympathized with activists (1)-(3). Second, some tweets that apparently expressed positive sentiment toward the activists' actions in fact ironically criticized it (4)-(6). Third, some tweets (7)-(8) that were assigned neutral and/or positive sentiment, due to irony, conveyed negative sentiments toward the activists (7) or mocked the incident (8).

Since irony recognition is largely dependent on the reader's subjectivity [26], we resorted to a qualitative analysis to investigate the sentiment of this type of tweets.

	Original tweet	Translation into English	nrc polarity	TextBlob polarity
(1)	Ma parliamo un altro po' di quella vernice brutta brutta lanciata sul Senato!	But let's talk a bit more about that very nasty paint being thrown over the Senate!	-0.86	-0.27
(2)	#evasorifiscali in libertà e dei ragazzi per 2 graffiti con vernice lavabile in carcere	#taxevaders free and some boys for 2 graffiti with washable paint in prison	-0.22	-0.35
(3)	Stiamo vivendo una crisi climatica irreversibile ma povero muro del senato	We are living in an irreversible climate crisis but poor senate wall	-1	-0.4

Tweets in (1)-(3) imply ironically the opposite of what is explicitly communicated, minimizing the activists' actions and showing solidarity with them. In addition, the opposite of (1) conveys a sort of impatience regarding the prominence this news was given in the media. In (1), irony is signaled both by the exclamation point and the reduplication of *brutta* 'nasty'; this adjective led our tools to classify this tweet as negative. Reduplication, intensifying the adjective meaning, makes the literal interpretation of (1) too overstated to be taken as true (also, adjective reduplication is one of the features of Italian baby talk). The occurrence of a term expressing judgments and feelings (*brutta* 'nasty') is a further indication of irony [26]. In (2) mentioning tax evaders' freedom is unexpected, and such freedom clashes with the fact that young activists are in prison, despite using washable (and not permanent) paint. It is precisely such unexpectedness [27] that allows us to infer the opposite of what is stated. Similar considerations concern the tweet shown in (3). Indeed, based on the author's viewpoint, it seems senseless to engage in condemning a protest action while ignoring irreversible climate change.

	Original tweet	Translation into English	nrc polarity	TextBlob polarity
(4)	Buongiorno dal #Senato. Anche oggi abbiamo salvato il clima!	Good morning from the #Senate. Even today we saved the climate!	0.6	0.7
(5)	Con il Senato imbrattato, il clima sarà sicuramente migliore	With the Senate smeared, the climate will certainly be better	1.0	0.10
(6)	In realtà sono degli artisti lasciate che sfoghino liberamente la loro creatività	In fact they are artists let them freely vent their creativity	0.1	0.2

Tweets in (4)-(6) were classified as positive by both systems. The overstatements of (4) and (5) lead us to interpret these tweets as ironic. The same applies to (6). What is said in (6) seems to be a violation of the Maxim of Quality, as the locutor cannot *bona fide* assert that what happened was an artistic gesture. Thus (4)-(6) convey dissent towards Last Generation's action.

	Original tweet	Translation into English	nrc polarity	TextBlob polarity
(7)	ultima generazione lo spero proprio	last generation I really hope so	0.6	0.0
(8)	Coi secchi di vernice coloriamo tutti i muri, case, vicoli e palazzi, perché lei ama i colori. #Senato	With buckets of paint let's color all the walls, houses, alleys and buildings, because she loves colors. #Senate	0.0	0.0

Tweets (7)-(8) exemplify particular types of irony. Specifically, (7) is a case of sarcasm, a sharper kind of irony. To understand (7), one does not need to infer the opposite of what is asserted; irony is triggered by joking with the ambiguity [21] of *ultima generazione* 'last generation'. The expression is not used to properly name an environmentalist group but is to be interpreted literally. Therefore, the tweet is to be read as 'I hope this is the last generation of (this type) of activists'. So, the sentiment assigned by the tools is inconsistent with the sentiment that arises from what is implicitly communicated (ChatGPT also misclassified this tweet as neutral, see section 3).

In (8) a case of echoic irony [28] is exemplified. The Last Generation's act is described by echoing words of a well-known song by the singer Riccardo Cocciante. The hashtag *#Senate* refers to the incident. The irony arises as the motivations underlying the event described in the song (a romantic gesture) and the activists' actions are completely different. In (8), irony generates hilarious effects that tend to mitigate the disagreement.

4.2 The power of slurs

In many tweets, including (9)-(12), Last Generation activists were designated by slurs [29]. Slurs have a derogatory connotation that eases the recognition of a negative sentiment by the systems we used. The (persuasive) appeal of slurs goes beyond their explicit derogatory connotation, as these expressions convey implicit meanings [30]. In particular, slurs allow the locutors to implicitly evoke stereotypes associated with the criticized social groups, in this case Last Generation activists, and to indirectly communicate the superiority of their own group. As a result, slurs, through their association with stereotypes, have the power to (implicitly) elicit a greater degree of negative sentiment compared to other negative words. In addition, stereotypes pertain to the common ground shared between the addressees and the locutors and, therefore, they could not be computed by the sentiment analysis systems at hand.

	Original tweet	Translation into English	nrc polarity	TextBlob polarity
(9)	I tre gretini di Ultima Generazione sono stati già rilasciati	The three cretins with "g" of Last Generation have already been absolved	-0.07	0.0
(10)	Le bimbettole arcobalenate della merda buttata addosso agli altri [...]	The little rainbow girls of shit thrown at others [...]	-0.66	-0.1
(11)	Questi non sono ecoterroristi come al senato?	Aren't these eco-terrorists as in the senate?	-0.4	-0.5
(12)	[...] gli ecovandali imbrattano il Senato	[...] the ecovandals soil the Senate	-0.33	-0.4

For example, *gretini* (see section 2) and *bimbette arcobalenate* ‘little rainbow girls’ imply unpleasant and questionable stereotypes which may sound as follows: activists are childish people, not very smart, and live in an ideal world without connections to reality. Moreover, these stereotypes represent (9) and (10) as hilarious or facetious tweets [25] that increases the locutor’s pleasantness. Generally, the more pleasant the locutor, the greater their persuasive force. Instead, *ecoterroristi* ‘eco-terrorists’ (11) and *ecovandali* ‘eco-vandals’ (12) implicitly identify the actions performed by environmental activists as detrimental, and activists as members of a dangerous social group. However, this implicit interpretative level was ignored by the computerized sentiment analysis tools we used (and also by ChatGPT) that, therefore, were not always able to account for the complex rhetorical-argumentative strategies and stereotypes underlying the use of slurs.

5 Conclusion

In this paper, we compared different NLP tools at performing sentiment analysis on a corpus of Italian tweets related to the environmental blitz that occurred at the Senate in Rome on January 2nd, 2023. Our mixed-method approach, combining quantitative and qualitative analyses, indicated that while lexicon-based and Naïve Bayes classifier techniques provided interesting insights regarding the episode considered, they fell short in addressing inherent and notorious challenges for sentiment analysis, such as implicit language and indirect rhetorical-argumentative strategies. More sophisticated language models, such as ChatGPT, tackled these issues efficiently, by deciphering irony and sarcasm. In general, the results of our analyses indicated that the users’ attitudes were directed toward a variety of subjects and entities (the act itself, Italian politics, institutions, etc.) The sentiment expressed towards these entities varied within the corpus, without a clear majority position or stance in the Twitter debate. Nonetheless, upon conducting an in-depth qualitative manual analysis of a sample of 49 tweets, it appeared that the act of staining the Senate walls generated primarily negative sentiment among users.

References

- [1] Jelani Ince et al. The social media response to Black Lives Matter: how Twitter users interact with Black Lives Matter through hashtag use. // *Ethnic and Racial Studies*. – 2017. – Vol. 40, no. 11. – P. 1814-1830.
- [2] Ali Rao Hamza et al. A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election // *Journal of Big Data*. – 2022. – Vol. 9, article 79. – P. 1-12.
- [3] Kruspe Anna et al. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. // *Proceedings of the 1st Workshop on NLP for COVID-19 at the 58th Annual Meeting of the Association for Computational Linguistics*. – 2020. – P. 1-10.
- [4] Barrie Christopher and Ho Justin Chun-ting. *academicwitterR*: An R package to access the Twitter Academic Research Product Track v2 API endpoint. // *Journal of Open Source Software*. – 2021. – Vol. 6, no. 62, article 3272. – P. 1-2.
- [5] R Core Team. *R: A language and environment for statistical computing*. – 2022. – Vienna: R Foundation for Statistical Computing.
- [6] Sanguinetti Manuela et al. PoSTWITA-UD: An Italian Twitter Treebank in Universal Dependencies. // *Proceedings of the Eleventh International Conference on Language Resources and Evaluation – LREC2018*. – 2018. – P. 1397-1404.
- [7] Sanguinetti Manuela et al. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. // *Proceedings of the Twelfth International Conference on Language Resources and Evaluation – LREC2020*. – 2020. – P. 5240-5250.
- [8] Wijffels Jan. *udpipe*: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP Toolkit (Version 0.8.11). – 2023. – CRAN.
- [9] Wickham Hadley. *Tidy Data*. // *Journal of Statistical Software*. 2014. – Vol. 59, no. 10. – P. 1-23.
- [10] Mohammad Saif M. and Turney Peter D. Crowdsourcing a Word-Emotion Association Lexicon. // *Computational Intelligence*. – 2013. – Vol. 29. – P. 436-465.
- [11] Jockers Matthew. *Syuzhet*: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text (Version 1.0.6.). – 2020. – CRAN.
- [12] Silge Julia and Robinson David. *tidytext*: Text Mining and Analysis Using Tidy Data Principles. // *Journal of Open Source Software*. – 2016. – Vol. 1, no. 3. – P. 1-3.
- [13] Plutchik Robert. *The Emotions*. – 1991. – Lanham: University Press of America.

- [14] Plutchik Robert. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. // *American Scientist*. – 2001. – Vol. 4. – P. 344–350.
- [15] Loria Steven. TextBlob: Simplified Text Processing. (Version 0.16.0). – 2020. – GitHub.
- [16] Bird Steven et al. *Natural Language Processing with Python*. – 2009. – Sebastopol: O’Reilly Media.
- [17] Han SuHun. Googletrans: Free and Unlimited Python Library for Google Translate API. (Version 3.0.0). – 2020. – GitHub.
- [18] Sazzed Salim. Cross-lingual sentiment classification in low-resource Bengali language. // *Proceedings of the Sixth Workshop on Noisy User-generated Text - W-NUT 2020*. – 2020. – P. 50-60.
- [19] Hunter John D. Matplotlib: A 2D graphics environment. // *Computing in Science & Engineering*. – 2007. – Vol. 9, no. 3. – P. 90-95.
- [20] Dmitry Davidov et al. Semi-supervised recognition of sarcastic sentences in twitter and amazon // *Proceedings of the fourteenth conference on computational natural language learning* – 2011. – P. 107-116.
- [21] Reyes Antonio et al. From humor recognition to irony detection: The figurative language of social media. // *Data & Knowledge Engineering* – 2012. – vol. 74. – P. 1–12.
- [22] Sbisà Marina. *Detto non detto*. – 2007. – Roma-Bari: Laterza.
- [23] Grice Herbert Paul. *Logica e conversazione. Saggi su intenzione, significato e comunicazione*. – 1993. – Bologna: Il Mulino.
- [24] Colston Herbert and Gibbs Raymond. A brief history of irony // Gibbs Raymond and Colston Herbert (eds.). *Irony in Language and Thought*. – 2007. – London: Routledge. – P. 3-24.
- [25] Gianti Andrea et al. Annotating irony in a novel Italian corpus for sentiment analysis // *Proceedings of the Workshop on Corpora for Research on Emotion Sentiment and Social Signals – LREC 2012*. – 2012. – P. 1-7.
- [26] Shelley Cameron. The bicoherence theory of situational irony // Gibbs Raymond and Colston Herbert (eds.). *Irony in Language and Thought*. – 2007. – London: Routledge. – P. 531-578.
- [27] Lucariello Joan. Situational irony: A concept of events gone away // Gibbs Raymond and Colston Herbert (eds.). *Irony in Language and Thought*. – 2007. – London: Routledge. – P. 467-498.
- [28] Dan Sperber and Deirdre Wilson. *Relevance: communication and cognition*. – 1986. – Cambridge: Harvard University Press.
- [29] Jeshion Robin. Slurs and stereotypes // *Analytic Philosophy*. – 2007. – vol. 54, no. 3. – P. 314-329.
- [30] Lombardi Vallauri Edoardo. *La lingua disonesta. Contenuti impliciti e strategie di persuasione*. – 2019. – Bologna: Il Mulino.

Scaled Down Lean BERT-like Language Models for Anaphora Resolution and Beyond

Vladislav Bolshakov NTR Labs and Bauman Moscow State Technical University Moscow, Russia vbolshakov@ntr.ai	Rostislav Kolobov NTR Labs Tomsk, Russia rkolobov@ntr.ai	Eugene Borisov NTR Labs and Higher IT School of Tomsk State University Tomsk, Russia eborisov@ntr.ai	Nikolay Mikhaylovskiy NTR Labs and Higher IT School of Tomsk State University Moscow, Russia nickm@ntr.ai
--	--	--	---

Gyuli Mukhtarova
NTR Labs
Moscow, Russia
gmukhtarova@ntr.ai

Abstract

We study performance of BERT-like distributive semantic language models on anaphora resolution and related tasks with the purpose of selecting a model for on-device inference. We have found that lean (narrow and deep) language models provide the best balance of speed and quality for word-level tasks, and opensource¹ RuLUKE-tiny and RuLUKE-slim models we have trained. Both are significantly (over 27%) faster than models with comparable accuracy. We hypothesise that the model depth may play a critical role for performance as, according to recent findings each layer behaves as a gradient descent step in autoregressive setting.

Keywords: BERT, LUKE, Anaphora Resolution

DOI: 10.28995/2075-7182-2023-22-1032-1040

Поджарые BERT-подобные модели для разрешения анафоры и не только

Владислав Большаков
ООО “НТР”
МГТУ им. Баумана
Москва
vbolshakov@ntr.ai

Ростислав Колобов
ООО “НТР”
Томск
rkolobov@ntr.ai

Евгений Борисов
ООО “НТР”
Высшая ИТ-Школа ТГУ
Томск
eborisov@ntr.ai

Николай Михайловский
ООО “НТР”
Высшая ИТ-Школа ТГУ
Москва
nickm@ntr.ai

Гюли Мухтарова
ООО “НТР”
Москва
gmukhtarova@ntr.ai

Аннотация

Изучена эффективность BERT-подобных моделей на задачах разрешения анафоры и смежных задачах, чтобы выбрать модели для использования на оконечном устройстве. Выяснено, что поджарые (узкие и длинные) языковые модели дают оптимальное соотношение скорости и качества. Представлены модели RuLUKE-tiny и RuLUKE-slim с открытым исходным кодом. Обе заметно (более чем на 27%) быстрее, чем модели со сравнимой точностью. Предположено, что глубина модели может играть решающую роль для ее эффективности, поскольку, согласно недавним исследованиям, каждый слой ведет себя как шаг градиентного спуска в условиях авторегрессии.

Ключевые слова: BERT, LUKE, разрешение анафоры

¹<https://huggingface.co/vbolshakov/RuLUKE-tiny>
<https://huggingface.co/vbolshakov/RuLUKE-lean>

1 Introduction

1.1 Anaphora Resolution

Anaphora is the use of an expression (a pronoun or a noun phrase) whose interpretation depends upon a preceding expression in context (its antecedent). Anaphora and cataphora (which is the use of an expression that depends upon a postcedent expression) both are special cases of coreference, which occurs when two or more expressions in a text refer to the same person or thing.

Anaphora resolution is the problem of resolving what a pronoun, or a noun phrase refers to. We are specifically interested in resolving pronoun anaphora. It is a challenging task because it requires good understanding of the context and the ability to recognize complex relationships between words and phrases (Bolshakov and Mikhaylovskiy, 2023). However, this task is crucial in many applications of NLP, such as information retrieval (Schmolz, 2015), question answering (Castagnola, 2002), opinion mining (Jakob and Gurevych, 2010), and natural language understanding (Kilicoglu et al., 2016). In addition, anaphora resolution can be used to improve the readability of a text, by replacing repeated mentions of the same entity with a pronoun or other reference.

Recent anaphora and coreference resolution approaches typically use some fine-tuned pretrained language model. As coreference resolution approaches are reviewed in detail recently by (Bolshakov and Mikhaylovskiy, 2023), here we only list some specifically anaphora resolution work. The use of BERT-like models for anaphora resolution was likely first suggested by (Joshi et al., 2019). At about the same time (Mohan and Nair, 2019) suggested resolving ambiguous pronoun anaphorae using BERT and SVM, and (Wang, 2019) suggested a BERT-based approach for gendered pronoun resolution. (Hou, 2020) suggests an approach to bridging anaphora resolution via question answering based on SpanBERT (Joshi et al., 2020).

1.2 Downscaling Transformers

A lot of recent research have focused on laws of and approaches to scaling transformer (Vaswani et al., 2017) language models up (Hoffmann et al., 2022; Kaplan et al., 2020; Rae et al., 2021; Shoeybi et al., 2019). Significantly less effort is being devoted to building smaller and more compute-efficient models (Geiping and Goldstein, 2022). In this work we continue the latter line of research, with a focus on the use of transformers in anaphora resolution.

1.3 Our contribution

Our contribution in this paper is threefold:

- We cast the anaphora resolution problem in a form similar to named entity recognition and linking
- We empirically study the performance of varied language model architectures and training approaches and found that lean (narrow and deep) language models provide the best balance of speed and quality for word-level tasks,
- Finally, we open-source RuLUKE-tiny and RuLUKE-slim models we have trained that have better performance on our downstream tasks than comparable models, and the larger of two models we present performs on par with significantly larger models.

2 Anaphora Resolution Approach

For the anaphora resolution problem, we suggest an approach inspired by tagging named entities using embeddings extracted from the transformer model (see, for example, (Arhipov et al., 2019)). Instead of named entity BIO tags (introduced by (Ramshaw and Marcus, 1995), see also (Nadeau and Sekine, 2007)) we suggest the following four tags:

- 0) 0 - the tag of all the words that are not in an anaphoric connection with the target pronoun;
- 1) AT_B - the tag of the first token included in the antecedent;
- 2) AT_C - the tag of subsequent tokens included in the antecedent;
- 3) AF_B - the tag of the first token of the anaphora.

We measure the accuracy of models with F1 metric applied to each token type, producing 4 separate metrics. This approach allows to analyze the models' performance in detail.

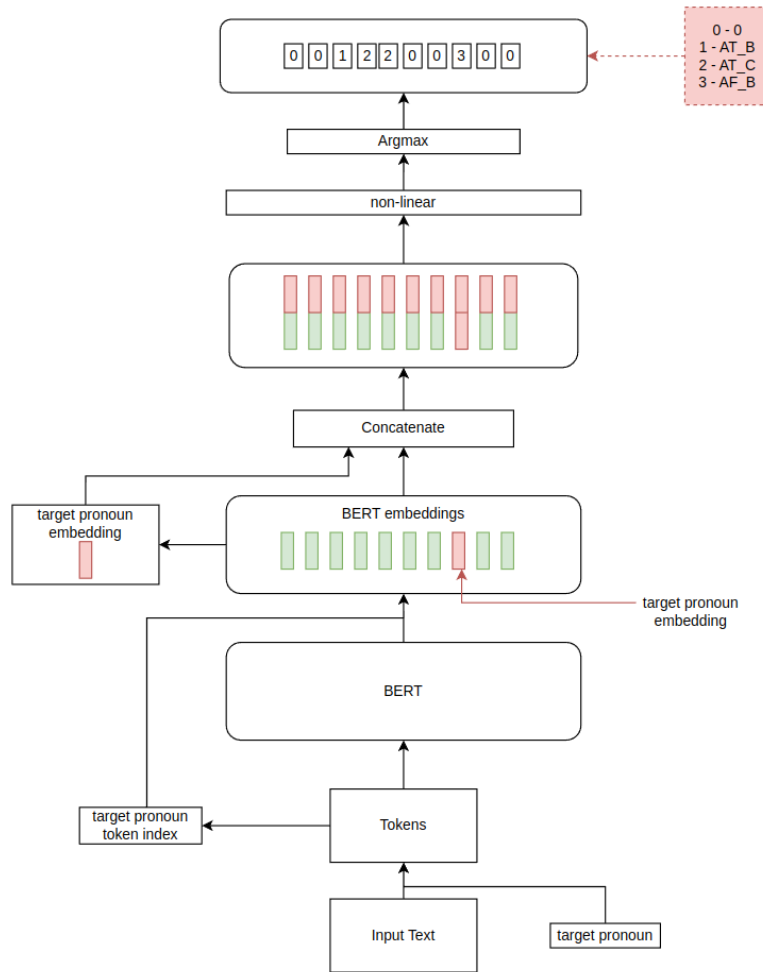


Figure 1: Architecture of the Anaphora Resolution model

Figure 1 depicts the suggested architecture for the anaphora resolution. Embeddings are generated for the tokens of the source text. The embedding of the token of the anaphora pronoun is concatenated with the embeddings of all tokens of the input text, and the result is passed to a fully-connected 2-layer network. At the output of the model, the argmax layer returns the indices of the most likely tags for each token. These tags are mapped to the words of the source text. In one pass over a window, the model finds antecedents for only one pronoun.

The largest (in terms of the number of words) antecedent is selected for each continuous span detected: "**Elizaveta Petrovna Kalinina** is the CEO of the company. **Liza** is responsible for a huge number of employees. Every day **she**...". The case and number of the antecedent are agreed with the pronoun. In cases where there were opening brackets/quotes in the antecedent, but their closing versions were not included, they are added.

3 Training approaches

We have benchmarked several approaches to training the models:

- Distillation
- Pretraining using SpanBERT approach (Joshi et al., 2020)
- Pretraining using LUKE approach (Yamada et al., 2020)

3.1 Distillation

We use knowledge distillation (Hinton et al., 2015) as a basic approach to training the models, following (Dale, 2021b). Knowledge distillation is the process of transferring knowledge from a large model (teacher) to a smaller one (student) (Gou et al., 2021).

We train bilingual English+Russian models and use two training sets:

- 2.5 million parallel English-Russian sentences collected from Yandex.Translate (Yan, 2022), OPUS-100 (Zhang et al., 2020) and Tatoeba (Tat, 2022; Tiedemann, 2012) corpora.
- 6.5 million sentences in Russian from ruswiki 2021, rusnews 2020 and rusweb 2019 collections from Leipzig corpus (Goldhahn et al., 2012) and Russian sentence pairs from XNLI (Conneau et al., 2018).

We use several losses and teachers. For the parallel corpus we, similarly to Dale (Dale, 2021b):

- distill CLS tokens, bringing their different projections closer to RuBERT (Kuratov and Arkhipov, 2019), LaBSE (Feng et al., 2022) and Laser (Artetxe and Schwenk, 2019) embeddings;
- distill the probability distribution of LaBSE (Feng et al., 2022) output tokens with MLM distillation loss, using the Kullback-Leibler divergence loss between mapped vocabularies of student and teacher models;
- minimize the whole-word MLM loss (Devlin et al., 2019) for English and Russian languages;
- minimize the translation ranking loss, as in LaBSE (Feng et al., 2022);

For the Russian corpus we

- minimize the per-token MLM loss with rubert-base-cased-sentence (Kuratov and Arkhipov, 2019);
- minimize the whole-word MLM loss to LABSE (Feng et al., 2022);
- minimize NLI loss.

We train models in three stages using Cosine Annealing with Warm Restarts on the first two stages. The details of the parameters on these stages are listed in Table 1. We have trained two models using this approach - one with rubert-tiny architecture but with extended dictionary (we call it distilRuBert-tiny) and the other with twice as much layers (we call it distilRuBert-lean). Input and output embeddings weights for these new models were partially copied from cointegrated/rubert-tiny2 and cointegrated/LaBSE-en-ru respectively. For distilRuBert-tiny we used cointegrated/rubert-tiny2 as a starting checkpoint.

	Stage 1	Stage 2	Stage 3
teachers	all	all	all but rubert-base-cased-sentence
steps	400000	800000	1100000
batch size for bilingual pairs	16	32	36
batch size for Russian texts	8	16	12
batch size for NLI	8	16	24
accumulation steps	4	4	1
learning rate	1e-5 to 0	1e-5 to 1e-6	1e-5

Table 1: Distillation stages description

3.2 Pretraining using SpanBERT approach

SpanBERT (Joshi et al., 2020) extends BERT by

- masking contiguous random spans, rather than random tokens, and
- training the span boundary representations to predict the entire content of the masked span, without relying on the individual token representations within it.

We only use the first option of these two. We have trained a model with rubert-tiny architecture using this approach, and refer to it distilRuSpanBert-tiny in this paper. We used cointegrated/rubert-tiny2 as a starting checkpoint.

Model	V	E	H	L	N
DeepPavlov/distilrubert-small-cased-conversational	119547	768	3072	2	106.4M
DeepPavlov/distilrubert-tiny-cased-conversational-v1	30522	264	792	3	10.3M
DeepPavlov/distilrubert-tiny-cased-conversational-5k	5031	264	792	3	3.6M
cointegrated/LaBSE-en-ru	55083	768	3072	12	127M
cointegrated/rubert-tiny2	83828	312	600	3	29.1M
cointegrated/rubert-tiny	29564	312	600	3	11.8M
(ours) distilRuBert-lean	55083	312	936	6	23.3M
(ours) distilRuBert-tiny	101520	312	600	3	34.3M
(ours) RuLUKE-tiny	83828	312	600	3	158.8M
(ours) RuLUKE-lean	55083	312	936	6	153.3M
(ours) distilRuSpanBert-tiny	101525	312	600	3	34.3M

Table 2: Parameters of models

3.3 Pretraining using LUKE approach

LUKE (Yamada et al., 2020) extends BERT by introducing:

- a new pretraining task that involves predicting randomly masked words and entities in a large entity-annotated corpus retrieved from Wikipedia
- an entity-aware self-attention mechanism that is an extension of the self-attention mechanism of the transformer, and considers the types of tokens (words or entities) when computing attention scores

We have trained two models using this approach - one with rubert-tiny2 architecture (we call it RuLUKE-tiny) and the other with distilRuBert-lean architecture (we call it RuLUKE-lean). For RuLUKE-tiny and RuLUKE-lean we used cointegrated/rubert-tiny2 and our distilRuBert-lean respectively as backbone transformers and starting checkpoints for further training. According to LUKE (Yamada et al., 2020), each model has additional entity vocabulary with top 500k entities from dump of Russian Wikipedia, that is why the disk size and the number of parameters of RuLUKE-tiny and RuLUKE-lean are larger compared to other models.

4 Experiments and results

4.1 Tasks and Datasets

We test the efficiency of the anaphora resolution approach overall and of each model in particular on the anaphora resolution subset of RuCoCo dataset (Dobrovolskii et al., 2022). To produce this subset we have sampled examples where one of the coreferences is a pronoun.

4.2 BERT-like Models

For our study, we selected small and medium sized BERT-like models that showed promising results in NLP tasks for the Russian language (Kolesnikova et al., 2022), (Dale, 2021b), based on the rating from Dale (Dale, 2022) and integrated well with spaCy (Honnibal and Montani, 2017). The size, performance and efficiency of BERT-like models depends on model architecture parameters and training approach. We treat the latter in Section 3. The key architectural parameters for BERT are:

- L - the number of hidden layers;
- H - the size of intermediate layer embeddings;
- E - the size of the output embedding;
- V - the size of vocabulary;
- N - the number of parameters (which is a function of the above parameters)

The Table 2 lists the architectural parameters of the key models we compare to and our models.

4.3 Inference Speed

We have benchmarked the performance of CPU inference of typical and potential architectures. The tests were run on the entire dataset. Time was measured in ms/sentence, mean of 3 runs, 1 loop each on an

Intel(R) Core(TM) i5-10400, 2.90GHz processor based computer with 6 cores. Batch size was set to 1, and torch.utils.data.DataLoader used $num_workers = 0$. The results are listed in Table 3. It is easy to see that the performance only slightly depends on the vocabulary size and intermediate embedding dimension, grows linearly with the number of layers and slower - with the embedding dimension.

		$E = 264$				$E = 528$				$E = 768$			
	L	3	6	9	12	3	6	9	12	3	6	9	12
V =	$H = E * 2$	10.7	19.4	28.5	37.2	15.5	29.5	43.0	56.7	23.9	45.9	67.9	90.0
29	$H = E * 3$	11.3	20.6	30.2	39.3	17.1	32.3	47.7	62.8	28.5	54.9	82.7	108.9
564	$H = E * 4$	11.4	21.1	31.0	40.7	20.3	39.1	61.8	77.6	33.8	65.0	96.4	127.7
V =	$H = E * 2$	11.8	21.5	30.9	40.6	16.2	30.6	45.2	59.6	24.6	47.1	69.8	95.1
83	$H = E * 3$	11.9	22.2	32.5	42.9	17.9	33.9	50.7	67.2	29.7	57.6	85.5	113.4
828	$H = E * 4$	12.3	22.8	33.3	43.8	21.2	40.7	60.2	79.9	35.0	68.1	101.1	133.7
V =	$H = E * 2$	11.8	21.8	32.1	41.9	16.5	31.9	49.1	61.7	24.5	46.8	69.1	92.1
119	$H = E * 3$	12.5	23.1	33.4	43.8	18.1	34.3	51.6	66.8	28.8	55.7	82.7	109.8
547	$H = E * 4$	12.6	23.3	35.2	44.5	21.2	40.8	60.1	79.6	34.0	66.0	98.2	130.3

Table 3: Dependence of performance (ms/sentence) on model architecture parameters L - the number of hidden layers, H - the size of intermediate layer embeddings, E - the size of the output embedding, V - the size of vocabulary

Table 4 shows CPU inference speed of discussed models on two benchmarks:

- *Benchmark1* - is the CPU speed task from (Dale, 2021a)
- *Benchmark2* - shows the performance of models on anaphora resolution task when running on CPU using the same data as in Section 4.1 and architecture as in Section 2

For both benchmarks we report mean inference time in milliseconds per sentence and standard deviation, collected on 3 runs.

Model	<i>Benchmark1</i>	<i>Benchmark2</i>
DeepPavlov/distilrubert-small-cased-conversational	5.22 ± 0.13	28.15 ± 0.67
DeepPavlov/distilrubert-tiny-cased-conversational-v1	2.87 ± 0.06	15.92 ± 0.76
DeepPavlov/distilrubert-tiny-cased-conversational-5k	3.18 ± 0.07	15.36 ± 0.63
cointegrated/LaBSE-en-ru	28.10 ± 0.55	134.07 ± 1.20
cointegrated/rubert-tiny2	3.09 ± 0.12	15.70 ± 0.08
cointegrated/rubert-tiny	3.23 ± 0.09	14.86 ± 0.07
(ours) distilRuBert-lean	6.12 ± 0.09	20.73 ± 0.53
(ours) distilRuBert-tiny	3.23 ± 0.07	12.40 ± 0.47
(ours) RuLUKE-tiny	3.31 ± 0.13	11.79 ± 0.57
(ours) RuLUKE-lean	6.12 ± 0.09	22.01 ± 1.18
(ours) distilRuSpanBert-tiny	3.38 ± 0.18	12.29 ± 0.17

Table 4: CPU inference speed of models

4.4 Accuracy

We list the results of accuracy evaluation in two groups - tiny (Table ??) and larger (Table 5) models. The results for LaBSE-en-ru are listed with larger models for comparison. The best results in each category are highlighted in bold.

5 Conclusion

Popular tiny Russian BERT models are trained primarily with sentence-related tasks in mind. Thus their accuracy on word-related tasks is significantly worse than on sentence-related tasks. It is hard to fine-

Model	AF_B F1	AT_B F1	AT_C F1	0 F1	Avg F1	Time
(ours) RuLUKE-tiny	0.966	0.372	0.484	0.509	0.583	11.79
(ours) distilRuSpanBert-tiny	0.973	0.290	0.438	0.471	0.543	12.29
(ours) distilRuBert-tiny	0.974	0.315	0.430	0.458	0.544	12.40
cointegrated/rubert-tiny	0.974	0.327	0.416	0.447	0.541	14.86
DeepPavlov/ distilrubert-tiny-cased-conversational-5k	0.972	0.333	0.425	0.461	0.548	15.36
cointegrated/rubert-tiny2	0.964	0.318	0.437	0.468	0.547	15.70
DeepPavlov/ distilrubert-tiny-cased-conversational-v1	0.972	0.380	0.473	0.511	0.584	15.92
distilRuBert-lean	0.975	0.422	0.490	0.521	0.602	20.73
RuLUKE-lean	0.975	0.411	0.500	0.528	0.604	22.01
DeepPavlov/ distilrubert-small-cased-conversational	0.972	0.382	0.499	0.541	0.599	28.15
cointegrated/LaBSE-en-ru	0.987	0.713	0.770	0.822	0.823	134.07

Table 5: Accuracy and speed of the models

tune/distill such models to achieve better accuracy on word-related tasks than model trained from scratch with word-related tasks in mind. LUKE improves performance on word-related tasks to be on par with the best similarly-sized models, but but is much faster so RuLUKE-tiny is 35% faster than DeepPavlov/distilrubert-tiny-cased-conversational-v1 that has about the same accuracy. SpanBERT training does not improve performance on anaphora resolution task.

For lean models, the accuracy improvement achieved by LUKE training is more noticeable, and the speedup compared to DeepPavlov/distilrubert-small-cased-conversational is 28%. Still, 6 layers seems to be an inadequate number to match the performance of full-fledged, 12-layer models such as LaBSE. We believe this might be connected with the recent finding that transformers learn in-context by gradient descent in the domain of autoregressive problems (von Oswald et al., 2022). In the latter setting each layer behaves as a gradient descent step. While our formulation of anaphora resolution task is not autoregressive, a similar mechanism may also be present. This is a matter of the future research.

Acknowledgements

The authors are grateful to their colleagues at NTR Labs ML division for the discussions and support, Prof. Sergey Orlov and Prof. Oleg Zmeev for the computing facilities provided, and anonymous referees who have provided valuable feedback that allowed to significantly improve the paper.

References

- Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. // *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, P 89–93, Florence, Italy, August. Association for Computational Linguistics.
- Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Vladislav Bolshakov and Nikolay Mikhaylovskiy. 2023. Pseudo-labelling for autoregressive structured prediction in coreference resolution. // *Dialogue-2023*.
- Luciano Castagnola. 2002. Anaphora resolution for question answering.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. // *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- David Dale. 2021a. encodechka-eval. <https://github.com/avidale/encodechka/>.

- David Dale. 2021b. Small and fast BERT for russian language, in russian. <https://habr.com/ru/post/562064/>.
- David Dale. 2022. Rating of russian-language sentence encoders, in russian. <https://habr.com/ru/post/669674/>.
- Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, P 4171–4186.
- Vladimir Dobrovolskii, Mariia Michurina, and Alexandra Ivoylova. 2022. Rucoco: a new russian corpus with coreference annotation.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 878–891, Dublin, Ireland, May. Association for Computational Linguistics.
- Jonas Geiping and Tom Goldstein. 2022. Cramming: Training a language model on a single gpu in one day. *arXiv preprint arXiv:2212.14034*.
- Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. // *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, P 759–765, Istanbul, Turkey, May. European Language Resources Association (ELRA).
- Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. // Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, *Advances in Neural Information Processing Systems*.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Yufang Hou. 2020. Bridging anaphora resolution as question answering. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 1428–1438, Online, July. Association for Computational Linguistics.
- Niklas Jakob and Iryna Gurevych. 2010. Using anaphora resolution to improve opinion target identification in movie reviews. // *Proceedings of the ACL 2010 Conference Short Papers*, P 263–268, Uppsala, Sweden, July. Association for Computational Linguistics.
- Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. Bert for coreference resolution: Baselines and analysis. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, P 5807–5812, 01.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C. Rindfleisch. 2016. Sortal anaphora resolution to enhance relation extraction from biomedical literature. *BMC Bioinformatics*, 17(1), apr.
- Alina Kolesnikova, Yuri Kuratov, Vasily Konovalov, and Mikhail Burtsev. 2022. Knowledge distillation of russian language models with reduction of vocabulary.

- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.
- Monisha Mohan and Jyothisha J. Nair. 2019. Coreference resolution in ambiguous pronouns using bert and svm. // *2019 9th International Symposium on Embedded Computing and System Design (ISED)*, P 1–5.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30:3 – 26, 08.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. // *Third Workshop on Very Large Corpora*.
- Helene Schmolz. 2015. *Anaphora Resolution and Text Retrieval, A Linguistic Analysis of Hypertexts*. De Gruyter, Berlin, München, Boston.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.
2022. Tatoeba, a collection of sentences and translations. <https://tatoeba.org/>.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, P 2214–2218.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // *Advances in Neural Information Processing Systems*, P 5999–6009.
- Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2022. Transformers learn in-context by gradient descent.
- Zili Wang. 2019. MSnet: A BERT-based network for gendered pronoun resolution. // *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, P 89–95, Florence, Italy, August. Association for Computational Linguistics.
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 6442–6454, Online, November. Association for Computational Linguistics.
2022. English-russian parallel corpus (version 1.3), in russian. <https://translate.yandex.ru/corpus>.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation.

Diachronicon: a new resource for the study of Russian constructions in a microdiachronic perspective

Evgeniya Budennaya
HSE University
Institute of Linguistics,
jane.sdrv@gmail.com

Maxim Bazhukov
HSE University
mbazhukov@hse.ru

Liubov Barkova
HSE University
labarkova@hse.ru

Darya Kharlamova
HSE University
dskharlamova@edu.hse.ru

Akhmed Dugrichilov
HSE University
akhmed1797@yandex.ru

Tatiana Reznikova
HSE University
tanja.reznikova@gmail.com

Anastasia Yakovleva
HSE University
avyakovleva@hse.ru

Kristina Litvintseva
HSE University
tinalitvina@gmail.com

Anastasia Andreeva
HSE University
aaandreeva_15@edu.hse.ru

Abstract

The article is devoted to the linguistic characteristics of the database "Diachronicon" and describes the features of the diachronic markup of Russian language constructions, as well as tags specially designed for searching through a diachronic database. A special comment field used in the database is separately justified. In addition, the computer interface of the "Diachronicon" is presented and described.

The developed resource provides extensive opportunities for systematic study of not only specific constructions, but also general mechanisms of idiomatization and grammaticalization. The database allows the researcher to simultaneously compare several separate plots, search through a list of constructions and their characteristics in diachrony, track the history of syntactic and semantic changes and limitations of compatibility of different constructions.

Keywords: database; construction; Russian language; diachrony

DOI: 10.28995/2075-7182-2023-22-1041-1051

Диакроникон: новый ресурс для изучения русских конструкций в микродиакронической перспективе

Евгения Буденная
НИУ ВШЭ
Институт языкознания РАН
jane.sdrv@gmail.com

Максим Бажуков
НИУ ВШЭ
mbazhukov@hse.ru

Любовь Баркова
НИУ ВШЭ
labarkova@hse.ru

Дарья Харламова
НИУ ВШЭ
dskharlamova@edu.hse.ru

Ахмед Дугричилов
НИУ ВШЭ
akhmed1797@yandex.ru

Татьяна Резникова
НИУ ВШЭ
tanja.reznikova@gmail.com

Анастасия Яковлева
НИУ ВШЭ
avyakovleva@hse.ru

Кристина Литвинцева
НИУ ВШЭ
tinalitvina@gmail.com

Анастасия Андреева
НИУ ВШЭ
aaandreeva_15@edu.hse.ru

Аннотация

Статья посвящена лингвистическим характеристикам базы данных «Диахроникон». В статье описаны особенности диахронической разметки конструкций русского языка, а также теги, специально разработанные для поиска по диахронической базе данных. Отдельно обосновывается особое поле комментария, используемое в базе. Кроме того, представлен и описан компьютерный интерфейс «Диахроникона».

Разработанный ресурс предоставляет обширные возможности для системного изучения не только конкретных конструкций, но и общих механизмов идиоматизации и грамматикализации. База данных позволяет исследователю одновременно сравнивать несколько отдельных сюжетов, проводить поиск по списку конструкций и их характеристик в диахронии, отслеживать историю синтаксических и семантических изменений и ограничений сочетаемости разных конструкций.

Ключевые слова: база данных; конструкции; русский язык; диахрония

1 Введение

«Диахроникон» представляет собой проектируемую базу данных конструкций русского языка в микродиахронической перспективе (XVIII–XXI вв.), с подробным систематизированным описанием синтаксических и семантических изменений, происшедших с каждой из конструкций за исследуемый период. Под *конструкциями*, в рамках проекта, подразумеваются идиоматизированные единицы на пересечении лексики и грамматики, состоящие из неизменяемой (якорной) части и слотов, которые могут заполняться различными лексемами [Fillmore et al., 1988; Croft, 2001; Рахилина, 2010; Эндерсен и др., 2020]: N-Gen.PL Cop (*хоть пруд пруди, не дай бог* (NP-Dat) VP-Pfv.Inf, NP-Dat *не улыбается* VP/NP и др. По своей сути, «Диахроникон» углубляет ранее разработанный ресурс «Русский конструктик»¹ [Janda et al. 2018; Эндерсен и др. 2020] в область исторических изменений и является первой базой данных для системного диахронического изучения фразеологизированных конструкций русского языка.

Объем «Диахроникона» значительно превышает аналогичные проекты для других языков, где внимание в первую очередь уделяется синхронному состоянию: так, если синхронные системы конструкций (*конструктики*, в терминологии Ч. Филлмора [Fillmore, 2008]), в последние годы активно изучаются и разрабатываются², то диахронические изменения исследовались более фрагментарно и в первую очередь только для определенного вида конструкций (см., например, работы [Падучева, 2018; Сай, 2010, 2014] для русского языка). С этой точки зрения, «Диахроникон», основанный на данных Национального корпуса русского языка (НКРЯ)³, Генеральном интернет-корпусе русского языка (ГИКРЯ, использовался как дополнительный источник в случае недостаточной выборки по данным НКРЯ)⁴ и сервисе Google Ngram Viewer⁵ (использовался как агрегатор текстов, а не как корпус; к этому источнику обращались только в случае недостаточной выборки, с дальнейшей верификацией даты вхождения), представляет собой самый полный ресурс, позволяющий одновременно сравнивать множество отдельных сюжетов и проводить поиск по заданному списку конструкций и их характеристик в динамике, а также отслеживать историю синтаксических и семантических изменений, которые происходили с конструкцией.

Заделом для базы данных «Диахроникон» послужили корпусные исследования студентов Школы лингвистики НИУ ВШЭ по диахронии идиом, выполненные в рамках курса «Академическое письмо» в 2015–2022 гг. [Klezovich, Golosov, 2018; Клезович, Голосов, Фесенко, 2018; Добрушина, Ахапкина, 2019] (в работах использовались выборки из 300 примеров). На данный момент с помощью коллектива преподавателей и студентов Школы филологических наук и Школы лингвистики НИУ ВШЭ, ведется активная работа по разработке и редактированию системы разметки конструкций, а также созданию компьютерной оболочки базы данных. Объем размеченных единиц составляет более 100 конструкций. Все конструкции были взяты из собрания [Баранов, Добровольский, 2016] и содержат единицы, в той или иной мере обозначающие количество, время, отрицание и дейксис. Критерием отбора конструкций была презентативность их

¹ Проект разработан исследователями Школы лингвистики НИУ ВШЭ (Москва) совместно с Арктическим университетом Тромсё (Норвегия) и расположен по адресу <https://constructicon.ruscorpora.ru/>

² См., например, ресурсы <https://spraakbanken.gu.se/konstruktikon> для шведского языка, <http://jfn.st.hc.keio.ac.jp> для японского, <http://webtool.framenetbr.ufjf.br/index.php/webtool/report/cxn/main> для бразильского португальского, <https://gsw.phil.hhu.de/constructicon/> для немецкого и уже упомянутый «Русский конструктик» для русского языка.

³ <https://ruscorpora.ru/>

⁴ <http://www.webcorporar.ru/>

⁵ books.google.com/ngrams

вариативности и различных синтаксических и семантических изменений. При разметке проводился ручной отбор и верификация примеров с целью устранить ошибки распознавания и, в некоторых случаях, отделить примеры с нужной конструкцией от примеров с похожим на неё сочетанием слов.

Настоящая статья посвящена лингвистическим характеристикам базы данных «Диахроникон». В разделе 2 будут описаны особенности диахронической разметки конструкций. Раздел 3 будет посвящен компьютерной оболочке базы данных. Заключительный раздел 4 суммирует состояние проекта на текущий момент и обозначает его дальнейшие перспективы.

2 Диахроническая разметка

В ходе работы над диахронической разметкой команда проекта учитывала необходимость ее дальнейшей интеграции в систему поиска: вводимые пометы позволят выявлять закономерности в семантической и синтаксической эволюции конструкций, обнаруживать регулярно воспроизводимые модели их развития, в частности, пути конструкционализации свободных сочетаний и де-конструкционализации устойчивых фразем.

2.1 Цели и задачи разметки

При разработке разметки основной упор делался на ее мнемоничность и интерпретируемость для потенциального пользователя.

Строки разметки для одной конструкции соответствовали этапам её развития; *этапом* в базе «Диахроникон» называется состояние конструкции в определённый период, являющееся результатом некоторого изменения.

Для каждого этапа размечались:

1. Индекс предыдущего этапа (*former change*), позволяющий устанавливать отношения преемственности между этапами;
2. Новый вид той части конструкции, которая подверглась изменению (*part of construction changed*). Это позволило отмечать, какие изменения затрагивают якорь конструкции, какие – её изменяемую часть, а какие – синтаксис и семантику всей конструкции в целом;
3. Тип (*type of change*) и подтип изменения (*subtype of change*), благодаря которому возник данный этап, а также уровень языка (*level*), к которому относится это изменение (синтаксис или семантика, *synt vs. sem*);
4. Теги для обозначения морфосинтаксических и семантических категорий (*morphosyntaxs* и *semantags*), характеризующие конструкцию на данном этапе ее развития;
5. Пример первого и последнего вхождения конструкции в НКРЯ; дата первого и последнего вхождений тем не менее условна, НКРЯ постоянно пополняется новыми примерами и работа строится в соответствии с тем, чтобы при обновлении базы употреблений разметка не теряла своей релевантности.
6. Текстовый комментарий, более детально описывающий суть конкретного изменения.

Пример разметки можно увидеть в разделе 2.4.

2.2 Типы и подтипы изменений

Для начального этапа, соответствующего первому зафиксированному употреблению конструкции в НКРЯ, в качестве "типа изменений" вносилась помета "*source*", т.е. источник. Он может быть как композиционным, так и идиоматическим (если идиома, ставшая якорем конструкции, впервые появляется в корпусе именно в некомпозиционном значении).

В рамках проекта каждый новый этап развития конструкции характеризовался изменением на синтаксическом или семантическом уровне. Следует отметить, что в некоторых случаях, по

данным НКРЯ, и синтаксическое, и семантическое изменения фиксировались фактически одновременно, в одном и том же примере⁶.

Для каждого семантического и синтаксического изменения выделялись тип и подтип. Список типов является закрытым, а список подтипов – открытым.

Типы синтаксических изменений размечались в зависимости от того, в какой части конструкции произошло изменение. Так, оно могло произойти:

1. В ядре конструкции (*change in anchor*).

Подтипы этого изменения описывают усиление или ослабление синтаксического и семантического единства якоря, а также изменения в составе якоря.

Так, при добавлении компонента (*adding a component*) в якорь конструкции “как не бывало” получает возможность добавляться усилительная частица *и*:

(1) *Потом вдруг как и не бывало литературных обозрений!* [В. Г. Белинский. Русская литература в 1842 году (1843)]⁷

2. Во внутреннем синтаксисе конструкции (*change in inner syntax*).

Под внутренним синтаксисом конструкции подразумеваются те возможные наборы составляющих, которые зависят от наиболее вложенной составляющей якоря конструкции. В большинстве случаев они представляют собой слоты в изменяемой части конструкции. Подтипы этого изменения описывают появление и исчезновение зависимых от якоря или изменения в наборе синтаксических типов существующих зависимых. Так, при подтипе *change in government: clause* слот в конструкции с якорем “без понятия” со значением информации раньше мог заполняться только предложной группой, а теперь заполняется и финитной клаузой:

(2) *И вновь нас зомбируют, только уже в другом направлении, хотя без понятия как это влияет на верующих* [ГИКРЯ]

3. Во внешнем синтаксисе конструкции (*change in outer syntax*).

Под внешним синтаксисом конструкции подразумеваются те возможные наборы составляющих, в которые может быть вложена наименее вложенная часть якоря конструкции.

Подтипы этого изменения описывают изменения в полярности конструкции и в наборе синтаксических ролей, в которых она может выступать.

Так, при изменении поляризованности (*change in polarity*) конструкция “без царя в голове”, которая раньше была положительно поляризованной, стала употребляться и в отрицательных предложениях:

(3) *Нужны люди, а я человек бывалый, опытный и не без царя в голове, чего еще?* [М. Е. Салтыков-Щедрин. Пестрые письма (1884-1886)]

Типы семантических изменений выделялись исходя из более широкого набора факторов. Так, были выделены:

1. Новые идиоматичные употребления конструкций (*new idiomatic use*).

Подтипы этого изменения описывают семантические переходы: метафора, метонимия, ребрендинг, а также расширение и сужение значения.

⁶ Строго говоря, даже в этом случае два изменения происходят последовательно, так как одно из них является причиной другого (например, ребрендинг является результатом расширения семантической сочетаемости (Rakhilina, et al. 2010)). Тем не менее, на данный момент в разметке отражены только те факты о развитии конструкций, которые фиксируются в выдаче НКРЯ - интерпретация же этих фактов пока полностью предоставляется пользователю. Поэтому пока что мы размечаем несколько изменений, впервые зафиксированных в одном примере, как произошедшие одновременно. Установление подлинного порядка произошедших в таких случаях изменений остается задачей пока что не решенной, но входящей в перспективы развития проекта.

⁷ Этот и все дальнейшие примеры взяты из НКРЯ.

Так, при подтипе *widening* (расширение) конструкция *не дай бог* ранее содержала интенсивность и негативную оценку; а на новом этапе – только интенсивность:

(4) — *Я, кажись, ни за что бы не выдержал, потому я до лошадей, не дай бог, какой охотник! Ежели увижу где хорошего коня, у меня душа мрет, а у самого так руки и чешутся украсть его.* [Ф. Ф. Тютчев. На скалах и долинах Дагестана (1903)]

2. Изменения семантической сочетаемости конструкции (*change in semantic compatibility*). Подтипы этого изменения описывают изменения в наборе семантических типов зависимых конструкции или вершин конструкции. Так, при подтипе *extension: new type of NP dependent* (расширение множества имён-зависимых, с которыми может сочетаться выражение) в конструкции *NP-Gen не напасешься* в качестве аргумента используется слово, не обозначающее множество, хотя ранее это было невозможно:

(5) *На всех проходящих и Маслова не напасешься.* [И. Т. Кокорев. Савушка (1847)]

3. Изменения в прагматике конструкции (*change in pragmatics*). Подтипы этого изменения описывают прагматикализацию и депрагматикализацию дискурсивных формул, а также изменения в иллокутивной цели употребления конструкции. Так, при подтипе *depragmaticalization of a routine* конструкция *ни-ни* ранее употреблялась как дискурсивная формула, а теперь может быть частью предложения:

(6) — *Нет, брат! играть с тобой еще можно, но позволять тебе карты сдавать — ни-ни!* [М. Е. Салтыков-Щедрин. Помпадур и помпадурши (1863-1874)]

2.3 Теги

Особо следует сказать об используемой в разметке системе тегов: она была введена с целью найти в базе данных этапы, имеющие отношение к тем или иным морфологическим, синтаксическим и семантическим явлениям. В связи с этим при разметке использовались два набора тегов: морфосинтаксические и семантические.

Инвентарь **морфосинтаксических** тегов включает в себя следующие подгруппы:

1. Частеречные теги (Noun, Adj, Verb и т.д.).
2. Теги для грамматических категорий глаголов (например, Infinitive, Past Tense, Subjunctive и т.д.), имен и местоимений (Animacy, Singular, Dative и т.д.).
3. Теги для синтаксических категорий составляющих (например, Sentential Actant, Argument, Discontinuity – разрывность и т.д.).
4. Теги для поляризованных единиц (NPI, PPI).

Большая часть **семантических** тегов заимствована из Конструктикона (см. описание в [Endersen et al. 2020]); среди них – тег *Maleficiary*, обозначающий участника, которому действие приносит вред. Использование данного тега в диахронической разметке можно увидеть в разделе 2.4.

Однако некоторые семантические теги были разработаны для нужд данного проекта. Среди них выделяются следующие группы:

1. Теги для явлений, связанных с квантификацией (Large Quantity, Small Quantity, Mass Noun).
2. Теги, обозначающие стиль (Obscenity).
3. Теги для режимов дискурса (Dialogue, Narrative).
4. Теги для семантического класса всей конструкции или её элемента (в частности, тег *Negation*, обозначающий отрицание какой-либо пропозиции в семантике конструкции).

2.4 Пример разметки: *NP-Dat не улыбаться NP*⁸

change	former change	part of construction changed	level	type of change	subtype of change	semantags	morphosyntags
1		NP-Dat не улыбаться NP		source	compositional source	Negation	
2	1	NP	sem	change in semantic compatibility	extension: new type of NP dependent	Negation, Maleficiary	
3	1	entire construction	sem	new idiomatic use	rebranding	Negation, Maleficiary	Noun
4	2, 3	не Adv улыбаться	synt	change in anchor	deidiomatization	Negation, Maleficiary	Discontinuity
5	4	не улыбаться CI	synt	change in inner syntax	change in government: clause	Negation, Maleficiary	Verb, Infinitive
6	5	NP-Dat улыбаться NP	synt	change in outer syntax	change in polarity	Maleficiary	NPI

Таблица 1: Фрагмент разметки выражения *NP-Dat не улыбаться NP*

Описанную выше разметку можно проследить на примере выражения *не улыбается* (см. Таблицу 1). За свою историю оно претерпело несколько существенных изменений: изначально мы видели, что конструкция была композиционной (этап 1) и имела несколько актантов: того, кто улыбается, и того, кому улыбаются.

(7) *Да ведь и Иван Самойлыч молодой человек, а он уж и хил, и желт, и согнут, да и актриса ему не улыбается...* [М. Е. Салтыков-Щедрин. Невинные рассказы / Запутанное дело (1848-1863)]

Затем сочетаемость расширилась (этап 2): зафиксированы употребления, в котором NP обозначает кого-либо, кто не может улыбаться в буквальном смысле этого слова⁹.

(8) *Им не улыбается мысль, что лучше быть первым в деревне, нежели вторым в Риме; им не приходит в голову даже то совершенно естественное предположение, что, сделавшись участником столичного движения, они не только не внесут никакой новой струи, но сами утонут в департаментском соре.* [М. Е. Салтыков-Щедрин. Письма о провинции (1868-1870)]

Этап 1 стал причиной ребрендинга (этап 3): наблюдаем вхождение, где выражение стало идиомой и субъект начал обозначать что-либо неприятное для объекта:

(9) *Министр финансов хотел бы созвать совещание с участием гг. Гирса и Абазы, и перспектива необходимости вернуться в город до своего ближайшего доклада, т. е. до 4 июня, вовсе не улыбается моему дорогому начальнику.* [В. Н. Ламсдорф. Дневник (1891)]¹⁰

Далее, в НКРЯ фиксируется вариант, где между частями выражения теперь может ставиться модификатор (этап 4, развивающийся из двух предыдущих этапов - см. таблицу 1).

⁸ В разметке использован инфинитив глагола “улыбаться”, чтобы отразить вариативность форм глагола, которые могут занимать данную позицию.

⁹ Возможное влияние метафорического употребления “улыбаться” в значении абстрактных существей (например, удача мне (не) улыбается). Тем не менее, изменения, происходящие с конструкцией “NP-Dat не улыбаться NP” (и отмеченные в Таблице 1) не влияют на вариант с метафорическим употреблением.

¹⁰ Данное вхождение не является первым примером употребления конструкции в НКРЯ; первый пример этапа 3 совпадает с первым примером этапа 2.

(10) *Хотя мне эта комбинация не особенно улыбается, но, вероятно, придется прибегнуть к ней.* [В. О. Ковалевский, С. В. Ковалевская. Письмо А. О. Ковалевскому (1874)]

Подобный тип изменений, связанных с тем, что якорь получает возможность изменяться там, где раньше он был неизменяемым, то есть теряет одно или несколько свойств идиомы¹¹, мы назвали *деидиоматизацией*.

На следующем этапе у конструкции фиксируется возможность присоединять предикаты (этап 5):

(11) [Телемахов Прокопий Евсеевич (муж, профессор)] *Мне вовсе не улыбается, Валентин Николаевич, быть вестником несчастья и врываться в вашу... идиллию.* [Л. Н. Андреев. Профессор Сторицын. Драма в четырех действиях (1912)]

Последний этап заключается в изменении якоря конструкции (этап 6): фиксируются варианты, где *не* перестает быть постоянной частью якоря и может заменяться на другие лицензоры NPI:

(12) *Это мне не улыбается — да и вряд ли кому вообще улыбается.* [Г. С. Эфрон. Дневники. Т. 1. 1941 (1941)]

Общую схему микродиакронического развития конструкции *не улыбается* можно видеть на Рис. 1.

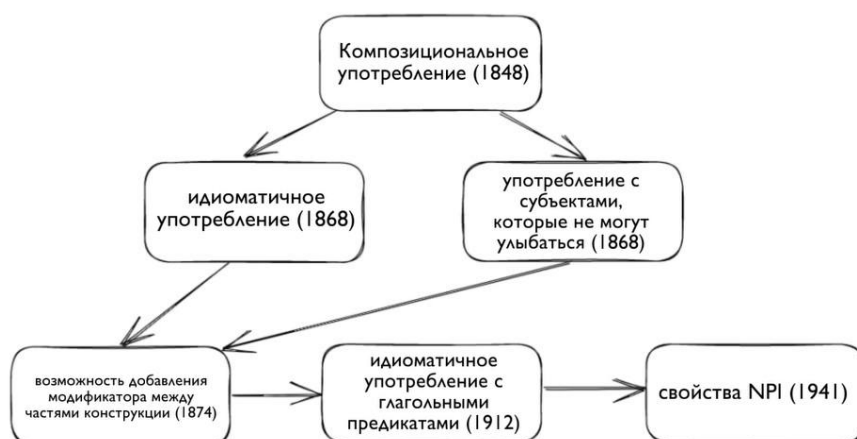


Рис. 1: Этапы микродиакронического развития конструкции *не улыбается* по данным НКРЯ

3 База данных с веб-интерфейсом

В рамках проекта предполагается, что база данных «Диакроникон» будет доступна онлайн. На данный момент для неё разрабатываются реляционная архитектура (в тестовой версии она представляет собой базу данных SQLite) и веб-сервер, который предоставит возможность просмотра данных в удобном виде и поиска на основе описанной выше разметки. Сервер создан при помощи фреймворка Flask для языка Python. Ниже будет кратко описана архитектура базы данных, а также реализованные и планируемые способы поиска и представления данных.

Для диахронических изменений в конструкциях в базу данных интегрирована таблица изменений, с разметкой описанной в Разделе 2. Помимо этого, в базе данных представлена разметка современного этапа **конструкции**, где вместо типов изменения представлены параметры «современное значение» (*contemporary meaning*) и «синтаксическая функция якоря конструкции» (*synt_function_of_anchor*). Кроме того, здесь добавлена схема (глоссирование) якорной части

¹¹ Например, в якорь конструкции порядок слов из фиксированного становится свободным; лексический состав якоря при этом обычно остаётся неизменным.

конструкции, и альтернативные формулы для современной конструкции (*variation*). Предусмотрены указатели для связи с Русским Конструктиконом.

Кроме того, отдельно размечаются синтаксические и семантические **ограничения** на конструкцию в целом и на ее элементы, которые данным этапе имеют вид текстовых комментариев. При этом сохранены метаданные и **общая информация** о конструкции: в частности, в базе данных представлена выборка, на основе которой получена разметка, и указаны разметчики конструкции.

Все эти данные подробно отражены на страницах отдельных конструкций, см. Рис. 2.



Рис. 2: Страница конструкции *пруд пруди*

На Рис.2 представлена базовая информация о конструкции, включающая большую часть перечисленных выше параметров: значение, якорь и альтернативные формулы. Здесь также доступен линейный график периодов в истории конструкции, а в нижней части страницы изменения отражены в более подробном табличном виде (Рис. 3). Наконец, приведена информация о разметчиках конструкции и выборке.

стадия	уровень	тип изменения	первое	последнее	первое вхождение	последнее вхождение	комментарий
			вхождение	вхождение			
(дата)	(дата)						
V N-Acc (NP-Ins)	synt	source	1850	1886	Видеть она, что враги хотя́ть прудить ровъ и стрѣляютъ по стѣнамъ. [А. В. Дружинин. Письма иногороднего подписчика о русской журналистике (1850)]	<...> убирали хлеб и сено (на монастырских запашках), прудили пруды, сады оплетали, на невод ходили, хлебы пекли, пиво варили, лен пряли». [М. Ф. Владимирский-Буданов. Обзор истории русского права (1886)]	Фразеологизм происходит от сочетания форм глагола «прудить» (ставить плотину поперёк водоёма) и существительного «пруд». Материал, из которого ставится запруда, выражался творительным падежом.
physical action	sem	source	1850	1886	Видеть она, что враги хотя́ть прудить ровъ и стрѣляютъ по стѣнамъ. [А. В. Дружинин. Письма иногороднего подписчика о русской журналистике (1850)]	<...> убирали хлеб и сено (на монастырских запашках), прудили пруды, сады оплетали, на невод ходили, хлебы пекли, пиво варили, лен пряли». [М. Ф. Владимирский-Буданов.	

Рис. 3: Фрагмент таблицы с изменениями на странице конструкции

Общая информация	Якорь	Изменения конструкции	Ограничения
id	Синт. функция якоря	Формула в этот период	Элемент
Формула	Схема якоря	Этап в истории конструкции	Выберите тип ограничения
Значение	Якорь (рус.)	Синтаксическое	Ограничение
<input type="checkbox"/> Есть в конструктиконе	Якорь (англ.)	Подтипы изменений	
id (Конструктикон)		Длительность это (≤, ≥ или =)	
		Длительность периода	
		Первое вхождение в таком виде	
		Последнее вхождение в таком виде	

Рис. 4: Поддерживаемые поля для поиска

По большинству указанных выше параметров возможен поиск (Рис. 4). Поля «значение» и «синт. функция якоря», «подтип изменения» допускают поиск по одному из предусмотренных в разметке значений. Поле «подтип изменения» позволяет находить конструкции по диахронической разметке, описанной в разделе 2.2. В совокупности с другими полями это учитывает большую часть вариативности конструкций.

Возможен поиск и по формуле. Он учитывает конструкцию, а также ее альтернативные формулы, варианты, выявленные при разметке, и допускает использование символа-джокера *. Ниже приведен пример запроса (12). При поиске по формуле с именной группой в любом падеже, что выражено символом *, на выходе выдаются конструкции с отдельными «подрезультатами» (данная особенность связана с тем, что часть конструкций имеют альтернативные формулы, которые тоже содержат именную группу).

(12) [Запрос] Формула: пр*
[Фрагмент выдачи]:

Результатов по искомым параметрам: 21 (конструкций — 14)

1. NP Cop не что (иное) как NP:
 - 1 NP Cop не что (иное) как NP
 - 2 NP Cop не кто (иной) как NP

Наконец, раздел “изменения конструкции” позволяет найти все изменения с определёнными характеристиками: формулой заданного вида, типа изменений (синтаксических vs. семантических), начавшихся или закончившихся в определённом году, либо длившихся заданный период времени.

Веб-интерфейс пока поддерживает не всю описанную в предыдущем разделе разметку. Тем не менее он уже сейчас позволяет отвечать на разные вопросы. Так, интересны могут быть «стабильные» этапы, на которых конструкции задерживаются долго без изменений. Найти такие этапы можно удобным запросом в (13), его результат представлен на Рис.5.

(13) [Запрос] Длительность периода это: Минимум (\geq)
(&)Длительность периода: 150

Результатов по искомым параметрам: 24 (конструкций — 16)

- 1 **N-Gen.PI Cop (хоть) пруд пруди**
1 1831 — 1984
synt : adding concessive particle *NP-Ins хоть N-Acc V-Imper*
- 2 **в точности PronDem**
2 1716 — 2019
synt : *source* *VP в точности*
- 3 **Prep N-Dat.Sg не по адресу**
3 1841 — 2021
sem : metaphor *negative assessment*
- 4 **NP-Gen не*густо**
4 1795 — 2017
sem : *source* *entire construction*
- 5 1830 — 2018
sem : metaphor *entire construction*

Рис. 5: Конструкции с этапами изменений длиной более 150 лет, найденные через веб-интерфейс

Так, видно, например, что для конструкций *в точности*, *пруд пруди* долго фиксируется употребление в исходном синтаксически или семантически виде (*source*).

Логичным расширением функционала видится возможность поиска по нескольким изменениям и обобщающих поисков. Так, можно найти конструкции, с которыми произошло меньше 2 изменений или конструкции, с которыми произошло больше 4 семантических изменений за период, определяемый конкретными годами или своей длительностью. Возможность выделить период на основе конкретных годов важна, чтобы выявлять года, когда среди разных конструкций в целом изменений происходит много и года, когда наоборот, конструкции устойчивы. Такие вопросы могут задаваться как лингвистами, так и филологами. Эти функции активно разрабатываются.

4 Заключение

На данный момент в рамках проекта “Диакроникон” разработана детальная и многоуровневая разметка диахронического развития конструкций, позволяющая обобщить и систематизировать изменения их синтаксиса и семантики, а также создан веб-интерфейс для поиска конструкций в базе данных и наглядного отображения изменений и свойств конструкций.

Разработанный ресурс предоставляет обширные возможности для системного изучения не только конкретных конструкций, но и общих механизмов идиоматизации и грамматикализации. База данных позволяет исследователю одновременно сравнивать несколько отдельных сюжетов, проводить поиск по списку конструкций и их характеристик в диахронии, отслеживать историю синтаксических и семантических изменений и ограничений сочетаемости разных конструкций. Ближайшей задачей проекта является пополнение инвентаря конструкций в базе и уточнение разметки имеющихся конструкций, а также дальнейшее совершенствование методов разметки и классификации диахронических изменений. Что касается развития веб-интерфейса, планируется разработка функции поиска по нескольким изменениям, которая позволит исследователям запрашивать в базе конструкции, с которыми произошло определенное количество изменений за конкретный период времени. Такое расширение функционала сделает ресурс особенно ценным инструментом для системного изучения языковых изменений.

Опыт работы над Диахрониконом может быть полезен также для создания аналогичных ресурсов для других языков и изучения механизмов конструкционализации в типологической перспективе.

Acknowledgements

Статья подготовлена по результатам проекта «База данных по историческим изменениям русских конструкций «Diachronicon»» при поддержке Фонда академического развития ФГН НИУ «Высшая школа экономики» в 2022- 2024 году.

References

- [1] Croft W. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press, 2001.
- [2] Fillmore C. J. *Border Conflicts: FrameNet Meets Construction Grammar*. // *Proceedings of the XIII EURALEX International Congress 2008*. — Barcelona, Spain, 2018. — P. 49–68.
- [3] Fillmore Ch. J., Kay P., O'Connor M. C. *Regularity and idiomatity in grammatical constructions: The case of let alone* // *Language* 1988. — Vol. 64(3). — P. 501–538.
- [4] Janda L. A., Lyashevskaya O., Nessel T., Rakhilina E., Tyers F. M. *A constructicon for Russian: Filling in the gaps*. // *Constructicography: Constructicon development across languages*. Amsterdam: John Benjamins, 2018. — P. 165–181.
- [5] Klezovich A.G., Golosov F.V. (2018a), *The Database “Phrasal Diachronicon”: constructional change in Russian expressions with quantitative semantics* // *Proceedings of Constructional semantics: Cognitive, functional and typological approaches 2018*. — Helsinki, Finland, 2018. — Poster session.
- [6] Баранов А. Н., Добровольский Д. О. *Основы фразеологии (краткий курс)*. — Москва: Флинта Наука, 2016.
- [7] Добрушина Е. Р., Ахапкина Я. Э. *Пограничный русский язык: как рождаются экспрессивные кванторные выражения*. — СПб.: Алетейя, 2019.
- [8] Клезович А. Г., Голосов Ф. В., Фесенко В. П. *База исторической фразеологии (Phrasal Diachronicon): отражение изменений в функционировании лексических единиц с семантикой количества*. // *Симпозиум по Русскому Конструктикону в филиале НИУ ВШЭ в Вороново, 17-23 марта 2023*. — М.: 2018.
- [9] Падучева Е. В. *Из наблюдений над языком Л. Толстого (к вопросу о малых диахронических сдвигах)*. // *Вопросы языкознания*. — 2018 — Вып. 5. — С. 49–63.
- [10] Рахилина Е. В. (ред.) *Лингвистика конструкций*. — М.: Азбуковник, 2010.
- [11] Рахилина Е. В., Резникова Т. И., Карпова О. С. *Семантические переходы в атрибутивных конструкциях: метафора, метонимия и ребрендинг*. — М.: Азбуковник, 2010. — С. 396–455.
- [12] Сай С. С. *Динамика развития обстоятельств времени со значением предшествования на интервал в русском языке*. // *Избыточность в грамматическом строе языка*. [Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН. — 2010. — Вып. 6(2)]. / Воейкова М. Д. (ред.). — СПб.: Наука, 2010. — С. 131–183.
- [13] Сай С.С. *Лексические механизмы грамматического дрейфа: конструкции с дативным субъектом в русском языке XVIII – XXI веков* // *Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН*. / Казанский Н. Н. (отв. ред.). СПб.: Наука, 2014. — Т. 10(3). — С. 568–610.
- [14] Эндерсен А., Жукова В., Мордашова Д., Рахилина Е., Ляшевская О. *Русский конструктикон: новый лингвистический ресурс, его устройство и специфика*. // *Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2020»*. — М.: РГГУ, 2020. — Т. 19. — С. 226–241.

Who is answering to whom? Modeling reply-to relationships in Russian asynchronous chats

Igor Buyanov
FRC CSC RAS
Moscow, Russia

Darya Yaskova
MTS AI
Moscow, Russia

Ilia Sochenkov
FRC CSC RAS
Moscow, Russia

buyanov.igor.o@yandex.ru dary.yaskova@gmail.com

sochenkov@isa.ru

Abstract

The study highlights the asynchronous nature of modern group chats and related problems such as retrieving relevant information on the asked question and understanding reply-to relationships. In this work, we formalize the reply recovery task as a building block toward solving described problems. Using simple heuristics, we try to apply the result reply recovery model to a thread reconstruction problem. As a result, we show that modern pre-trained models such as BERT show great results on the task of reply recovery compared to more simple models, though it cannot be applied to thread reconstruction with just simple heuristics. In addition, experiments have shown that model performance depends on the chat domain. We open-sourced a model that can automatically predict which message the particular reply responds to and provide a representative Russian dataset that we built from Telegram chats of different domains. We also provide a test set for a thread reconstruction task.¹

Keywords: Russian language, asynchronous chats, group discussion, thread reconstruction, reply recovery, BERT

DOI: 10.28995/2075-7182-2023-22-1052-1060

Кто кому отвечает? Моделирование взаимоотношений сообщений в асинхронных чатах на русском языке

Игорь Буянов
ФИЦ ИУ РАН
Москва, Россия

Дарья Яськова
MTS AI
Москва, Россия

Илья Соченков
ФИЦ ИУ РАН
Москва, Россия

buyanov.igor.o@yandex.ru dary.yaskova@gmail.com sochenkov@isa.ru

Аннотация

В исследовании поднимается тема асинхронной природы современных групповых чатов и связанных этим проблем, таких как получение соответствующей информации и понимание того, кто кому отвечает. В этой работе мы формализуем задачу восстановления ответов как базовый блок в решении описанных проблем. Используя простые эвристики, мы стараемся применить полученную модель восстановления ответа к проблеме реконструкции тредов сообщений. В результате мы показали, что современные предварительно обученные модели, такие как BERT, показывают отличные результаты на задаче восстановления ответов по сравнению с более простыми моделями. Тем не менее тесты показали, что использование модели с простыми эвристиками не дают хороших результатов на задаче реконструкции тредов. Кроме того, эксперименты показали, что производительность модели зависит от домена чата. Мы опубликовали модель и набор данных на русском языке, который мы создали из чатов Telegram из разными доменов, для задачи восстановления ответов. Мы также опубликовали тестовый набор для задачи реконструкции тредов.

Ключевые слова: русский язык, асинхронный чат, обсуждения в группе, реконструкция тредов, восстановление ответов, BERT

¹https://github.com/Astromis/research/tree/master/reply_recovery

1 Introduction

The spread and availability of the Internet allow people to be in touch with each other, regardless of their location. Via Web, one can ask for help about a particular problem or discuss any topic with other people. To make this process more ordered, people start to form communities and online forums that dedicate to some declared theme.

With further development of web technologies and, in particular, smartphones, several quick message applications gained popularity such as Telegram and WhatsApp. The fact that smartphones are always near hand coupled with a live time regime of receiving and sending messages allows for speeding up online conversations. Although at first these applications were intended to be for personal conversations, the group chat feature was also introduced, thus opening doors for group discussions.

However, the speed of these chats that allows the users to quickly ask and get help also is a drawback as useful information is flushed away. Another issue that information seekers could encounter is difficulty to track messages that relate to a topic of interest because in such chats users can discuss several topics simultaneously. Both of these drawbacks lead to a tangle, e.g. a situation when a group chat newcomer asks about a topic that was already discussed. Usually, the only thing he gets is a response about this fact, but not at least a message from which the topic begins. On the other hand, if you want to read messages that are in one dialogue, you have to swipe up unrelated messages. Although some chat applications such as Telegram have a feature to show messages that are connected via replies, users may just not use the "reply to" function that breaks any sense of this feature.

From the scientific perspective, the study of online chats can help us to pursue our understanding of discourse and dialogue phenomenon, as these chats are a rich source of how people are communicating with each other using text and other modalities. The fact that these chats can have several topic discussions simultaneously with intertwined participants makes it a great challenge to automatically analyze them, while humans easily keep track of discussed topics they follow. The investigation of these chats, for example, can help us to develop dialog agents that can actively operate in group conversations rather than in personal ones.

In this work, we investigate the simple relation between messages in these chats, specifically, what message is replying to another. We will refer to this task as reply recovery. We present a model that can automatically predict such relations and try to generalize it to the more complex task of reconstructing separate threads in chats. Our contribution can be summarized as follows:

1. We formalize a task of reply recovery, and provide a representative Russian dataset that we built from Telegram chats of different domains. We also study several methods to solve this task.
2. We create a small benchmark from chats where all "reply to" relations were annotated. We propose a greedy wrapper upon reply recovery model to test its performance in a thread reconstruction setting.

2 Related work

The first work that defined the problem of thread structure reconstruction is (Wang et al., 2008). They used a dataset constructed from forum conversations. The authors proposed a simple unsupervised method that relies on a graph-based text representation. The graph is constructed in a way that all messages are connected to all previous messages, with edge weight calculated as TF-IDF between message texts. Having a complete edge matrix, authors apply a threshold to filter out weak connections. The authors also propose some penalizing strategies based on data observations.

This work started productive research in this direction. The supervised methods appeared that works with emails (Dehghani et al., 2013) and blog comments (Balali et al., 2013). In the work (Louis and Cohen, 2015) authors pointed out a topic as an aspect of thread structure that represents which theme is discussed. They segment messages into several topics and model their treelike structure with different types of context-free grammar.

In the work (Nguyen et al., 2017) the authors adapt a well-known coherence model based on the entity grid in a way that can operate in asynchronous conversation. They pointed out that traditional coherence models can't be applied to this task as they assume a chronological, synchronous flow. On the top of the entity grid, they apply a convolutional network trained with a pairwise ranking loss on choosing the valid

thread tree.

In the paper (Guo et al., 2018) authors for the first time considered the task of predicting "reply-to" relations. They provide two LSTM-based models. The first model use just messaged words, while the second one operates on a sentence level, using the previous model as a core. In addition to forum data, they use data from quick messages systems, in particular WeChat, which is similar to WhatsApp.

In the most recent work (Ji et al., 2021) authors propose a complex solution that tries to capture latent factors such as topic consistency and discourse dependency. To do that, they combine two modules. The first one jointly learns latent topics and discourse, while the second one makes actual predictions about relations.

3 Method

In this work, we consider the task of determining whether one message can be a reply to another. We call this task reply recovery. On top of that, we test simple methods for thread reconstruction. We describe them in separate chapters.

3.1 Reply recovery task

We will use notation and definitions from (Guo et al., 2018). We are given a group chat corpus which is an ordered list of messages $M = \{m_1, m_2, \dots, m_N\}$ where N is the total number of messages. The messages come from different users that can participate in several conversations at once. That leads to a situation when messages relating to a particular conversation come asynchronously, so it becomes unclear to say what a particular message responds to. We say $m_i \prec m_j$ if m_j has a "reply to" relation with m_i for $\forall m_i, m_j \in M, m_i \neq m_j$. The task is a binary classification, the objective of which is to predict whether the pair of messages $m_i, m_j, j > i$ has a "reply to" relation.

3.2 Thread reconstruction

Let's assume that messages M contain I threads that we define as a subset of messages $T \in M$ that are related to each other by meaning and thus form a coherent dialogue between users on some topic. The messages from different threads are intertwined as they come asynchronously. So the task of thread reconstruction is to divide the message list M into I threads. We hypothesize that by having a complete map of "reply to" relations, we can get these I threads.

4 Dataset collection

We found out that real chats can be used as a natural source of the data for this task, as people do use the "reply to" feature, explicitly declaring to whom they answer. That could be used as positive labeling. On the other side, it might be thought that the message with its reply should be read coherently. The usual way to make a sequence of text incoherent is to replace some parts with a random sample. We do the same, combining messages randomly.

As a message source, we use a publicly available list of opened Telegram chats. Using the Telegram API, we gather messages from these chats. To ensure the dataset diversity, we manually picked chats with different topics: two chats of women who recently become mothers and who lived in Bali, a chat about football, two chats on IT topic, one house tenant chat, and two teen chats about gaming and suicidal game "Siniy kit" ("Blue whale"). We choose these chats randomly from what we had scraped, except the suicidal game chat, as it is very hard to collect data dedicated to a particular topic. Another criterion was a high ratio of "reply to" messages, although the resulting message counts come with a high disbalance.

4.1 Chat data analysis and preparation

As we have our data, we do a basic analysis to fulfill our interests. First of all, we check the offset between messages and their replies (Fig. 1, left). It can be seen that an offset equal to one (the right after the reply message) is the most common case, and on the other hand, the offset with a distance of 14 is almost diminished. A similar distribution is presented in paper (Guo et al., 2018).

We also examine the number of trees and chains that can be formed from existing "reply to" relations

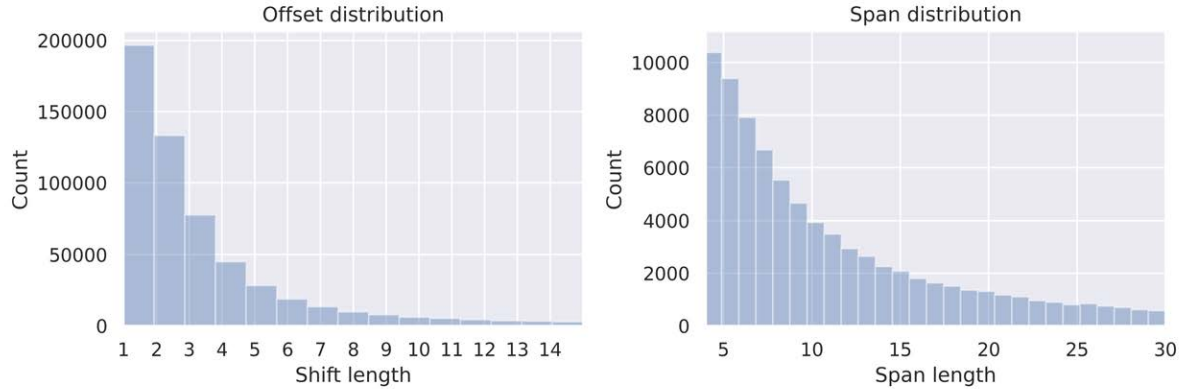


Figure 1: Distributions of message offsets (left) and spans (right) in chats

and related distributions. In the whole dataset, we find 173064 chains and 62067 trees. It’s not surprising the superior number of chains have only two nodes, whereas trees have three nodes. At last, we examine the distribution of the distance between the start node and the end node in threads, which we refer to as a span distribution (Fig. 1, right). We exclude threads with less than three nodes as a trivial variant. From this distribution, we can roughly say that most threads lie between 30 messages.

To construct our reply-to dataset, we collect all pairs of messages and their replies. Next, we filter pairs where a message or reply length is less than three symbols and more than 1000. We also filter out pairs that contain more than one Latin symbol pair, as we will use a pretrained model that is not multilingual. We assume these pairs to be positive. As we mentioned above, to make negative pairs, we randomly combine messages with the restriction that they must be from the same chat. We hypothesize that such pairs would be harder than if we pick them from different domains. In total, we have 894997 pairs that we divide in train and evaluation set in proportion 89 and 10 percent. We leave 1 percent for a test set.

4.2 Test set annotation for reply recovery

During the analysis, we figure out that not all explicitly marked message pairs are coherent, and surprisingly not all randomly combined messages are incoherent. To be sure of the good quality of the test set, we decided to annotate it with crowdworkers at the Yandex.Toloka platform that is widely used in post-Soviet country space. For example, it was used to create a large-scale dataset of crowdsourced audio transcriptions named CrowdSpeech (Pavlichenko et al., 2021)

Each task consists of two texts representing the message and its potential reply. The crowdworkers were asked a question if a second text can be a reply to the first or not. The project setup included a restriction of using only the top 10 percent best-rated tolokors and annotation overlap equals to three. The negative and positive examples are annotated separately. While we are aware that such a setting could make an annotator biased, we didn’t figure out how to preserve identification information on the platform to be able to map examples with a preannotation. By the end of the annotation process, we got inter-annotator agreement by Krippendorff’s alpha (Krippendorff, 2011) for negative being equal to 0.182 and for positive being equal to 0.280. We think that such a low score relates to the noisy nature of the crowdsource, as the task was pretty simple. In addition, we didn’t perform an annotator training procedure. Actually, we rely on the Dawid-Skene model (Dawid and Skene, 1979) as a method of obtaining true labels from noisy crowd labels. After completing a pilot start, the investigation of the model confidence result distribution shows that the sufficient part of examples has a very high confidence score, although the distribution has a notable tail. We establish a threshold of 80 percent, above which we consider the example annotation to be reliable. Another finding is that there are some negative examples out of a positive set. Some of them have a strong confidence that’s of high interest. A similar picture is observed with the negative set.

Having all annotations completed, we manually validate messages that were supposed to be negative but were annotated as positive and vice versa. We also examine messages that have weak confidence. We can highlight the reason why that occurs: foreign language, domain misunderstanding, and, in particular, meme phrases, uninformative messages, and, at last, simple mistakes.

In the end, we got 4693 positive and 3997 negative messages with balanced domain distribution, although one category has slightly more messages due to data management mistakes.

4.3 Test set annotation for thread reconstruction

As we intended to use our model to thread reconstruction, we create a small test that consists of chats from the above-mentioned topics. We randomly pick 10 slices from each topic with 100 messages in range. We ensure that slices do not intersect and that all messages contain text. As the task is much harder than just deciding whether two texts are related and, to the best of our investigation, the technical restriction of the platform, we hire two annotators to manually label all connections in each chat using the Label Studio platform (Tkachenko et al., 2020). The main criterion of messages being connected is coherence between them. We conduct a workshop where annotators were instructed to keep in mind that related messages should have a meaning if we would discard all other messages, and also that related messages have common words. It's important to note that unlike in some other works, we assume that the response has only one head. However, some messages have a clear sense of addressing many chat participants. These messages were asked to mark as self-connected. It contradicts the work (Guo et al., 2018) where messages are self-connected if they have no replies. In our schema, such messages just didn't have annotation.

Another notable phenomenon is several messages followed by one another from one user. The annotators were instructed to connect these messages, except when some message replies to a message outside the monologue. In the end, we got 89 annotated dialog slices with 100 messages in each sample. It has to be noticed that the random manual inspection of the dialogues shows that the annotation is quite noisy.

5 Models description

5.1 Reply recovery

As a starting point, we use **Logistic regression** on top of concatenated vectors of texts within the pair. The vector representation is a term-document matrix. The hypothesis behind the use of such a simple model is the fact that related messages often reuse words that represent a subject or object. This is the ground of entity grid representation (Barzilay and Lapata, 2008).

Following the (Guo et al., 2018) we train **LSTM** model on word level, but we augment it with self-attention mechanism (Vaswani et al., 2017). Given the model of concatenated text pairs, we expect that the attention mechanism allows the model to learn better discourse dependencies. We exclude the hierarchical sentence level variant from those words, as it didn't give notable gain.

The defined task aligns closely with the next sentence prediction (NSP) loss that is used to train the BERT model (Devlin et al., 2019). We can assume that messages are often about one sentence, though we assume that integrating more than one sentence is an applicable strategy.

We use **Conversational RuBERT²** (ConvBERT) that is fine-tuned on social media texts RuBERT (Kuratov and Arkhipov, 2019). It's important that it has a vocabulary based on this data. In turn, we fine-tuned the Conversational RuBERT on our data with NSP loss with 3 epochs and 1e-5 learning rate. Thinking about how the model will be used in a thread reconstitution task and the fact that two related messages should be coherent, we try to use **Siamese architecture** (Neculoiu et al., 2016) with different bases. As a reminder, the Siamese network architecture consists of two identical basic networks with shared parameters. In the usual setting, they consume two objects and the main goal is to distinguish the difference between them. Usually, it trains with a contrastive or a triplet loss. In our work, we use LSTM and our fine-tuned ConvBERT as the basic network. The latter architecture is also known as SentenceBERT or SBERT (Reimers and Gurevych, 2019). Instead of the above-mentioned losses, we use CrossEntropyLoss as our task is a classification.

²<http://docs.deeppavlov.ai/en/master/features/models/bert.html>

As can be seen in the next section, ConvBERT shows the best result but due to our approach to thread reconstruction and as a matter of fact that BERT-like models are computationally expensive, we also experiment with a distilling knowledge from ConvBERT into LSTM by the teacher and student paradigm (Tang et al., 2019). This would allow us to disentangle embeddings for two messages and reduce computations. The question is whether a quality loss would be affordable.

5.2 A note about tokenization

In social media, people tend to reduce some words and make abbreviations of frequent phrases to save typing time. This phenomenon can be seen in this work (Buyanov and Sochenkov, 2022) where authors study the language of Twitter posts of persons with suicidal tendencies. Another thing is that people try to mimic speech methods of emotional expression, thus they multiply vowels ("whaaaaat?"). In particular, Russian speakers express laughter like "axaxax" ("ahahah"). Due to a combination of different sequence lengths, sequence permutation errors, and mistyping (sometimes intended) the vocabulary of laughing can be very large, and the normalization of this vocab is a tough task. All described facts lead to an enormous vocabulary size, where many entities have a low frequency. So we think that using the BERT tokenizer, which works on a subword level, with Logistic Regression and LSTM models could be beneficial in terms of vocab efficiency.

In our experiments, we compare two ways of tokenization. In first one is a using tokenizer that does not use subword tokenization. For that purpose, we use a tokenizer from NLTK and will refer to it as **simple tokenizer**. In a second way, we use the **BERT tokenizer** and will refer to it as its name.

5.3 Thread reconstruction method

As a baseline, we consider a dummy heuristic that connects the present message with the previous one. Following our assumption, we need to reconstruct all “reply to” relations that would reveal threads. Having a model that can predict the relation between two messages, we would apply this model to all pairs of messages inside a chat of length N . Although, with our models, it is prohibitively expensive as we need to proceed a $O(N^2)$ pairs. Likely, from the data analysis we know that the biggest part of the pairs have a distance of less than 15 messages, so we can restrict the search space to this count reducing the count of pairs to $O(15 * (n - 15))$.

It is worth noting that we would recompute the embeddings of almost every text 15 times. While using the Siamese network, we can escape this computational overhead by precomputing embeddings of all texts with the basic network. Having a matrix, we then can just use classification head on pairs of rows, which is much cheaper to compute. Unfortunately, to compute BERT in the NSP regime we have to form a string with a special format, so here we can’t escape these expenses.

We organize the predictions in the $N \times N$ probability matrix, from which we must derive a valid adjacency matrix. We experiment with two simple heuristics. The first is to keep the nearest predicted reply w.r.t. current message position. For example, if the 7th message 2nd, 5th, and 6th messages were predicted as replies, then we chose the 6th. We will call this *greedy binary*. The second heuristic is to choose the message as a reply that has the highest probability score among all possible ones. We will call it *greedy probas*. Moreover, we can variate how many messages before the current position we will consider. Having the processed matrix, we can derive threads as a collection of connected graph components. We also can use this matrix to compare it with human annotation.

6 Results

6.1 Reply recovery

As we have a standard classification task, we report precision, recall, and F1 score. In Table 1 the result on the entire dataset is presented. We can see that ConvBERT shows the best result overall models. The simple linear regression fails to discover any useful relations to tackle the problem. Another observation is that Siamese architectures perform worse and for the LSTM the performance drop is very significant compared to SentenceBERT. We also see that using the BERT tokenizer for the LSTM model is beneficial compared to simple token dictionary tokenization. As for distillation, we see that model fails to learn

knowledge from BERT. We leave the study of why to further research, but it’s probably a good idea to focus on DistilBERT (Sanh et al., 2019) as a base block for SentenceBERT.

An important observation is that the model performance depends on a chat domain. In Table 2 we see the models variate significantly. For example, the ConvBERT model ranges up to 10 score points. We think that one of the factors is a chat topic and goal that influences symbol length. The chats where participants ask questions or discuss complex living situations tend to have more message symbol length, like *sling38* where participants are young mothers. Conversely, game chats are about fun which is not required to write a lot. The Pearson correlation between average pair length and ConvBERT results is 0.62 with a p-value of 0.076. Another factor is probably the lexicon specificity of chats. However, these factors do not cover all cases. Although, *cotedazuchat* is a chat of Russian emigrants in France where they discuss various topics, and it has medium average text length compared to others, it has the worst performance score.

Model name	Precision	Recall	F1
with BERT tokenizer			
LSTM	0.651	0.719	0.628
Siames LSTM	0.668	0.491	0.539
Logistic regression	0.475	0.529	0.501
with simple tokenizer			
LSTM	0.602	0.606	0.566
Siames LSTM	0.507	0.537	0.505
Logistic regression	0.474	0.528	0.500
ConvBERT	0.822	0.846	0.833
DisilledLTSM	0.459	1.000	0.630
SentenceBERT	0.786	0.838	0.797

Table 1: Model scores overall reply recovery test set.

Chat name	Avg pair len	LSTM-ST	SLSTM-BT	LSTM-BT	SLSTM-ST	SBERT	ConvBERT
balichat_woman	140.948	0.695	0.688	0.659	<u>0.580</u>	0.857	0.883
borussia_chat	81.697	0.585	0.627	0.605	<u>0.568</u>	0.798	0.821
chat_suicidnikov	72.400	0.538	0.576	0.579	<u>0.533</u>	0.784	0.826
cotedazurchat	93.785	0.575	0.628	0.588	<u>0.536</u>	0.762	0.793
easypeasycodachat	94.804	0.702	0.519	0.555	<u>0.278</u>	0.859	0.885
openwrt_ru	101.410	0.556	0.620	0.566	<u>0.527</u>	0.788	0.885
orange_sosedi	151.584	0.625	0.647	0.658	<u>0.533</u>	0.817	0.849
sling38	174.404	0.686	0.655	0.668	<u>0.512</u>	0.842	0.890
terrariaphone	69.732	0.578	0.624	0.618	<u>0.544</u>	0.801	0.841

Table 2: Model scores for different chat domains. BT is the BERT tokenizer, ST is the simple tokenizer. The best score is in bold, and the worse one is underlined.

6.2 Thread reconstruction

As we can see from Table 3 none of the proposed heuristics can beat the baseline. Comparing two proposed heuristics, we could say that the selection of the highest probability generally performs worse than taking the first predicted message. Another note is that with increasingly considered messages, *greedy probas* performs significantly worse than *greedy binary*. The probable explanation is that *greedy binary* is aligned with the observation that in general, the reply is a previous message. In contrast, in *greedy probas* relies on predictions of not so excellent model. Based on these results, we see that simple heuristics are not enough to restore threads with the reply recovery model of the proposed quality. We hypothesize that the model for thread reconstruction should consider the context of the dialogue, or the

reply recovery model should have better performance.

Model name	Precision	Recall	F1
dummy	0.776	0.814	0.793
greedy_binary_15	0.740	0.749	0.743
greedy_binary_3	0.799	0.735	0.762
greedy_binary_7	0.753	0.747	0.749
greedy_probas_15	0.640	0.645	0.642
greedy_probas_3	0.768	0.710	0.735
greedy_probas_7	0.680	0.676	0.677

Table 3: Scores of methods on thread reconstruction test set

7 Conclusion and future work

In this work, we investigate the asynchronous chats in the Russian language. We show that modern pre-trained models show great results on the task of reply recovery compared to more simple models. We also experiment with a thread reconstruction task based on restored “reply to” relations with simple heuristics. The results show that these heuristics are not enough, despite the relatively good result of the model.

As a future work, we can highlight the investigation of pipelines that would produce a dataset of better quality, as we found that using replies only and random sampling is not guaranteed to have good labeling. Possibly weak supervision approach can be used to better filter out broken pairs. Speaking of thread reconstruction, our small benchmark is far from ideal, so it certainly can be improved. Also, using the reply recovery model, one can create a noisy train dataset on thread reconstruction. Such datasets can reduce the cost of annotation of large-scale data. From the model point of view, the graph neural networks can probably benefit from the natural graph structure of chat conversations.

Acknowledgements

We thank our annotators Anastasiya Dyachenko and Alexander Lugarev for their work for the thread reconstruction dataset. The research was funded by the Ministry of Science and Higher Education of the Russian Federation in accordance with agreement № 075-15-2020-907 of 16.11.2020. The grant was provided for state support for the creation and development of a World-class Scientific Center “Agro-technologies for Future”.

References

- Ali Balali, Hesham Faily, Masoud Asadpour, and Mostafa Dehghani. 2013. A supervised approach for reconstructing thread structure in comments on blogs and online news agencies. *Computación y Sistemas*, 17.
- Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34:1–34.
- Igor Buyanov and Ilya Sochenkov. 2022. The dataset for presuicidal signals detection in text and its analysis. *Computational Linguistics and Intellectual Technologies*.
- A. Philip Dawid and Allan Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, 28:20–28.
- Mostafa Dehghani, Azadeh Shakery, Masoud Asadpour, and Arash Koushkestani. 2013. A learning approach for email conversation thread reconstruction. *Journal of Information Science*, 39:846 – 863.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.
- Gaoyang Guo, Chaokun Wang, Jun Chen, and Pengcheng Ge. 2018. Who is answering to whom? finding “reply-to” relations in group chats with long short-term memory networks.

- Lu Ji, Jing Li, Zhongyu Wei, Qi Zhang, and Xuanjing Huang. 2021. Who responded to whom: The joint effects of latent topics and discourse in conversation structure. *ArXiv*, abs/2104.08601.
- Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *ArXiv*, abs/1905.07213.
- Annie Louis and Shay B. Cohen. 2015. Conversation trees: A grammar model for topic structure in forums. // *Conference on Empirical Methods in Natural Language Processing*.
- Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with siamese recurrent networks. // *Rep4NLP@ACL*.
- Tien Dat Nguyen, Shafiq R. Joty, Basma El Amel Boussaha, and M. de Rijke. 2017. Thread reconstruction in conversational data using neural coherence models. *ArXiv*, abs/1707.07660.
- Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. Crowdspeech and vox diy: Benchmark dataset for crowdsourced audio transcription. // J. Vanschoren and S. Yeung, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.
- Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. Open source software available from <https://github.com/heartexlabs/label-studio>.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.
- Yi-Chia Wang, Mahesh Joshi, William Cohen, and Carolyn Rosé. 2008. Recovering implicit thread structure in newsgroup style conversations. 01.

Binary classification model as a tool to detect sentences with microsyntactic units¹

Chaga A. V.

Institute for Information Transmission Problems (Kharkevich Institute),
Russian Academy of Sciences, Moscow, Russia
chagachaga@gmail.com

Abstract

We consider a model of binary classifier predicting occurrence of microsyntactic units in sentences. The model is based on AWD-LSTM architecture with an encoder pre-trained on the Russian version of Wikipedia and further trained on a dataset built from the SynTagRus corpus supplied with a microsyntactic markup. We present the structure of the model and discuss its output. The study showed that binary classification allows targeting of microsyntactic markup and helps to significantly improve its recall.

Keywords: microsyntax; binary classification; AWD-LSTM

DOI: 10.28995/2075-7182-2023-22-1061-1070

Бинарный классификатор как инструмент для поиска предложений, содержащих конструкции микросинтаксиса

Чага А. В.

Институт проблем передачи информации РАН им. А. А. Харкевича,
Москва, Россия
chagachaga@gmail.com

Аннотация

В данной статье рассматривается модель бинарного классификатора, предсказывающего наличие единиц микросинтаксиса в предложениях. Модель построена на основе архитектуры AWD-LSTM с предобученным энкодером на русскоязычной версии Wikipedia и дообученной на наборе данных из синтаксического корпуса СинТагРус, имеющего микросинтаксическую разметку. В работе приводится структура модели, а также рассматриваются результаты её работы. В процессе исследования выяснилось, что бинарный классификатор позволяет таргетировать микросинтаксическую разметку и существенно увеличить её полноту.

Ключевые слова: микросинтаксис; бинарная классификация; AWD-LSTM

1 Вводные замечания

Элементы, которые входят в область микросинтаксиса, исследовались и продолжают активно исследоваться лингвистами, но ввиду своего положения на стыке грамматики и лексики, а также ввиду своего специфического устройства с большим трудом поддаются систематизации и полноценному анализу. Тем полезнее представляется работа над созданием микросинтаксического словаря и микросинтаксической разметкой, а также широкое и обстоятельное исследование микросинтаксических единиц русского языка, проводимое на протяжении двух десятков лет в Лаборатории компьютерной лингвистики Института проблем передачи информации им. Харкевича РАН. СинТагРус является полностью отредактированным

¹ This work was done with the financial support of a grant from the Russian Science Foundation, No. 22-28-01941 “Development of the infrastructure and the first phase of the semantic corpus for Russian”.

экспертами-лингвистами корпусом текстов на русском языке с аннотацией на морфосинтаксическом уровне, предоставляя ценный материал с высоким качеством разметки как для теоретических исследований, так и для задач компьютерной лингвистики.

Л.Л. Иомдин (2015, 2019) предложил термин «микросинтаксис» для обозначения множества неоднословных языковых элементов той или иной степени идиоматичности, для которых характерно нестандартное синтаксическое поведение. Во многом эта область сближается и пересекается с классической фразеологией, в особенности, грамматической фразеологией. Для микросинтаксических конструкций типична семантическая некомпозициональность, высокая степень устойчивости, нерегулярность, то есть выход за рамки общих грамматических правил, реинтерпретация грамматических характеристик, когда один или несколько компонентов выражения меняют свой категориальный статус, как, например, междометие *была не была*, составленное из двух глаголов и частицы.

Примерами конструкций микросинтаксиса выступают разнообразные адвербиалы (*как можно лучше, как бы то ни было*), сложные союзы (*так как, потому что*), составные предлоги (*по отношению к, во главе*), частицы (*что ни на есть, нет-нет да и*), вводные выражения (*надо сказать*), дискурсивные единицы (*вот те на, а то*), различные синтаксические конструкции, лексическое наполнение которых имеет определенную степень свободы (*не наХ-оваться: не могу налюбоваться, не набегаешься*) и др.

Сложность идентификации единиц микросинтаксиса заключается в том, что в языке существуют тысячи синтаксических идиом, и, имея набор заданных показателей, не всегда легко даже вручную определить границу между свободным словосочетанием или иной регулярной конструкцией и единицей микросинтаксиса. Не всегда последовательности слов, по внешним признакам удовлетворяющих описанию конструкции, к ним относятся. Речь идёт, в первую очередь, о «ложноположительных» единицах, когда идентичные по форме лексические элементы не образуют единицы микросинтаксиса. Ср.:

- (1) (a) Так ещё и **надо сказать** определённым образом и достаточно внятно, чтобы она поняла. [Rozetked Discuss. telegram Rozetked Discuss (09.12.2021)]²
- (1) (b) Муравьи, **надо сказать**, всегда вызывали у некоторых из нас немалый интерес, как и другие социальные насекомые, хотя большинство относится к ним с раздражением, особенно когда они воруют у нас сахар. [Алексей Петрович Цветков. Муравьиный космос (2020)]
- (2) (a) День, когда это началось, был **тем самым** днем, когда терпение царя и царедворцев лопнуло. [Татьяна Георгиевна Щербина. Терпение лопнуло (2006)]
- (2) (b) Смысл выживания данной инфекции в том, что она оперативно меняет свою генетическую культуру - и **тем самым** выживает в среде человека. [Геннадий Григорьевич Онищенко. Зима без гриппа: Оценить серьезность проблемы (1999)]
- (3) (a) Ознакомившись с телеграммой, Шевченко явно растерялся и ушел от меня **в состоянии** протрации. [Олег Трояновский. Через годы и расстояния (1997)]
- (3) (b) Не надо думать, что все **в состоянии** освоить такие вещи. [Т. В. Ершова, Н. А. Никифоров. Качество работы госслужащего измеряется не наградами, а полезными результатами для людей // «Информационное общество», 2011]

В приведённых выше предложениях (1 - 3) (a) представлены свободные словосочетания, омонимичные микроединицам, но ими не являющиеся, а в (1 - 3) (b) выступают собственно микросинтаксические конструкции.

Учёт периферийных синтаксических явлений языка способствует адекватному анализу текста и его качественному переводу на другие языки. Установление и описание явлений

² Здесь и далее примеры взяты из Национального корпуса русского языка (ruscorpora.ru)

микросинтаксиса представляется важным для лингвистики, в том числе для решения практических задач в прикладных областях, таких как межъязыковая типология, автоматический перевод и семантический анализ текстов.

2 Цель исследования

Цель нашей работы состояла в разработке и проверке эффективности бинарного классификатора, построенного на основе нейронной сети и способного предсказывать наличие хотя бы одной микросинтаксической единицы в предложении естественного языка. Это первый шаг на пути к автоматической детекции конкретных единиц микросинтаксиса, а также поиску новых микросинтаксических конструкций русского языка.

Метод, который мы применили к задаче автоматического поиска фразеологии, ранее к такой задаче не применялся, хотя и использовался для решения других, во многом очень похожих задач, например, для анализа тональности текста (Katsarou et al., 2022), классификации идиоматичных фраз (Briskilal, Subalalitha, 2022) и др. Несмотря на простоту поставленной задачи, результаты модели имеют прикладную пользу, о чём будет сказано ниже.

3 Подготовка данных, создание и обучение модели

Для автоматического поиска предложений, содержащих микросинтаксические конструкции, мы использовали языковую модель глубокого обучения AWD-LSTM (Merity et al., 2017), реализованную в библиотеке `fastai` с использованием `PyTorch`³. AWD-LSTM расшифровывается как `ASGD Weight-Dropped Long Short-Term Memory`: модель с усреднённым стохастическим градиентным спуском, регуляризацией весов и долгой краткосрочной памятью. Это нейронная сеть, которая использует рекуррентный блок LSTM, а также различные стратегии регуляризации и оптимизации, такие как `DropConnect` для снижения риска переобучения сети путём введения разреженности весов модели, метод стохастического усреднённого спуска, метод усечённого обратного распространения ошибки при обновлении весов, регуляризацию активации и другие приёмы, позволяющие модели эффективно обучаться, сохраняя нужные паттерны, выученные из прошлого контекста и выбрасывая из памяти ненужное.

Архитектура AWD-LSTM была выбрана, поскольку она показала свою эффективность в некоторых задачах классификации, схожих с нашей (Briskilal, Subalalitha, 2022), (Kiran, Shashi, Madhuri, 2022), (Tao, et al., 2019).

В своей работе мы по большей части применяли методы, представленные в библиотеке `fastai`, адаптируя их к русскому языку. Языковая модель была обучена на русскоязычной версии `Wikipedia`. Для этой задачи мы использовали стандартный набор методов преобразования текста, используя встроенные функции и стандартные настройки из библиотеки `fastai`: замена переноса строки, приведение всех букв к строчным с последующим добавлением специальных токенов и другие. Токенизатор был взят из библиотеки `spaCy` для русского языка. В процессе первоначального обучения модели был составлен словарь объёмом в 60000 токенов, что соответствовало стандартным рекомендациям и нашим требованиям ко времени, затраченному на обучение.

Далее модель прошла дообучение на подготовленном нами наборе данных, полученных из материала корпуса `СинТагРус`, имеющего, помимо других видов аннотации, микросинтаксическую разметку. Все тексты `СинТагРус` представлены в формате XML, где каждое предложение, содержащее хотя бы одну микросинтаксическую конструкцию, имеет соответствующий тэг. Корпус состоит из 107132 предложений, входящих в 1305 текстов. Каждое предложение получило метку о наличии либо отсутствии микроединицы в своём составе. Таким образом, для дообучения языковой модели использовались сырые предложения из `СинТагРус` с тэгами о наличии микроединиц. Токенизация корпуса не учитывалась.

В момент написания настоящей работы в корпусе выделено и размечено 41697 единиц микросинтаксиса в 31322 предложениях, а всего словник микросинтаксических элементов содержит 3119 единиц. Набор собранных нами размеченных данных был разделен на

³ <https://github.com/fastai/fastai>

обучающую, валидационную и тестовую выборки в соотношении 80, 10 и 10% от общего объёма данных соответственно. Тестовая выборка использовалась только для оценки качества работы классификатора и не использовалась для обучения модели.

Бинарный классификатор строился с помощью функции `text_classifier_learner`⁴ из библиотеки `fastai`, которой в качестве аргументов передаётся набор данных для обучения, архитектура AWD-LSTM, и значения гиперпараметров и метрик по умолчанию. В дообучении использовался словарь, собранный во время обучения языковой модели. Исходя из характеристик имеющегося оборудования, а также учитывая диапазон длин предложений в собранном нами датасете, мы использовали длину обрабатываемой последовательности (`seq_len`) в 72 токена. Обучение проходит за 4 эпохи, поскольку на этом этапе значение ошибки на валидационной выборке оказывается самым низким.

Получив на вход строку с предложением, на выходе бинарный классификатор предсказывает для этого предложения наличие микроединицы в его составе и присваивает ему соответствующую метку без уточнения конкретной конструкции.

4 Оценка качества предсказаний модели бинарного классификатора и его сравнение с базовой моделью

Для формальной оценки результатов бинарной классификации были посчитаны точность, прецизионность, полнота и F-мера (см. Таблицу 1).

Для того, чтобы убедиться в целесообразности использования нейросетевой архитектуры при построении бинарного классификатора, мы создали простейшую (базовую) модель классификатора и сравнили результаты, полученные на тех же наборах данных, которые были использованы в основной модели.

Базовый классификатор устроен следующим образом: из списка предложений, входящих в обучающую и валидационную выборку, был извлечён список всех микросинтаксических единиц, где каждая конструкция представлена в виде пары {начальный элемент + конечный элемент}, ср.: *по причине, по меркам, ...* {'по': 'причине', 'меркам', ...}, *в мгновение ока, в момент времени, ...* {'в': 'ока', 'времени', ...}, *абы как* {'абы': 'как'} и т.д.

Таким образом был составлен словарь из 37157 единиц. Далее мы проверяли все предложения из тестовой выборки на наличие в них конструкций из собранного словаря. В случае, если предложение содержало оба элемента конструкции с соблюдением порядка их следования, то оно отмечалось как содержащее микроединицу, в противном случае предложение помечалось как не имеющее микроединиц в своём составе. В базовом классификаторе мы использовали токенизацию корпуса `СинТагРус`, все элементы конструкций приводились к нижнему регистру.

Для оценки качества базовой классификации на тестовой выборке были посчитаны те же метрики, что и для основной модели. Результаты сравнения приведены ниже.

Модель классификатора	Прецизионность	Полнота (recall)	F-мера	Точность
базовый классификатор	0.45	0.93	0.61	0.64
AWD-LSTM классификатор	0.87	0.76	0.81	0.89

Таблица 1: Сравнение результатов классификации на тестовой выборке

Ожидаемым образом, метрика полноты базового классификатора имеет более высокий уровень по сравнению с основной моделью. С одной стороны, базовый классификатор идентифицирует большее число микроединиц, а с другой, имеет серьёзный недостаток по сравнению с основной моделью. Число ложноположительных случаев оказывается чересчур высоким. См. Таблицу 2.

4 https://docs.fast.ai/text.learner.html#text_classifier_learner

Модель классификатора	true_positive	true_negative	false_positive	false_negative
базовый классификатор	2976	3881	3625	231
AWD-LSTM классификатор	2453	7149	754	357
объём тестовой выборки: 10713				

Таблица 2: Количественное сравнение предсказанных меток по тестовой выборке

5 Результаты исследования

Базовый классификатор наиболее точно идентифицирует предложения, содержащие неразрывные конструкции с фиксированными первым и последним элементами вроде *в X-овой мере* (в значительной / какой-то / немалой мере), *в X-овом смысле* (в указанном / прямом / узком смысле), *до сих пор*, *с точки зрения* и т.д. Тем не менее, как уже было сказано выше, количество ложноположительных предложений у базового классификатора оказывается неприемлемо высоким. В частности, не будут различены случаи вроде (4), где имеется микроединица *и всё*, и (5) и (6), где совпадают первый и последний элемент, но ни о каком обороте нет речи:

(4) Не могла простить *и все*.

(5) Папа, мама *и* Толя, *все* уехали в город, я осталась одна.

(6) *И все* поют.

Можно было бы несколько улучшить базовую модель, но она по определению не сможет идентифицировать микроединицы со свободным лексическим наполнением, вроде *X за X-ом* (*ступенька за ступенькой, препятствие за препятствием*), или предсказывать в предложениях новые конструкции микросинтаксиса.

Модель бинарного классификатора, построенная на архитектуре AWD-LSTM, показала существенно более высокий уровень прецизионности, нежели базовый классификатор. Это выражается в том, что количество ложноположительных случаев существенно меньше. В частности, оба предложения (5) и (6) были классифицированы верно. В целом нейросетевая модель оказывается более полезной в прикладном плане. Далее мы будем рассматривать только её.

Для быстрой первоначальной проверки качества классификации текстов на наличие микроединиц проверялись случайные предложения, не представленные в синтаксическом корпусе. Было взято 10 предложений из детских рассказов В.И. Драгунского. Модель правильно классифицировала 8 из них. Все предложения, действительно не содержащие микросинтаксические конструкции, получили правильное значение False для меток MICROSYNT, а все прочие – значение True. Два предложения, содержащие микроединицы, были ошибочно классифицированы как не имеющие таких единиц. Также специально выбирались предложения, содержащие микроединицы, не ещё представленные в корпусе СитТагРус, и модель идентифицировала два случая:

(7) Он её спас, а Чапку постегал прутиком — *для виду*, конечно.

(8) *Вот тебе раз!*

Всё же, поскольку поиск новых единиц микросинтаксиса изначально не входил в цель исследования, и представляет собой отдельную задачу, далее мы анализировали только примеры из корпуса СинТагРус.

Из тестовой выборки были случайно извлечены и вручную проверены 600 предложений. Сначала были проанализированы случаи, в которых классификатор показал свою эффективность. Это случаи с редкими микроединицами:

(9) Аня увидела *самое себя*.

Микроединица *‘самое себя’* была размечена в корпусе лишь один раз.

(10) ***Ничего подобного!***

Микроединица *‘ничего подобного’* размечалась в корпусе три раза.

Стоит отметить, что далеко не все микроединицы в корпусе имеют высокую частотность: 80% из них, или 2519 единиц словника, имеют 10 и менее отмеченных вхождений, а 37,5% микроединиц представлены в корпусе всего один раз. Тем не менее, классификатор успешно выделяет предложения с такими конструкциями.

Поскольку мы имеем дело с бинарным классификатором, нет возможности проверить, какие именно последовательности слов (цельные или разрывные) влияют на результат предсказания алгоритма, но судя по коротким предложениям, можно с высокой долей вероятности предположить, что классификатор способен учитывать и запоминать некоторые редкие паттерны.

Программа правильно классифицировала некоторые предложения, содержащие ранее не встречавшиеся выражения микросинтаксиса, но похожие по своей структуре на уже установленные:

- *в логике*, ср.:

(11) ***В этой логике*** каждый "недоплаченный" бюджетный рубль оборачивается рублем прибыли для того начальника, который может найти способ продать необходимую услугу населению.

В корпусе размечалось выражение *“по логике вещей”*.

- *в границах*, ср.:

(12) Все здесь думают ***в границах*** определенных рамок.

В корпусе отмечались единицы типа *“за границу”*, *“за границей”*, *“из-за границы”*, а также *“в рамках”*.

Стоит упомянуть, что классификатор успешно выделяет предложения со сложными союзами, где элементы далеко отстоят друг от друга:

(13) ***Чем*** большее расстояние мог охватить взгляд, ***тем*** быстрее хотелось достигнуть далеких вершин и с них оглядеть новые непокоренные места.

(14) Уже во второй половине дня 7 мая, сразу после инаугурации Владимира Путина, стали известны ***не только*** кандидатура на пост главы правительства (Госдуме предложено пере назначить Дмитрия Медведева), ***но и*** основные вице-премьеры (их назвал будущий премьер-министр на встрече с фракцией "Единой России" в Госдуме), и основные приоритеты нового Белого дома.

(15) ***Как*** колхоз сдавал зерно на хлебозаготовку по цене, которая никак от него не зависела, ***так и*** российские корпорации продают свои ресурсы по цене, которая никак от них не зависит.

В микроединице *не только ... но и* между первым и последним элементом 11 слов, а в *как ... так и* – 13.

Классификатор также выделяет предложения с разрывными единицами вроде конструкции *в порядке*, внутрь которых довольно часто вставляются другие лексические элементы, причем в корпусе данный пример с тремя вставленными элементами был пропущен разметчиком:

(16) Но эти папские милости ***в особом и закрытом порядке*** сделали людей предметом торга, разменной монетой на переговорах.

Конструкция *в свете* представлена в корпусе 8 раз, и лишь один пример содержит вставной элемент (*в их свете*), в данном же примере в конструкцию вошло три слова. Ср.:

- (17) Под влиянием непостижимого предубеждения всё самое простое и обыкновенное представилось *в каком-то таинственном, враждебном свете*.

Несмотря на то, что классификатор плохо обнаруживал предложения с повторяющимися элементами, некоторые предложения с новыми паттернами он всё же выделил:

- (18) Нагнулся раз, нагнулся другой...

Также классификатор успешно идентифицировал предложения с конструкциями, вроде *речь идет о*, в которых глагол может довольно сильно варьироваться:

- (19) *Речь зашла о* "теории эмбрионального поля", предложенной профессором Гурвичем.

Рассмотрев случаи, когда наш бинарный классификатор успешно справляется с поставленной задачей, необходимо также рассмотреть примеры, где он систематически допускает ошибки. Для этого проверялись предложения, где были обнаружены расхождения в аннотации между имеющимися метками, проставленными в процессе ручной разметки, и результатами предсказания классификатора.

Во-первых, это случаи с так называемыми «ложноположительными» единицами, когда идентичные по форме лексические элементы не образуют конструкции микросинтаксиса, а являют собой свободные словосочетания или регулярные конструкции:

- (20) Эпоха метамодерна предполагает колебания "между модернистским стремлением к смыслу и постмодернистским *сомнением в смысле всего этого*", "между иронией и энтузиазмом, между сарказмом и искренностью, между эклектичностью и чистотой, между разрушением и созиданием".

- (21) А вечером приезжают сюда на джипах и, не выходя из них, наблюдают за косолапыми *в свете фар*.

- (22) Взросление нынешней молодежи *пришлось на время формирования* в России общества потребления.

- (23) Еще не было ни одного *заседания по делу*, кроме предварительного, а Романа уже уволили с работы.

Во-вторых, мы обнаружили, что модель практически не выявляет конструкции с переменными, такими как *взять и X-овать*, а также микроединицы с повторяющимися или частично повторяющимися элементами: *пропади она пропадом, шёл и шёл* и т.д.:

- (24) Я бы, не скрою, молился ему, он же, сказав, что сказал, *взял и умер*.

- (25) Вариации: зеленый ключик высоты передается *от вершины к вершине* и каждая новая гряда запирает лощину на замок.

Мы также обратили внимание, что алгоритм неправильно размечал предложения, содержащие конструкции с определенными опорными элементами вроде существительного *суд*.

6 Практическое применение бинарного классификатора

Несмотря на то, что мы работаем над созданием модели для автоматической идентификации микросинтаксических конструкций в новых текстах, и отдаём себе отчёт в том, что бинарная

классификация не может в полной мере решить поставленную задачу, оказалось, что применительно к уже размеченному корпусу, такая классификация может быть полезна.

Мы собрали предложения не только из тестовой выборки, но также из тренировочной и валидационной выборки и отобрали случаи, когда классификатор предсказывал наличие хотя бы одной микросинтаксической единицы в предложении, не имеющем аннотации. Из 107132 предложений корпуса было выявлено 3927 случаев такого рода. Далее, вручную было проверено 400 случайно выбранных предложений из этой группы случаев (чуть более 10%) и установлено, что из них 82.7% действительно имеют хотя бы одну микроединицу в своём составе, а для 17,2% предложений предсказание классификатора оказалось неверным. Таким образом, мы обнаружили примерно 3% потенциально недоразмеченных предложений.

Дело в том, что микросинтаксическая разметка производится на протяжении нескольких лет, в течение которых реестр микросинтаксических единиц всё время пополняется, а сама работа проводится силами разных специалистов с разной степенью подготовки. Учитывая постоянное изменение состава словника и человеческий фактор, наличие некоторого числа недоразмеченных предложений в корпусе неизбежно.

Возможно, поэтому среди правильно классифицированных недоразмеченных предложений выявляется много случаев со сложными союзами вроде *не только ... но и, как ... так и, чем ... тем, если ... то* и др. В процессе ручной разметки довольно легко пропустить такие сложные союзы, поскольку из-за своей частотности и далеко отстоящих друг от друга элементов они легко выпадают из поля зрения аннотатора.

На текущем этапе мы не ставим целью полностью автоматизировать процесс идентификации микроединиц. Во-первых, ввиду сложности исследуемого объекта, обилия и разнообразия конструкций микросинтаксиса, встречающихся в текстах естественного языка, мы рискуем упустить тонкие и нетривиально устроенные единицы, а во-вторых, ручное аннотирование даёт самое высокое и надёжное качество разметки и позволяет устанавливать новые виды конструкций, тем самым пополняя перечень микроединиц.

Проанализировав результаты классификации на всём объёме корпуса, было обнаружено несколько новых микроединиц, например, *'на таком-то году жизни'*, *'брать за образец'*, *'тому есть X'*:

(26) Он умер в минувшую пятницу *на 95-м году жизни*, отдав всего себя без остатка Израилю.

(27) Кого *взять за образец*.

(28) И *тому есть немало оснований*.

Как можно было заметить, классификатор учитывает самые разные типы конструкций. На результат предсказания не сильно влияет частотность микроединицы в выборке, её морфологические характеристики, положение во фразе, наличие вставных элементов и количество токенов между первым и последним элементом конструкции. В список потенциально содержащих микроединицы фраз попадают как длинные, так и короткие предложения.

Чаще всего модель правильно классифицирует предложения, содержащие микросинтаксические конструкции с фиксированными элементами вроде: *до сих пор, помимо прочего, хотя и, хотя бы, по поводу, в области, кроме того, до тех пор, на вид, потому что, на взгляд, вместо того, на время, с успехом, на основе, в принципе* и т.д. Судя по всему, запомнив паттерн, модель успешно классифицирует конструкции и со вставными элементами: *по поводу* и *по этому поводу, по виду* и *по внешнему виду*.

В случайной выборке из 200 предложений, проверенной нами вручную, не было правильно классифицировано ни одного предложения, содержащего микроединицу с повторами того или иного рода, а всего их было 19. Это конструкции как с максимально свободным лексическим наполнением вроде *звонит и звонит, слушали и слушали, из книги в книгу, от лотка к лотку, от киоска к киоску, возраст возрастом, но*, так и довольно специфичные конструкции типа *всякая всячина, издавший виды, мокры-мокрешеньки, сам не свой, вновь и вновь, пропади она пропадом, нос к носу* и др.

Было обнаружено, что классификатор систематически игнорирует конструкции с опорными словами *дело, вечер, суд, Бог, чёрт*, и их производными. В контрольной выборке присутствовало 7 микроединиц, содержащих конструкции вида *Бога ради, слава Богу, Бог с вами*, 3 конструкции со словом *чёрт*: *чёрт знает куда, какого чёрта*, 4 конструкции типа *к вечеру* и *по вечерам*, и ни одно из предложений с такими микроединицами не было выделено.

Производя сортировку конструкций словника по опорным лексическим элементам, мы заметили, что вокруг некоторых слов образуются целые группы разнообразных конструкций. В нашем словаре рекордсменом по количеству образованных конструкций является лемма *'время'* – 91 микроединица, от леммы *'раз'* образуется 61 конструкция, от *'дело'* – 60, от леммы *'рука'* – 47, от *'чёрт'* – 31, от леммы *'Бог'* – 25 микроединиц. Для повышения качества автоматической идентификации предложений, содержащих микросинтаксические выражения, необходимо учитывать такие опорные слова.

Созданный нами бинарный классификатор с довольно высокой точностью предсказывает наличие искомым выражений в конкретном предложении текста, тем самым позволяя детектировать микросинтаксические единицы и существенно увеличить полноту разметки корпуса.

7 Выводы и перспективы

Мы не ставили задачу создания максимально эффективной модели для идентификации предложений, содержащих единицы микросинтаксиса в своём составе. Бинарная классификация не является нашей конечной целью, поэтому мы не проводили сравнения эффективности разработанной модели с другими нейросетевыми классификаторами. Вместо этого мы сразу использовали её в качестве инструмента для повышения полноты микросинтаксической разметки в корпусе СинТагРус.

При разработке классификатора мы не использовали лемматизацию, а также не учитывали информацию о конкретных микроединицах, встречающихся в корпусе. Не использовалась и информация о синтаксической структуре предложений. Тем не менее, несмотря на простое устройство и короткое время обучения модели, было достигнуто значение F-меры 0.81.

Благодаря идентификации 3% потенциально недоразмеченных предложений мы сможем повысить качество микросинтаксической разметки синтаксического корпуса СинТагРус, которая в скором времени станет доступна на сайте НКРЯ. Помимо этого, в процессе проверки работы классификатора, был обнаружен ряд новых микросинтаксических конструкций.

Данный бинарный классификатор можно рассматривать и как систему, интерпретирующую коллокационный уровень фразы: анализ результатов классификации показал, что предложения, потенциально содержащие микроединицы, действительно, как правило, содержат устойчивые словосочетания, имеющие коллокационный уровень выше среднего.

На следующем этапе работы мы планируем создание модели автоматической детекции конкретных единиц микросинтаксиса, содержащихся в предложениях, представленных в микросинтаксическом словаре, и для этих целей мы планируем использовать другие архитектуры и подходы, в частности, применяющиеся в задачах по распознаванию именованных сущностей. Попробуем использовать лемматизацию и учитывать всю доступную информацию, содержащуюся в синтаксическом корпусе СинТагРус. Также нам кажется целесообразным использовать словарь большего объема для обучения модели, что должно положительно сказаться на качестве её работы.

Возможно, нам удастся приблизиться к решению проблемы отделения случайного соположения элементов от действительных случаев употребления этих элементов в качестве единицы микросинтаксиса и распознавать микроединицы со свободным лексическим наполнением и с повторяющимися элементами.

Благодарности

Автор признателен своим коллегам Л. Л. Иомдину и А. А. Мовсесяну за ценные советы и замечания.

References

- [1] Avgustinova, T., Iomdin, L. Towards a Typology of Microsyntactic Constructions. In: Corpas Pastor, G., Mitkov, R. (eds) Computational and Corpus-Based Phraseology. EUROPHRAS 2019.
- [2] Briskilal, J., Subalalitha, C.N. (2022). Classification of Idiomatic Sentences Using AWD-LSTM. In: Jeena Jacob, I., Gonzalez-Longatt, F.M., Kolandapalayam Shanmugam, S., Izonin, I. (eds) Expert Clouds and Applications. Lecture Notes in Networks and Systems, vol 209. Springer, Singapore. https://doi.org/10.1007/978-981-16-2126-0_11
- [3] Chaga A. (2021). On a specific Russian construction with saturative verbs and negation. Annual International Conference DIALOGUE 2021, student session, Moscow.
- [4] Howard, J. and Gugger, S. (2020). Deep Learning for Coders with Fastai and Pytorch: AI Applications Without a PhD, O'Reilly Media, Incorporated.
- [5] Iomdin, Leonid (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. // Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8-18. (<http://aclweb.org/anthology/W/W16/W16-38.pdf>). ISBN 978-4-87974-706-8.
- [6] Iomdin L.: Konstruktsii mikrosintaksisa, obrazovannye russkoj leksemej raz. [Constructions of microsyntax built by the Russian word raz.]. SLAVIA 2015, Časopis pro Slovanskou filologii, ročník 84, sešit 3, pp. 291-30. Praha (2015). (in Russian).
- [7] Iomdin, Leonid L. "Microsyntactic Annotation of Corpora and its Use in Computational Linguistics Tasks." Journal of Linguistics/Jazykovedný casopis 68 (2017): 169 - 178.
- [8] Katsarou Katerina, Sunder Sukanya, Woloszyn Vinicius, Semertzidis Konstantinos. (2022). Sentiment Polarization in Online Social Networks: The Flow of Hate Speech. 10.1109/SNAMS53716.2021.9732077.
- [9] Merity, Stephen & Keskar, Nitish & Socher, Richard. (2017). Regularizing and Optimizing LSTM Language Models. In International Conference on Learning Representations.
- [10] Sirra Kanthi Kiran, M. Shashi, K. B. Madhuri, "Multi-stage Transfer Learning for Fake News Detection Using AWD-LSTM Network", International Journal of Information Technology and Computer Science (IJITCS), Vol.14, No.5, pp. 58-69, 2022. DOI:10.5815/ijitcs.2022.05.05.
- [11] Y. Tao, et al., FineText: text classification via attention-based language model fine-tuning (2019). arXiv preprint arXiv:1910.11959.
- [12] Ziheng, Zeng & Bhat, Suma. (2021). Idiomatic Expression Identification using Semantic Compatibility. Transactions of the Association for Computational Linguistics. 9. 1546-1562. 10.1162/tacl_a_00442.

Corpus of Accentuated Byzantine Written Monuments and Methods of Its Markup

Evdokimova A. A.

Institute of Linguistics, Russian Academy of Sciences /
1 bld. 1 Bolshoy Kislovsky Lane, 125009 Moscow
arochka@gmail.com

Abstract

This article presents a corpus of Byzantine accentuated texts (BGAT) created since 2008. It currently includes 1010 Byzantine inscriptions, 950 papyri from various collections from the 1st to the 9th centuries, 132 seals from the collection of Dumbarton Oaks, and a selection of 100 Athos manuscripts from the 8th to the 15th centuries. Based on the collected data, we developed a method for markup such texts, which later makes it possible to create a database of accentuated texts from them and, based on the entire corpus, to train neural networks for classifying texts according to accentuation systems and recognizing them in images. As a result of marking texts, in addition to the previously known Alexandrian, Byzantine, and Dorian systems of accentuation, new accentuation systems were identified, including logical or semantic, with a shift of an accent mark to the right, with a displacement of an accent mark to the left, and mixed. For each group of monuments, their variants of using accentuation systems, especially the Alexandrian one, were identified, which show different aspects of the accentuation of the Byzantine Greek language. When creating a glossary of accentuated word forms based on the corpus, we determined that there were lexemes that retained their accentuation regardless of the influence of the dialect, meter, or traditions characteristic of the masters. However, a comparison of identical texts, even found in the same region of the Byzantine Empire, showed that the accentuation was not replicated when quoting.

Keywords: corpus BGAT, markup methods, accentuation systems, Byzantine Greek language, diacritical marks, Greek epigraphy, Byzantine sigillography, Byzantine manuscripts, papyri.

DOI: 10.28995/2075-7182-2023-22-1071-1081

Корпус акцентуированных византийских письменных памятников и методы его разметки

Евдокимова А. А.

Институт языкознания РАН /
125009, Москва, Большой Кисловский пер. 1 стр. 1
arochka@gmail.com

Аннотация

В данной статье презентуется корпус византийских акцентуированных текстов (BGAT), создаваемый с 2008 г. В его состав на сегодняшний день входят 1010 византийских надписей, 950 папирусов из разных коллекций с I по IX вв., 132 печати из собрания Dumbarton Oaks и выборка из 100 афонских рукописей от VIII по XV вв. На основе собранных данных была разработана методика разметки такого рода текстов, позволяющая в дальнейшем сделать из них базу акцентуированных текстов и на основе всего корпуса обучать нейронные сети классификации текстов по системам акцентуации и распознаванию их на изображениях. В результате разметки текстов кроме известных ранее александрийской, византийской и дорийской систем акцентуации были выявлены новые системы, среди которых логическая или смысловая, со сдвигом ударения вправо, со сдвигом ударения влево и смешанная. Для каждой группы памятников были выявлены свои варианты использования систем акцентуации, в частности александрийской, которые показывают разные аспекты акцентуации византийского греческого языка. При создании глоссария акцентуированных словоформ по материалам корпуса было определено, что существовали лексемы, сохраняющие свою акцентуацию независимо от влияния диалекта, стихотворного размера или традиций, свойственной мастерам. Однако, сопоставление одинаковых текстов даже в одном регионе показало, что акцентуация не заимствовалась при цитировании.

Ключевые слова: корпус BGAT, методика разметки, система акцентуации, Византийский греческий язык, диакритические знаки, греческая эпиграфика, византийская сфрагистика, византийские рукописи, папирусы.

[1] Введение

Акцентуированные греческие надписи византийского периода (IV-XV вв.), особенно граффити, показали, что, несмотря на использование в рукописях с X в. последовательной акцентуации, названной византийской, параллельно с ней существовали другие системы акцентуации. Среди них самая старая александрийская [3, 7, 9, 13, 15, 30], характерная для папирусов I-III вв., которая встречается и в надписях XIII-XV вв. из Северной Греции, Каппадокии, Грузии [18, 22].

С распространением технологий музеи и архивы занялись фотофиксацией своих коллекций с последующим представлением их в виде баз данных или электронных корпусов. Так в научный оборот стали поступать ранее не учтенные источники, хранившиеся в запасниках, и памятники, известные по изданиям, в которых отсутствовали фотографии. Как следствие, одной из актуальных задач для византистики стало описание всех коллекций в онлайн-формате и создание корпусов, разделенных по жанровым, географическим и иным принципам [ср. 11, 12, <http://papyri.info/>]. Частично это было реализовано при создании базы данных РНИ-7, где в виде текстов представлены папирусы, надписи и печати. Анализ этой базы с точки зрения акцентуации показал, что исследователи могут быть непоследовательны и отражать только те случаи акцентуации, которые показались им значимыми или унифицировать их, игнорируя данные критического аппарата в изданиях. К орфографии памятников применялся другой принцип. Она сохранялась, как в изданиях, откуда собирались надписи, что позволяет посчитать статистику для типов замен одних графем на другие. Поскольку существовала традиция при публикации византийских памятников письменности унифицировать знаки акцентуации в пользу позднейшей византийской системы или вовсе их пропускать, то сейчас в результате цифровизации источников появились новые данные по бытовавшим в Византии системам акцентуации и стало возможно их более детальное изучение. Однако, до сих пор не существует электронной базы акцентуированных греческих надписей и папирусов, включающей в себя фотографии, прорисовки, расшифровки, переводы, географические карты, лингвистические, палеографические и исторические комментарии, а также изданного корпуса подобных текстов или представленного онлайн. Все это делает затруднительным решение вопроса, какие системы акцентуации бытовали в византийский период, как они между собой коррелировали и как они влияли на произошедшие фонетические изменения. Однако, решение данного вопроса необходимо не только для истории систем акцентуации в греческом языке, но и для истории греческого языка в целом, а также как дополнительный источник для датировки памятника. Что привело нас к необходимости собрать корпус акцентуированных памятников в виде изображений и разметить их. В процессе сбора корпуса, начатого в 2008 г., и разметки данных проводились исследования, посвященные акцентуированным греческим памятникам разных регионов Византийской империи [18, 19, 20, 22, 23, 24, 25].

[2] Материал и метод

Собранный корпус акцентуированных памятников содержит:

- 1010 надписей из Каппадокии, Константинополя, Киева, Грузии, Болгарии, Северной Греции, Афин, Италии, большая часть которых датируется X-XV веками. Для каждого региона производилось ранжирование по типу надписи (лапидарная, граффито [22], дипинти etc) и по типу материала: на металле, на камне, на стеатите, на дереве, на фреске [25].
- 950 папирусов из разных коллекций с III по IX вв. Примерное соотношение акцентуированных папирусов и неакцентуированных иллюстрируют коллекции Кельнского университета - из 420 папирусов 60 акцентуированы [19].
- выборку из рукописей, хранящихся на Афоне (ок. 100 экземпляров от VIII по XV вв.) [24].
- 142 акцентуированные печати, отобранные из 11984 экземпляров коллекции Dumbarton Oaks <https://www.doaks.org/resources/seals>. Распределение акцентуированных печатей по векам показало, что на XI-XIII вв. приходится рост акцентуированных сфрагистических памятников [23].

Все памятники в корпусе при разметке были расклассифицированы в отношении 8 систем акцентуации, существовавших в Византии: александрийской¹ [1, 4, 9, 15], дорийской [3], византийской² [16, 18], смысловой или логической [24, 25], со сдвигом вправо [23], со сдвигом влево, соединения систем со сдвигами в одну систему, обусловленную поэтическим метром [20, 23], и смешанной, соединяющей черты любых из них.

Для классификации по системам учитывались диакритические знаки:

- акут,
- гравис,
- циркумфлекс,
- придыхание (легкое и густое),
- трема.

[3] Технические вопросы и подготовка к разметке

3.1 Связанные с текстовым представлением акцентуации

Традиционно для древнегреческого языка шрифты построены на последовательном проставлении диакритических знаков над ограниченным набором гласных. Так возникает проблема представления в уникоде различных знаков акцентуации и диакритики, если они отличаются от традиционно принятых, например, циркумфлекс над *ο* или трема над *ε*. Кроме передачи знаков в случае систем со сдвигами ударения, когда они оказываются над согласными или между двумя буквами, важной оказывается возможность поставить знак так, как он представлен в надписи, для возможности дальнейшего поиска по корпусу всех этих особенностей. Например, *λε'ω* выглядит, достаточно наглядным, однако шрифта, где ударение можно было бы сместить еще правее и написать его над любым согласным не существует. Другой вариант решения кроме разработки шрифта, содержащего все необходимые символы, это представить все слова в виде буквенных формул на латинице и задавать в них через определенные правила местоположения знаков, как и сами типы знаков в более общем виде. Один из способов представления таких формул приводится нами в разделе 4.3.

3.2 Обработка картинок и сопутствующих иллюстративных материалов

Каждый акцентуированный памятник представлен в корпусе в виде фотографии или прорисовки. В случае, если авторы изданий опубликовали и то, и другое, то оба материала учитываются, так как в них бывают разночтения. Чаще прорисовка отражает ту интерпретацию, которой придерживается издатель. При беглом анализе фотографий из одного региона видно различие в материале, на котором выполнена надпись (рис.1 и рис.2), и в использовании разного типа письма, маюскульного (рис. 2) или минускульного (рис. 3).



Picture 1: Италия, надпись на мраморе



Picture 2: Италия, надпись на металле

1 два и более ударений в слове, маркировка безударных слогов, ударение дальше, чем третий слог от конца, ударение на первый элемент дифтонга, ударения на сонорные согласные, гравис как маркер безударного слога, сдвиг ударения влево.
2 ударение точно над гласной, не далее третьего слога от конца, четкие правила употребления разных знаков акцентуации.



Picture 3: Италия, надпись на металле минускулом

Все фотографии фиксируются в учетной документации, которая помимо экстралингвистических данных (дата, регион, палеографические характеристики etc) указывает и некоторые формальные параметры для дальнейшего анализа систем акцентуации:

- наличие акцентных знаков над всеми словами,
- смещение ударения, с указанием типа смещения,
- наличие придыхания внутри слова,
- ударение над сокращениями,
- ударение над первым элементом дифтонга,
- ударения над безударными слогами,
- ударение дальше, чем третий слог от конца.

Каждая надпись или папирус имеет справочную карточку в общем учетном файле вида (рис. 4)³:

	<p>ВСН1590 – 1455-1512 гг., Мистра (bch23.121.10.1) Палеография: маюскульная надпись, буквы с имитацией объема, сокращение "K," для <u>κα</u> внутри слова. Акцентуация: ударения стоят почти над всеми словами и над сокращениями. Густое придыхание внутри слова над второй из следующих подряд гласных, ударение в дифтонге со смещением влево на первый гласный.</p>
--	--

Рис. 4: Пример сопроводительной карточки к надписи.

Каждая из акцентуированных печатей или рукописей представлена в виде фотографий с лингвистическими комментариями и гиперссылкой на электронную базу:



Феофилакт (XI в.) BZS.1958.106.1725
 +ΟΜΩ
 ΝΥΜΟΝ
 ΣΟΝΟΙΚΕΤ
 ΜΑΚΑΡ
 ΣΚΕΠ
 ὁμόνυμόν σον οἰκέτ(την), μάκαρ, σκέπ(οις)
 Придыхание ровно над гласной, акут над 3 и 4 словом, в 4 слове сдвинут вправо.

Рис. 5 BZS.1958.106.1725 Феофилакт (XI в.)

3 В скобках ссылка на файл с разбором всех акцентуированных слов

[4] Результаты и дискуссия

4.1 Теги и их варианты при разметке в зависимости от категории

В результате анализа данных, вошедших в корпус на сегодняшний день, были выявлены 8 систем акцентуации. Для удобства пользования каждой из систем приписывался тег (alex(andrine), dor(ic), byz(antine), mix(ed), log(ic), l-or (left-oriented), r-or (right-oriented), l-r-or (left-right-oriented)), и все акцентуированные слова маркировались этими тегами. В случае, если акцентуация в слове могла быть интерпретирована с точки зрения разных систем, то приписывались теги всех систем. За норму отсчета принималась позиция ударения по словарю, т.е. как принято в византийской системе акцентуации. Поскольку акцентуация литературных папирусов, стихотворных легенд печатей и текстов надписей часто связана с метрическими законами, использовалась отдельная система тегов для поэтических размеров (hex(ametr), 12 (=dodekasyllabos) etc.) и маркировки типов текстов (pr(osaic), po(etic), doc(umental) etc). Для надписей указывался их жанр: f(unerary) погребальные, bu(ilding) строительные, d(edication) посвяжительные, inv(ocation) молитвы etc.

Независимо от системы акцентуации, все акцентуированные позиции маркировались по критериям:

- соответствует ударению по словарю или правилам грамматики: tg (true grammatic = соответствует форме в словаре или акцентному типу согласно морфологической парадигме) или нет — ntg;
- произошло ли смещение ударения (shift), если нет, то тег nsh;
- куда произошло смещение ударения влево — sh_l (shift_left) или вправо — sh_r (shift_right) и на сколько;
- в случае дифтонга: первый или второй компонент оказался под ударением, d1 и d2 соответственно. Если знаков несколько и они оказались на разных частях дифтонга, то для каждого знака указывается его вариант через /;
- соответствует типу ударения, ожидаемому в этой позиции tt (true type) или нет — ntt.

Такие маркеры помогают определить при анализе, к какой из систем будет отнесена та или иная словоформа. Что в дальнейшем позволит не только выстроить удобную систему поиска, но и выявить, какие слова сохраняют собственную систему акцентуации независимо от общей системы акцентуации памятников, в которых они встречаются.

4.2 Соотношение орфографии и акцентуации

В византийских греческих надписях нет унификации орфографии и понятия нормы [ср. 14, 27, 28, 29], поэтому нельзя однозначно сказать, меняется ли функция у ударения в случае монофтонгизации дифтонгов или сокращения долготы гласной или ее формальной замены при написании на другой более краткий или долгий вариант. Поскольку часто в разбираемых источниках слова пишутся не так, как в словарях, это приводит к возникновению нетипичных сочетаний ударений. Например, так появляется циркумфлекс над о или над ε или два и больше ударения в слове. Эти особенности также маркируются отдельно для удобства поиска следующими тегами: orph (для отличной от словаря орфографии) и n1ac (no one accent, для нескольких ударений в слове). Так достигается основная задача разметки — учесть особенности написания каждого слова и при этом в сводном глоссарии представить эти особенности в зависимости от системы акцентуации. Сам глоссарий строится как дерево, в котором от унифицированной по словарю словоформы идут ветви по выбранной системе акцентуации. Для удобства анализа, создания системы поиска и дальнейшего использования при создании базы данных или обучения нейронной сети распознаванию и распределению фрагментов текстов, каждая словоформа представлена в виде картинка, вырезанной из конкретного памятника. Такая картинка сопровождается комментариями: 1) лингвистическим; 2) палеографическим; 3) акцентологическим со всеми необходимыми для дальнейшей работы тегами. Примеры слов из глоссария, вырезанных из афонских рукописей, представлены на рисунках 6-9:

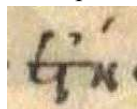


Рис. 6: *ετη*

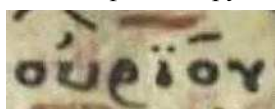


Рис. 7: *οὐρῖου*

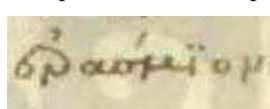


Рис. 8: *ἐράσμιον*

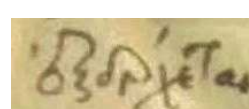


Рис. 9: *ἐξέρχεται*

Такая разметка показывает, что в некоторых случаях акцентуация копируется вместе с орфографией. Можно сказать, что ряд слов имели устойчивое написание по этим двум показателям (акцентуация и орфография) в определенные периоды или в конкретных регионах. В случае регионов мы можем

отнести эти особенности к письменной фиксации диалектного произношения. Встречаются также примеры акцентуации, которые объясняются фонологически, как произошедшие под влиянием произношения. Среди них появление густого придыхания внутри слова на стыке двух гласных, после дифтонга или гласной над следующей гласной, чтобы избежать зияния, например, *δέησιν* или *παλαιόλογος*. Что, в свою очередь, помечается тегом *pron* (=pronunciation).

4.3 Подготовка к использованию алгоритма для полуавтоматической разметки и принципы кодирования текста

Анализ результатов распознавания надписей и фрагментов рукописей разными онлайн программами OCR для древнегреческого языка показал, что знаки акцентуации были распознаны неточно и часто унифицировались согласно принятым в византийской системе акцентуации правилам. Поэтому использование результатов этого распознавания в дальнейшей работе на данном этапе невозможно. Это привело к необходимости после выполненной вручную разметки ряда надписей создать алгоритм полуавтоматической разметки и потом проверить его работу на практике.

На данном этапе работы все тексты корпуса фиксируются в общем файле, разделенном по разделам: рукописи, папирусы, лапидарные надписи, фресковые надписи, надписи на предметах прикладного искусства, граффити, печати. К каждой надписи составляется отдельный учетный лист (рис. 10), где представлены все акцентуированные слова по строкам, и к каждому слову указаны соответствующие теги.

расшифровка	прав. написание	перевод	формула	место находки	дата	знаки	граммат.	сдвиг	дифтонг	без трема	соответ. типа	орф. тип	акцентуации
ΑΙΣΘΗΣΙΝ	αἰσθησίν	αἰσθησίν	чувство	Феры	1104	sb, a, t	ntg	sbsb_l, nash	d1sb/d2a	nfr	tt		alex
ΑΥΤΑ	αὐτά	αὐτά	эти	Феры	1104	sb, g	ntg	sbsb_l, ngsh	d1		tt		alex
ΔΕΣΠΟΝΗ	δεσπότην	δεσπότην	госпожина	Феры	1104	a	ntg	ash_lb			tt		alex, l-or
ΕΜΠΙΚΡΑΙΝΩΝ	ἐμπικραίνων	ἐμπικραίνων	рассердившись	Феры	1104	sb, t, a	ntg	ash_lb	db	nfr	tt		alex, l-or
ΕΥΘΑΛΕΣ	εὐθαλές	εὐθαλές	пышно цветущий	Феры	1104	sb, a	ntg	rbsh_l, ash_rb	d1		tt		byz
ΘΗΣΑΥΡΙΑΙΣ	θησαυρίσους	θησαυρίσους	сохраненным	Феры	1104	ta	tg	nsh		nfr	tt		byz
ΚΑΙ	καί	καί	и	Феры	1104	g	ntg	gsh_lb	d1b		tt		alex, l-or
ΚΑΤΑΦΥΓΕΥΣΑΙΣ	καταφυτεύσους	καταφυτεύσους	обсаженным	Феры	1104	a	tg	nsh	d2		tt		byz
ΠΑΛΙΝ	πάλιν	πάλιν	обратно	Феры	1104	cir, t	ntg	cirsh_r		nfr	ntt		byz, r-or
ΤΩΝ	τῶν	τῶν	артиклъ В.п.	Феры	1104	g	ntg	ash_r			tt		r-or
ΤΟΥΤΟΝ	τούτον	τούτον	вот этого	Феры	1104	a	ntg	ash_l	d1		tt		r-or
ΤΡΟΦΗΣ	τροφής	τροφής	роскошь	Феры	1104	cir, t	tg	nsh			ntt		byz

Рис. 10: Надпись из Адрианополя 976-1064 гг.

Из таких учетных листов, собранных вместе и ранжированных по алфавиту, собирается общий глоссарий. Приведем фрагмент, сделанный на основе надписей разного типа для лексемы «έτος» (рис. 11):

расшифровка	прав. написание	перевод	формула	место находки	дата	материал	знаки	граммат.	сдвиг	дифтонг	бс	т	ор	соответ.	типа	тип акцентуации	жанр	тип текста	в формуле	название файла	публикация
	ἐτ' ἔτος	года	V'SrbCa(d)C	Адрианополь	976-1064	мрамор	rbl. a2	ntg	ash_r_c	nd				rbnnt, att	r-or		ded	pr	да	DSC02233adrianopol	Asdracha 108
	ἐτ	ἐτ(ουc)	года	aC'VS	Афины	1062	граффито	a/rb	ntg	a/rbsh_l	nd			att/rbnnt	l-or		f	pr	да	DSC00413	A. Капустин 11, №17
	ετ'	ἐτ(ουc)	года	C'VSa	Италия	1083-1084	мрамор	a	ntg	ash_r	nd			att	byz+r-or		bu	pr	да	DSC03248ital	Guillou 131
	ἐτ	ἐτ(ουc)	года	sbaVS(d)C	Комотини	4 февраля 1111	мрамор	arb	ntg	nsh	nd			rbnnt, att	byz+l-or		f	pr	да	DSC08791	Asdracha 63
	ἐτος	ἐτος	года	rbVSaCVSC	Италия	1131-1132	мрамор	arb	ntg	rbsh_l, nash	nd			rbnnt, att	byz+l-or		f	pr	да	DSC03221ital	Guillou 112
	ετ'	ἐτ(ουc)	года	C'VSa	Афины	1153	граффито	a	ntg	nsh	nd			tt	byz		f	pr	да	DSC02388athen	Orlandos 39
	ετ'	ἐτ(ουc)	года	C'VSa	Афины	1160	граффито	a	ntg	ash_r	nd			tt	byz		f	pr	да	DSC02389athen	Orlandos 40
	ἐτ	ἐτ(ουc)	года	rbaC'VS	Италия	7 августа 1186	мрамор	asb	ntg	sbsl_l+ash_l	nd			rbnnt, att	byz+l-or		f	po	да	DSC03212ital	Guillou 104
	ἐτόc	ἐτους	года	sbVS/C(d)C	Наксос	1309	фреска	sb	ntg	sbsl_l	nd			tt	byz+l-or		bu/f	pr	да	DSC08594	Μουτσopoulos 20.1
	ε'τ'(ουc)	ἐτ(ουc)	года	V'Sa(CdC)	Афины	1314	граффито	a	ntg	ash_r	nd			tt	r-or		f	pr	да	DSC02385athen	Orlandos 26
	ἐτ(ουc)	ἐτ(ουc)	года	sbaVS/C(d)C	Трикалаон, Метеора	1548	фреска	arb	tg	nsh	nd			rbnnt, att	byz		bu	pr	да	DSC08594	Μουτσopoulos 22.3
	ἐτα	ἐτα	года	V'SrbCa(d)	Трикалаон, Каламба	1766	фреска	asb	tg	sbsl_l, nash	nd			tt	byz+l-or		bu/f	pr	да	DSC08595	Μουτσopoulos 24.2

Рис. 11: Фрагмент подготовительного файла для глоссария с лексемой ἔτος

При этом к каждой лексеме делается сопроводительный формульный лист (табл. 1), составленный согласно византийской системе акцентуации для удобства анализа орфографических вариаций:

нач. форма / В.п. ед.ч.	перевод	формула	Р.п. ед.ч	формула	Д.п. ед.ч	формула	Им.п. мн.ч./ В.п. мн.ч.	формула	Р.п. мн.ч.	формула	Д.п. мн.ч.	формула
ἔτος	год	VSsbaCVSC	ἔτεος	VSsbaCVSVSC	ἔτεϊ	VSsbaCVSt	ἔτη	VSsbaCVL	ἔτων	VSsbCVLcirC	ἔτεσι	VSsbaCVSCV
			ἔτους	VSsbaCdC								

Табл. 1: Формульный лист для ἔτος

Для обработки нового материала используется следующий порядок действий:

Сначала ставится тег по названию знака, который использован: a=acute (´), g=gravis (˘), cir = circumflex (ˆ), t=trema (¨), rb = rough breathing (ˊ), sb = smooth breathing (ˋ). Для определения знака в дальнейшем с использованием нейросети для каждого из них собраны графические варианты с учетом всех палеографических особенностей. То же сделано и для случаев соединения знаков, которые прописываются отдельно, например, rba, sba, ta, tcig и т.д. В случае, если над одной лексемой присутствуют два знака на разных слогах, то проставляются теги по числу вхождений и, если они одинаковые, то для определения их порядка в слове им присваивается номер от 1 до 5. Если циркумфлекс оказывается над о или ε, то ставится дополнительный тег ntg. Такой же тег добавляется в случаях, когда придыхание зафиксировано не на первой гласной в слове, если только это не дифтонг. На втором шаге определяется местоположение знака. Если ровно над гласной, то ставится тег nsh (no shift), если нет, то маркируется сдвиг вправо sh_r или влево sh_l. При сдвиге цифра рядом с тегом показывает на сколько элементов сдвинуто: 1 — на предыдущую или последующую букву, 2 — на пробел дальше и 3 — на букву через одну. Если знак попадает между буквами, то добавляется тег b (=between). В случае, если знак длинный и попадает на две буквы, это маркируется отдельно l (=long), а если точкой, то p (=point). Для обучения нейронной сети автоматическому определению, находится ли знак над гласной или нет, собран из размеченного вручную материала отдельный корпус примеров для каждой гласной и каждого типа гласных, иллюстрирующий, когда знак ровно над гласной, а когда нет. Все случаи со сдвигами маркируются также тегом ntg.

На следующем этапе предлагается все слова приводить в базе к схематичному виду, например, для ἀνέκενϑοῦ — sbVCV'SaCVSVCV'LCCVt, где VS - краткий гласный (ε, ο), VL — долгий гласный (η, ω), V — гласный, который может быть и кратким, и долгим (α, ι, υ), C — согласный. Все буквы заданы списками с вариантами их написания по палеографическим схемам (пример таковых, [21]). Маленькими буквами обозначены акцентные знаки, зафиксированные после того знака, к которому

относятся. Если придыхание оказывается до гласного, то тег пишется до гласного. Если знак сдвинут и попадает между буквами, то в схеме после него ставится тег *b*. В случае горизонтальных лигатур (рис. 12) буквы, в них включенные, пишутся в строчном виде *Cv*, где первый знак с большой буквы, а второй с маленькой и в случае вертикальных лигатур (рис. 13) *C/C*, где первой пишется верхняя буква. Ряд классических лигатур для дифтонгов, как δ — *ov*, сразу записываются в виде *(d)*. В фигурных скобках пишутся принятые сокращения с их расшифровкой, например, вместо $\overset{\circ}{\circ} = \delta v$ {*VgC*}.

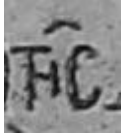


Рис. 12 пример горизонтальной лигатуры

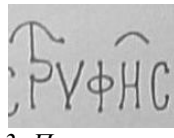


Рис. 13: Пример вертикальной лигатуры



Рис. 14: Пример для кодирования с дифтонгом и лигатурой

Если в слове две последовательные гласные, то определяется дифтонги они или нет. Для этого проверяется, является ли вторая гласная *υ* или *ι* и написана ли она без тремы. В случае дифтонга под ударением маркируется на какой элемент дифтонга падает ударение и ставятся соответствующие теги *d1* и *d2*. Если ударение оказывается между, то пишется тег *db*. Что приводит к корректировке схемы, если она изначально была взята не из ручной разметки, где эти теги проставляются сразу, например, для слова $\acute{\upsilon}\lambda\acute{\alpha}\tau\omicron\upsilon$ (см. рис. 14) *VrbCVaC/(VSV)* переходит в *VrbCVaC/(d)*.

После анализа состава слова по описанным выше схемам определяется корректно ли согласно грамматике и словарям место ударения, соответствует ли слово своему акцентному типу и акцентной парадигме. Для удобства определения можно использовать правила, описанные Тронским И.М. [30, 49-50] и представленные через введенные выше принципы кодирования. Например, для акута — окситоны (ударение на последнем слоге): *VCVCVSAc*, *CVCCVSCVLaC*, парокситоны (ударение на предпоследнем слоге): *d/VCd/VLaCVLC*, *VCVSAcCd*, *CVSCVaCVS*, пропарокситоны (ударение на 3 слоге от конца слова): *VaCCVLCVSC*, *CVSAcVSCVSC*, *CVSCVLaCVCV*. для циркумфлекса — периспомены (ударение на последнем слоге): *VCVSCVLcīrC*, пропериспомены (ударение на предпоследнем слоге): *VCVLCdcīrVSC*. для грависа — *CVSCVSG*, *VCVCVSGC*, *CVCCVSCVLgC*. Дополнить возможные схемы можно, используя файл, сделанный из словаря древнегреческого языка [8] путем кодировки по нашей системе, где один столбец слово, другой его код. При необходимости в третий столбец можно добавить морфологические показатели по классификатору. Тогда этот файл можно будет использовать как проверочный для идентификации возникших в слове орфографических вариантов.

Если после слова стоит энклитика и с нее переносится ударение, то это маркируется отдельно тегом *en* после всех тегов этого ударения. Это необходимо, чтобы разграничить случаи, когда в слове появлялось два и более ударения на безударных слогах, как было принято в александрийской системе акцентуации.

В случае наличия тремы маркируется тегом *fr*, если ее положение соответствует ее функциям, а если оно фонетически не обусловлено, то тегом *nfr*.

В случае наличия традиционных сокращений слов и ударения на них, ставится тег *abr*.

Для простановки тегов, фиксирующих типы систем акцентуации, используются следующие принципы:

alex — в случаях, когда в итоговой схеме гравис оказывается не на последнем слоге тег *gntg* или сочетание тегов *g*, *ntg*, *gsh_1l*. А также фиксируется ударение на первом элементе дифтонга *d1*, или тег *sh_1*, или слово содержит несколько ударений и тег *n1ac*, но за ним нет энклитики и как следствие тега *en*.

byz ставится, если есть теги *nsh*, *d2*, *tt*, *tg*.

r-or — если в графе со сдвигом находится тег, содержащий в своем составе сочетания *sh_r*.

l-or — если в графе со сдвигом находится тег, содержащий в своем составе сочетания *sh_1l*.

l-r-or — если в графе со сдвигом находится тег, содержащий в своем составе сочетания *sh_1l* и *sh_r*.

log — при маркировке ударением сокращений и наличия тега *abr*.

4.4 Одинаковые тексты и их акцентуация

Греческие надписи, особенно, на фресках часто содержат фрагменты одинакового текста. Это могут быть цитаты из литургических текстов или формульные словосочетания или фразы. Отдельно при разметке выявляются одинаковые тексты и их фрагменты в разных регионах, поскольку анализ в

рамках византийских надписей Грузии показал, что даже в одном регионе одинаковый текст может акцентуироваться по-разному. Рассмотрим на примере трех надписей (рис. 15, 16 и 17) из Грузии [5,6] и Касторьи [10]:



Рис. 15: Грузия XIVв.

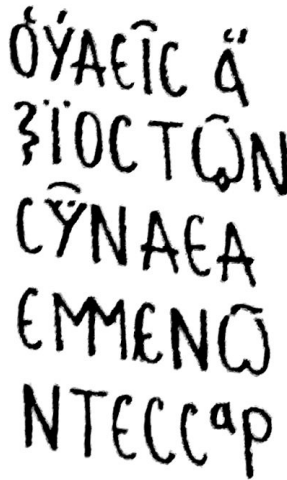


Рис. 16: Грузия XIVв.

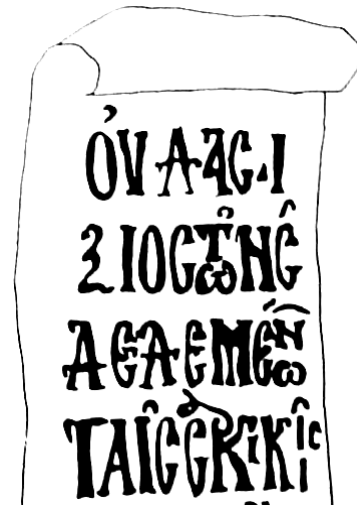


Рис. 17: Касторья 1432 г.

Близкие по времени три варианта акцентуации цитаты из херувимской песни: «Οὐδείς ἄξιός τῶν συνδεδεμένων ταῖς σαρκικαῖς» (d2sbCd2gC VsbaCVVSC CVLcirC CVCCVSCVSCV SaCVLC Cd2cirC CVCCVCd2cirC) показывают, что при разнице орфографии везде использованы элементы александрийской системы, но в разных местах. В таблице с формулами для каждой надписи подчеркнуты места, которые можно интерпретировать, как александрийскую систему (табл. 2):

Грузия XIVв.	Грузия XIVв.	Касторья 1432 г.
<u>d1rbb</u> Cd <u>Cg</u> VrbaCVLVSC CVScirbC C(VL <u>Cg</u>)CVSCVS()	<u>d1rb/d2a</u> Cd2cirC VrbaCVtVSC CVLcirC <u>CVcirt</u> CCVSVS(CC)VSC <u>VLcir</u> C CVSC CVC()	<u>d1sb</u> C(d)C VCVVSC C/VLcirC C()CVSCVSCV SaC/VL(C) Cd2cirbC V/C(CC)VC(<u>d1</u>)cirC

Табл. 2: Кодировка надписей, цитирующих одинаковый текст.

Выявление таких формульных обозначений для характерных элементов разных систем акцентуации позволит их в дальнейшем выделять автоматически.

4.5 Синонимия знаков акцентуации или проблема наложения значения

По данным разных памятников в процессе разметки было выявлено, что некоторые знаки акцентуации использовались в нескольких значениях одновременно, независимо от ведущей системы акцентуации [26]. Этот факт позволяет нам говорить применительно к знакам акцентуации о таких явлениях как синонимия, омонимия и других применимых традиционно к лексике понятиях. В таких случаях при разметке теги приводятся через «/», что потом позволит учитывать оба варианта и определить самые частотные сочетания. Например, волнистая линия могла быть циркумфлексом рис. 16 во второй строке или знаком сокращения, как на рис. 17 в третьей строке, а также знаком переноса (тег hy (=hyphen) [9, 26]), как на том же рис. 16 в 4 строке. При этом последний можно интерпретировать и как циркумфлекс на безударном слоге, что можно закодировать так: CVcirtCCVSVS(CC)VSCVLcir/hyC.

[5] Выводы

На данном этапе исследования для разработки системы тегирования и кодирования были вручную размечены надписи из Каппадокии, как граффити, так и фресковые, надписи из Грузии, надписи на металле из Италии, граффити из Киева, Константинополя, Афин, Кипра, Крита, надписи на фресках из Кипра, Северной Греции и ряд надписей из других регионов для сравнения. В результате было выявлено, что использование систем акцентуации в разных регионах обусловлено не только

материалом и типом надписи, но господствующим диалектом. В Каппадокии такой характеристикой стала особая традиция использования тремы, частые ударения на сокращениях, как элемент смысловой системы. Анализ подборки надписей на металле из Италии показал, что в них часто используется александрийская система акцентуации и различные вариации сочетаний систем со сдвигом вправо и влево, как в сфрагистических памятниках. В ряде случаев наблюдается ошибочный выбор знака или их путаница по палеографическим причинам, когда знак по начертанию похож на точку, что нами также было зафиксировано и на материале печатей. Грузинские надписи, цитирующие литургические тексты, предпочтительно используют элементы александрийской системы акцентуации. При этом не наблюдается связи между сохранением правильной орфографии и использованием правильной, с точки зрения византийской системы, акцентуации. Однако, некоторые частотные слова сохраняли свою акцентуацию независимо от других факторов.

Применение предложенного типа кодирования для остального корпуса позволит проверить, насколько бытовавшие в разное время варианты написания слов отклоняются от описанных ранее в исследованиях акцентных парадигм. Как влияет на эти отклонения выбранная автором система акцентуации и можно ли говорить о влиянии орфографических вариаций, в целом, на изменения в акцентуации. Последовательное кодирование памятников позволит автоматически тегировать источники по системам акцентуации и создать систему корректировки акцентуации в процессе распознавания такого рода текстов.

Предполагается в дальнейшем написать на основе сформулированных принципов тегирования программу, которая бы осуществляла частичную автоматическую разметку подобных текстов. Результаты этой разметки после верификации также можно будет использовать для обучения нейронных сетей классификации текстов по их системам акцентуации и распределению по акцентным парадигмам. Последнее применимо не только для византийского греческого, но и для других языков и письменных памятников, содержащих акцентные знаки. Что было нами выборочно проверено на материале славянских языков, представленного в книге Дыбо В.А. [16].

Acknowledgements

Выражаю огромную благодарность своим коллегам: Дыбо В.А., Казанскому Н.Н., Афиногенову Д.Е., Степановой Е.В., Степаненко В.П., Алексеенко Н.А., Б.Л. Фонкичу, Д.И. Коломацкому, которые поддерживали меня на разных стадиях работы над корпусом.

References

- [1] Biondi A. Gli accenti nei papyri greci biblici. — Roma: Papyrologica Castrociquitana, 1983.
- [2] Hinge G. Der dorische Akzent in den hellenistischen Korpora der archaischen Meliker. — 2001–2006 — Access mode: <http://alkman.glossa.dk/akzent.html>.
- [3] Fournet Jean-Luc. Hellénisme dans l'Égypte du Vie siècle. La bibliothèque et l'oeuvre de Dioscore d'Aphrodité MIFAO 115/1 Caire: 1999. P. 22.
- [4] Fournet Jean-Luc. L'influence des usages littéraires sur l'écriture des documents : perspectives, Proceedings of the 20th International Congress of Papyrologists. Copenhagen: 1994. P. 418-422.
- [5] Kauhchischvili T. Korpus der griechischen Inschriften in Georgien — Tbilisi: 1999 — Vol.1 (on Georgian).
- [6] Kauhchischvili T. Korpus der griechischen Inschriften in Georgien — Tbilisi: 2002 — Vol. 2-3 (on Georgian).
- [7] Laum B. Das Alexandrinische Akzentuationssystem unter Zugrundelegung der theoretischen Lehren der Grammatiker und mit Heranziehung der praktischen Verwendung in den Papyri. — Paderborn: Druck und Verlag von Ferdinand Schoeningh, 1928.
- [8] Liddell H.G., Scott R. A Greek-English Lexicon. Oxford: Clarendon Press, 1940. (https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3AAalphabetic+letter%3D*a%3Aentry+group%3D1%3Aentry%3D*a)
- [9] Murphy D. (1995) Hyphens in Greek Manuscripts, Greek, Roman, and Byzantine Studies, 36, pp. 293-314.
- [10] Mutsopulu N. K. Contribution a la morphologie de l'écriture hellénique: album d'inscriptions byzantines et post-byzantines. — Thessaloniki: 1977.
- [11] Reggiani N. Digital Papyrology I. Methods, Tools and Trends. — Berlin/Boston: 2017.
- [12] Reggiani N. Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri. — Berlin/Boston: 2018.
- [13] Seider R. Paläographie der griechischen Papyri. — Stuttgart: Anton Hiersmann, 1967. — Bd. 1.
- [14] Ševčenko I. (1981) Levels of style in Byzantine prose, Jahrbuch der Österreichischen Byzantinistik 31.1. P. 289–312.
- [15] Turner E. G. Greek manuscripts of ancient world. — Oxford: Clarendon Press, 1971.

- [16] Дыбо В.А., Замятина Г.И., Николаев С.Л. Основы славянской акцентологии. — Москва: Наука, 1990.
- [17] Евдокимова А.А. Языковые особенности греческих граффити Софии Киевской. — Москва: 2008 (диссертация).
- [18] Евдокимова А.А. Александрийская и византийская системы акцентуации в греческих акцентуированных надписях провинциального происхождения. // Индоевропейское языкознание и классическая филология 13. — Санкт-Петербург: 2009, с. 164–171.
- [19] Евдокимова А.А. Акцентуированные папирусы из коллекции Кельнского университета (предварительные замечания). // Индоевропейское языкознание и классическая филология 14. — Санкт-Петербург: 2010, с. 277–295.
- [20] Евдокимова А. А. Системы акцентуации греческих папирусов и Парфений Алкмана. // Аристей. — Москва: 2010, с. 41–52.
- [21] Евдокимова А.А. Палеография греческих граффити Софии Константинопольской, Софии Киевской и Софии Новгородской: сравнительный анализ. // Аристей VII. — Москва: 2013, с. 132–180
- [22] Евдокимова А.А. Диалог византийской и александрийской систем акцентуации в греческих граффити из разных балканских памятников. // Балканские чтения 15. — Москва: 2019, с. 31–38.
- [23] Евдокимова А.А. Системы акцентуации в легендах византийских моливдовулов XI–XII вв. Из коллекции Dumbarton Oaks. // В XII международной византийский семинар ΧΕΡΣΟΝΟΣ ΘΕΜΑΤΑ: Империя и полис. — Симферополь: 2021, с. 111–126.
- [24] Евдокимова А.А. Системы акцентуации в греческих рукописях из коллекции Афона. Ч. I. // Индоевропейское языкознание и классическая филология 25.1. — Санкт-Петербург: 2021, с. 341–368.
- [25] Евдокимова А.А. Системы акцентуации в византийских надписях на фресках богородичных церквей Троодоса (Кипр) // XIII международный византийский семинар ΧΕΡΣΟΝΟΣ ΘΕΜΑΤΑ: Империя и полис. — Симферополь: 2022, 83-92.
- [26] Евдокимова А.А. Синонимия акцентных знаков или знак переноса (ἐνωτικόν) в византийских надписях // Индоевропейское языкознание и классическая филология 26.1 — Санкт-Петербург: 2022, с. 387-401.
- [27] Казанский Н.Н. Между письменным и устным текстом: древнегреческое «наивное» письмо. // Исследования по языкознанию: к 70-летию члена-корреспондента РАН А.В. Бондарко. — Санкт-Петербург: 2001, с. 246–256.
- [28] Казанский Н.Н. Вариативность диалектной нормы и история древнегреческого языка. // Язык, литература, эпос: к 100-летию со дня рождения академика В. М. Жирмунского. — Санкт-Петербург: 2001, с. 82–99.
- [29] Кисилиер М.Л. Анализ средневековых текстов и социолингвистические факторы (на материале «Луга духовного» Иоанна Мосха) // Юрьевские чтения: Материалы междисциплинарной конференции молодых филологов. Выпуск II. — Санкт-Петербург: 2002, с. 41–48.
- [30] Тронский И. М. Древнегреческое ударение — Москва, Санкт-Петербург: 1962.

Exploring Evaluation Techniques in Controlled Text Generation: A Comparative Study of Semantics and Sentiment in ruGPT3large-Generated and Human-Written Movie Reviews

Anastasia Margolina
HSE University
Saint-Petersburg
avmargolina@edu.hse.ru

Anastasia Kolmogorova
HSE University
Saint-Petersburg
akolmogorova@hse.ru

Abstract

The paper describes the proposed strategy for evaluation controlled text generation with the sentiment as attribute. Our approach mainly consists of automatic sentiment analysis (ruBERT) and topic modelling (BERTopic), which are applied to a parallel corpus with artificially produced and human-written texts. The model for evaluation is fine-tuned on the parsed reviews from big Russian movie-related website ruGPT3Large with the sentiment as prompt. The results of the analysis demonstrate that the proposed methods can offer a more comprehensive understanding of the advantages and limitations in the context of semantics and sentiment. Additionally, the paper employs metrics such as BERTscore and self-BLEU to further evaluate the generated text. The proposed methodology provides a novel approach for evaluating the quality of generated text and may have implications for future studies in the field.

Keywords: controlled text generation, strategy for quality measurement, topic modelling, sentiment-analysis, movie reviews.

DOI: 10.28995/2075-7182-2023-22-1082-1090

Оценка контролируемой генерации текста: сравнительное исследование семантики и сентимента в отзывах на фильмы, написанных ruGPT3large и человеком

Марголина А.В.
НИУ ВШЭ
Санкт-Петербург
avmargolina@edu.hse.ru

Колмогорова А.В.
НИУ ВШЭ
Санкт-Петербург
akolmogorova@hse.ru

Аннотация

В статье предлагается новая стратегия оценки контролируемой генерации текста с тональностью в качестве атрибута. Наш подход включает автоматический анализ тональности (ruBERT) и тематическое моделирование (BERTopic). Эти инструменты применяются к параллельному корпусу, состоящему из пар “сгенерированный отзыв – реальный отзыв”. Модель используемая для оценки – ruGPT3Large, которая была ранее дообучена на собранных с Кинопоиска отзывах на фильмы с тональностью “вшитой” в заголовке. Результаты анализа демонстрируют, что использованные методы предлагают более полное понимание преимуществ и ограничений в контексте семантики и эмоциональной окраски языковой модели. Кроме того, в статье применяются такие метрики, как BERTscore и self-BLEU, для дополнительной оценки сгенерированного текста. Наша методология представляет новый подход для оценки качества генерируемого текста и может дать основу для будущих исследований в этой области. Ключевые слова: контролируемая генерация текста, стратегия оценки качества, тематическое моделирование, сентимент-анализ, кинорецензии.

Ключевые слова: контролируемая генерация текста, стратегия измерения качества, тематическое моделирование, анализ тональности, отзывы на кино

1 Introduction

In this paper we tackle the problem of controlled generation of text in Russian. Our experiments concern such a text genre as movie reviews and the attribute we initially control while generating is text sentiment.

To discuss the problem a few challenges, need to be consequently addressed. First of them - how to evaluate the quality of generation. Despite the exponential growth in the number of pre-trained generative language models (LMs), the problem of accurate metrics for measuring generated text quality persists. There are no studies that aim to explore the artificially made texts, although it could potentially reveal unseen differences and similarities between ‘made-up’ texts and the ‘actual’ ones and it could be used as a peculiar metric for the evaluation of the semantic quality of generated texts. This is linked to a certain limitation: to make such research happen one needs a generative model, which is fine-tuned on downstream tasks, and the dataset of real texts that can be directly compared to the dataset of produced data.

In our case, we analyse the effectiveness of two unsupervised metrics (BERTscore and Self-BLEU) and display the results of our experiments when applying Topic Modelling (TM) and Sentiment Analysis methods to compare sentences in two parallel corpora of movie reviews in Russian having the same prompts: written by human users and generated by fine-tuned ruGPT3Large model.

We focus on differences between human generated and AI generated texts of a specific genre. In this context, TM is not only the tool for linguistic research of the overall structure of movie reviews but it is also a strategy for evaluation how well, comparatively to humans, does the model construct the narrative. The Sentiment Analysis use is sanctioned by our desire to compare not the correctness of the label assigned to text by machine, but to verify its adequateness to human subjective expression in analogous text.

Our hypothesis is formulated as following:

1. More discrepancies we observe between topics having the most important weights in human-written texts and AI-generated texts, less qualitative is the controlled generation. To assess the degree of deference in topics we use values of Cosine similarity distance between vectorized representations of topics.
2. The higher is the difference between accuracy values returned by classification when estimating the sentiment in two parallel corpora, the less qualitative is the controlled generation.

2 Related Papers

One of the main challenges in evaluating the quality of generated text is the lack of accurate metrics. It is caused by several factors. First, the evaluation is commonly conducted in a reference-free setting because it is challenging to collect sufficient high-quality references for each input of control variables in this open-ended text generation task (Dathathri et al., 2020). This led to the situation when the majority of existing metrics measures the similarity of generated text against human-written references. Such metrics can be classified into unsupervised, supervised, and human evaluation-based methods, each with its own limitations and advantages. To overcome the shortcomings of classical single-score BLEU (Papineni et al., 2002), researchers propose a family of interpretable metrics for the key aspects of diverse tasks (summarization, style transfer, and dialogue) which either don’t require human references (Deng et al., 2022) or can model human assessment with rather high accuracy (Sellam et al., 2020).

An interesting approach tested on Russian language data was suggested in (P. A. et al., 2022): within the RuATD Shared Task 2022 the authors propose to use binary classification methodology designed to detect AI-generated texts to filter well-generated texts (with the high number of false positives in generated texts classified as human written) from bad-generated.

The benchmarking platform to support research on open-domain text generation models Tegygen (Zhu et al., 2018) also provides several groups of metrics: Document Similarity based Metrics, Likelihood-based Metrics and Divergence based Metrics (in our experiments we use one of them too).

In contrast with the mentioned above metrics, we suggest two metrics, mostly qualitative, but having a quantitative support, to evaluate the semantic and emotional consistency of human-generated and AI-generated movie reviews in Russian using TM and Sentiment Analysis methods.

3 Data and Methodology

The parallel dataset consists of 1200 actual reviews and 1200 generated reviews. The reviews made by the model were generated according to prompt, which is 5-6 words in the beginning of the actual review and the corresponding sentiment, which is a controlling attribute for the text generation. The dataset is normalised in the context of the sentiment: 400 reviews for each (positive, neutral and negative).

Topic modelling is widely applied in exploratory analysis as a tool for extracting hidden semantic relationships, topics in the set of textual data. Many researchers use this method to analyse not only social data but also literature in order to find covert patterns (Schöch, 2016; Ordun et al., 2020; Sherstinova et al., 2022). Needless to say, all this data is human-written: either the author of some book or the dataset of short-texts from twitter.

In this research such an implementation of topic modelling towards generative language models is presented. We explore the semantic distance between two sets of movie reviews: one generated by ruGPT3Large and one with reviews written by the users. This paper employs BERTopic, an unsupervised topic modelling technique (Grootendorst, 2022), to conduct exploratory data analysis on two distinct datasets of movie reviews. The analysis is conducted in three steps. In the first step, we analyse a dataset of real reviews, identifying overarching trends and topics in web-reviews on films. In the second one, we apply BERTopic to a dataset of generated reviews, revealing the typical semantic net of artificially produced reviews. Finally, we compute a cosine similarity distance between the vectorised representations of topics.

Aside from semantic validation of generated text, the important goal is to investigate whether the model creates appropriate texts in terms of controlling attribute, the sentiment. There are no sustainable metrics for the evaluation of controlled text generation tasks except human assessment. Nevertheless, the psycholinguistic experiment is time and resource consuming. To address this issue, we propose the use of automatic sentiment analysis on the parallel corpora, comparing the given sentiment of texts that have the same prompt and claimed sentiment. For this task the ruBERT¹ was fine-tuned with the movie reviews dataset (60k reviews, 20k for each sentiment) for the multilabel classification task. The final model achieves the accuracy of 95 percent on the test data.

4 ruGPT3Large Fine-Tune

The Russian version of GPT – the ruGPT3Large model² was chosen for the experiment. The architecture of ruGPT-3 is similar to that of GPT-2: it is a decoder-only transformer-based model, which makes it perfect for text generation (Radford et al., 2019). The data for fine-tuning was collected on Russian-language movie-related website. Aside from the text, the sentiment of the review was also parsed in order to then make a sample less biased. The original dataset consists of 199k reviews (148k positive, 28k neutral and 21k negative) but it stratified for training: 60k total number of reviews, 20k for each sentiment.

Data was transformed from a csv table to a textual file with prompts for model input. This format uses line breaks to separate reviews and special characters to mark start (<s>) and end (</s>) of each string. The structured data looks as follows:

```
<s>Тональность: [позитивная, нейтральная или негативная]\nТекст: [текст отзыва]\n</s>
```

```
Translation: <s>Sentiment: [positive, neutral or negative]\nText: [the text of the review]\n</s>
```

The objective of incorporating reviews with prompt into the model is to facilitate the memorization of patterns by the ruGPT3. This is achieved by utilising the second segment of the prompt, which serves as a continuation that the model must generate, namely, the review itself. The data then was split with the ratio of 0.3/0.7 for test/train.

¹<https://huggingface.co/Tatyana/rubert-base-cased-sentiment-new>

²<https://github.com/ai-forever/ru-gpts>

Table 1 showcases the parameters selected for the fine-tuning of ruGPTLarge to accommodate GPU memory constraints. Opting for the minimal batch size, as the table indicates, enhances the stability of training at the expense of per-step computation efficiency (Li et al., 2022). The learning rate adheres to the default setting. These configurations allowed the large model to complete fine-tuning within a time frame of six and a half hours using the GPU.

Parameter	Value
num train epochs	1
per device train batch size	1
per device eval batch size	1
block size	1024
learning rate	2.5e-4

Table 1: ruGPT3Large finetuning parameters.

The perplexity of the fine-tuned model, trained on movie reviews, has been measured to be 19.

Table 2 presents the parameter of generation. We choose them in order to make the model generate more 'creative' and less repetitive (temperature, top k, repetition penalty).

Parameter	Value
repetition penalty	5.0
top p	0.95
top k	5
temperature	1
no repeat ngram size	2.0

Table 2: ruGPT3Large generation parameters.

The result is evaluated by two metrics: BERTscore (Zhang et al., 2020) and Self-BLEU (Zhu et al., 2018).

The first one computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings. In the context of my research this score shows how different the LM generates the output compared to the original text with the same prompt. The median BERTscore for the parallel dataset (original review – generated review) of 1200 equals 69.7 percent. It means that the model generates similar to the original, however it tends to create its own narratives. At Table 3 we show the example of our parallel corpora: the text column contains human written texts, and the result column contains AI-generated texts. As can be seen they both have the same prompts.

The BLEU score, a classical metric for text evaluation, is not suitable for our task. It is designed for text-to-text tasks like machine translation, measuring translation accuracy. However, our research focuses on creative text generation, where the output can be diverse. To assess text diversity, we utilize a metric called self-BLEU score.

This metric shows how diverse the output of the generated model is: 'a higher Self-BLEU score implies less diversity of the document, and more serious mode collapse of the GAN model' (Zhu et al., 2018, p. 4). This score is also calculated on the parallel dataset. To evaluate the quality of both types of texts, we computed the self-BLEU-3 metric for each review, which reflects more structural diversity, and calculated the mean values across the entire dataset. Our results, presented at Table 4, reveal that while the metric is slightly better for the neural-generated texts, both the human-written and generated texts exhibit similar levels of diversity.

This finding is surprising, given the longstanding perception that human-written texts are of higher quality than their machine-generated counterparts (Hardcastle and Scott, 2008). However, we attribute the small difference in favour of the generated texts to the parameters used in the generation process, namely the repetition penalty and the no-repeat n-gram size. These parameters are designed to penalise

sentiment	Generated reviews	Human reviews
нейтральная [neutral]	Лично я ожидал большего. Фильм явно не дотягивает до того уровня, на который претендует. [Personally, I expected more. The film is clearly not up to the level it's supposed to be.]	Лично я ожидал большего. Фильм явно не доработан, не хватает продуманного смысла. [I was expecting more of the same. The film is clearly incomplete, lacking thoughtful meaning.]
положительная [positive]	Перед нами довольно бесцветная и тревожная, но в то же время захватывающая история о том, как группа людей... [We have before us a rather colorless and disturbing, but at the same time gripping, story of how a group of people...]	Перед нами довольно бесцветная и тревожная, но тем временем, жизнеутверждающая история... [We are faced with a rather colorless and disturbing, but in the meantime, life-affirming story...]
отрицательная [negative]	Признаюсь фильм решила посмотреть из-за трейлера. Он меня очень впечатлил и я ожидала от него чего-то невероятного. [I admit that I decided to watch the movie because of the trailer. It really impressed me and I was expecting something incredible from it.]	Признаюсь фильм решила посмотреть из-за трейлера. Вторая ошибка моей жизни. [I admit that I decided to watch the movie because of the trailer. The second mistake of my life.]

Table 3: Example of our parallel corpora of texts with the same prompts.

the model for repeating words and sequences, thereby encouraging the model to produce more diverse texts.

To explore semantic features of generated texts, we suggest the strategy based on Topic Modelling and Sentiment Analysis methods.

	Human-written reviews	Artificial reviews
mean	0.074323	0.032231
max	0.020033	0.011562
min	0.209597	0.127813

Table 4: Self-BLEU metric applied to human-written and neural-generated texts.

5 Controlled Text Generation Evaluation

5.1 Topic Modeling

This study aims to compare the topics extracted from generated by artificial intelligence (AI) texts and human-written texts using the BERTopic algorithm with multilingual embedding model (Reimers and Gurevych, 2019). The results, presented in Figures 1 and 2, reveal notable differences in the topics discussed in these two sets of texts.

The topics showed in Figure 1 reveal the most frequent topics in our dataset of 1200 human-written movie reviews. Although most topics are not very interpretable, there are several clusters that can be

analyzed. Reviewers tend to focus their comments on various aspects of the movie’s production, such as the soundtrack, acting performances, and visual effects.

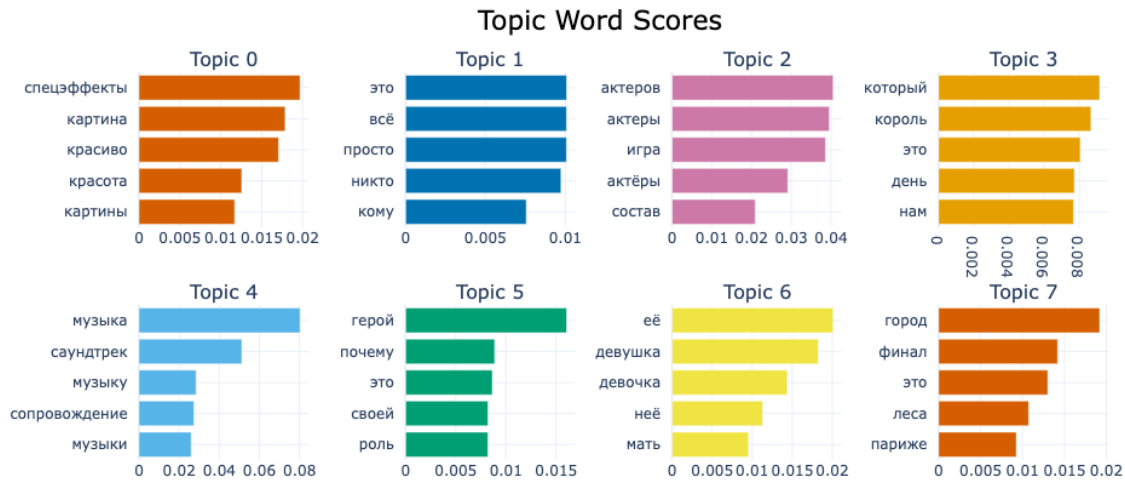


Figure 1: The most frequent topics in the dataset of human-written movie reviews.

We can also see that people tend to focus on the setting and main characters (Topics 1, 3, 6, and 7). However, the descriptions provided by reviewers are often vague and generalised, lacking details or named entities that could enhance nuance of their analysis.

On the other hand, the topics identified in AI-generated movie reviews focus on plot elements, action scenes, and character descriptions (Topic 1, 2 and 5). This suggests that the LM has a better grasp of narrative elements and character development. Interestingly, one named entity, Hans Zimmer, is present in the "musical" topic, which could indicate that the LM has prior knowledge of famous people in the domain.

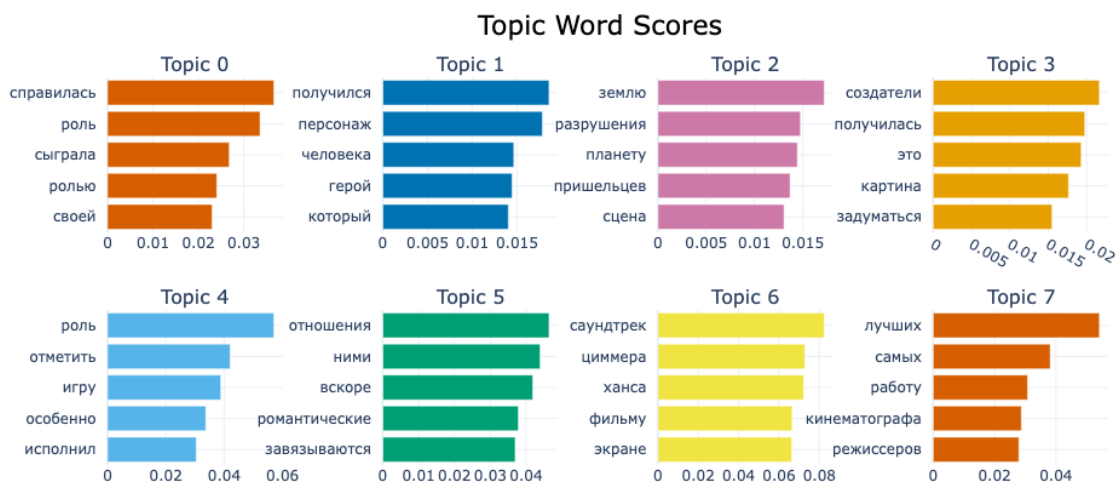


Figure 2: The most frequent topics in the dataset of AI-written movie reviews.

These findings highlight the strengths and limitations of both human and AI-generated movie reviews.

While human-written reviews are better at evaluating technical aspects of the movie’s production, AI-generated reviews excel in capturing plot and character details. Future research could explore ways to combine the strengths of both types to improve the quality of generated movie reviews.

We are interested in topic modelling not only as an exploratory data analysis tool but also as a metric for evaluating the similarity between two datasets’ topic distributions. To address it, we compute the Cosine similarity between vectorized representations of topics. This metric yields a value ranging from 0 to 1, where 0 denotes minimal similarity and 1 signifies complete identity. For our dataset, the cosine value is 0.56, indicating that while the topics are largely congruent, they also exhibit certain distinctions. Thus, we conclude that an ideal cosine similarity range lies between 0.50 and 0.70. Scores within this range indicate a balanced similarity level. A score of 1 would suggest overfitting, implying entirely identical token distributions, while a substantially lower score around 0 would suggest underfitting, indicating a lack of topic congruence and potentially suggesting that the topics are not related to movies.

5.2 Sentiment Analysis

The fine-tuned BERT is used to evaluate controlled attribute quality on a parallel corpus.

The classification achieves 74 percent accuracy on human data and 66 percent on neural data. This could indicate either inaccurate classification or issues with claimed sentiment. To investigate further, a confusion matrix is examined.

Figure 3 presents the confusion matrix for human-written review classification, with labels 0 (neutral), 1 (positive), and 2 (negative) sentiment.

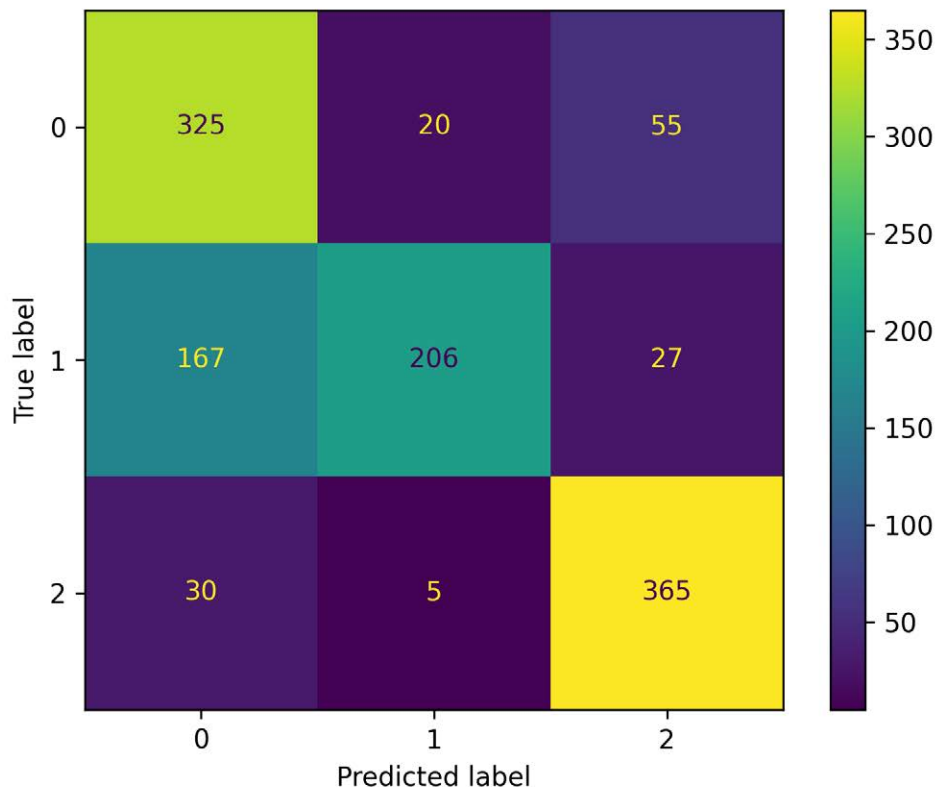


Figure 3: Confusion matrix for the results of human-written reviews classification by sentiment.

Confusion matrix analysis shows high accuracy for negative sentiment (365 out of 400 true negatives) but challenges in distinguishing positive sentiment. People’s tendency to use less explicit language in positive reviews creates a subtle and emotionless tone, leading to confusion with neutral sentiment.

Negative sentiment, requiring stronger conviction, is more clearly expressed.

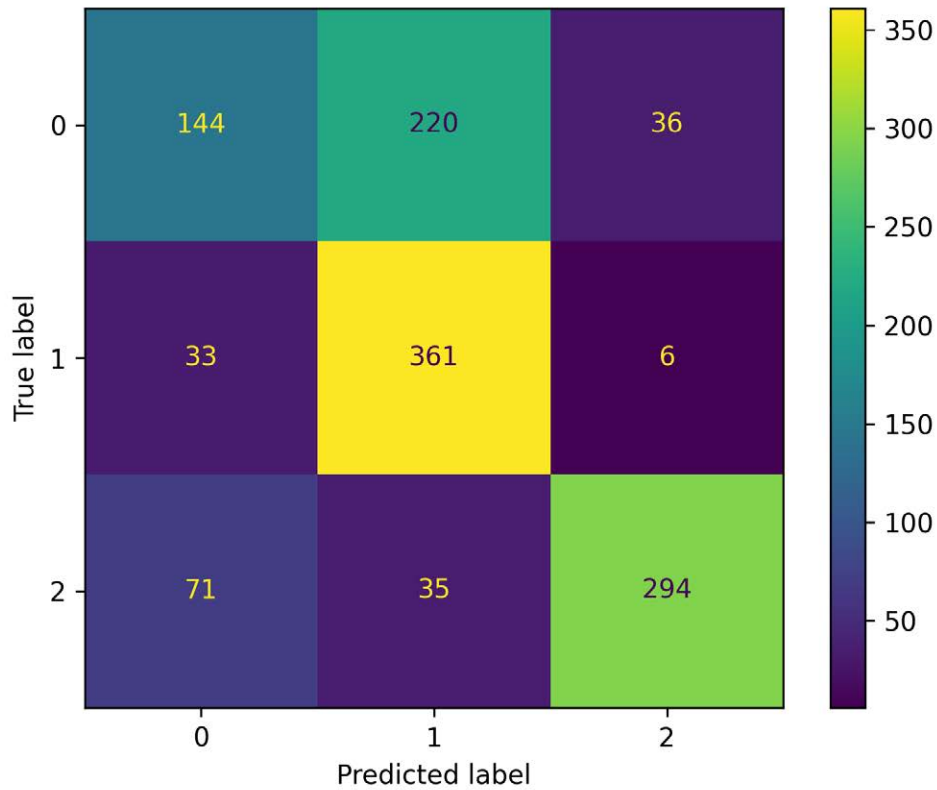


Figure 4: Confusion matrix for the results of AI-written reviews classification by sentiment.

The different distribution of true predicted labels is seen on the Figure 4, which displays confusion matrix of generated text by AI. In this case, the positive sentiment achieves the highest accuracy, while the neutral sentiment is the least accurate. This disparity provides valuable insights into the performance of the classifier. While Figure 3 might have suggested that BERT was biased towards neutral sentiment, the current findings indicate that the problem may be attributed to the nature of the texts themselves.

It is widely accepted that models may struggle to identify neutral sentiment. In this context, it is notable that our fine-tuned model appears to generate neutral sentiment less accurately than positive and negative sentiment.

Upon manually analyzing the neutral texts, we discovered a recurring pattern where many of them ended with the phrase "highly recommend to watch" or included a mention of "10 out of 10" ratings. This observation suggested that the model has a bias towards generating positive reviews even when the sentiment should have been neutral.

6 Conclusion

In this study, we have presented results of an approach to the validation of controlled text generation, which involves the use of popular natural language processing methods as reliable metrics to investigate the success of LM's generation. Our experiments showed the potential of Topic modeling and Sentiment Analysis tools to provide a deeper and more accurate estimation of the semantic consistency of generation validated on a parallel dataset that includes the controlled attribute, the original human-written text, and the generated text with the same beginning as in the original review.

Our approach has been implemented using the decoder's part (ruGPT3) of transformer architecture as a generative model and the encoder part (ruBERT) as a validation tool. Our findings offer important

insights into the structure of movie reviews in general.

Funding: This article is an output of a research project “Text as Big Data”: modeling of convergent processes in language and speech by digital methods” implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University) in 2023.

References

- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation, March. arXiv:1912.02164 [cs].
- Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2022. Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation, January. arXiv:2109.06379 [cs].
- Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March. arXiv:2203.05794 [cs].
- David Hardcastle and Donia Scott. 2008. Can we Evaluate the Quality of Generated Text? January.
- Conglong Li, Minjia Zhang, and Yuxiong He. 2022. The stability-efficiency dilemma: Investigating sequence length warmup for training gpt models.
- Catherine Ordun, Sanjay Purushotham, and Edward Raff. 2020. Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs, May. arXiv:2005.03082 [cs].
- Posokhov P. A., Skrylnikov S. S., and Makhnytina O. V. 2022. Artificial text detection in Russian language: a BERT-based Approach. // *Computational Linguistics and Intellectual Technologies*, P 470–476. RSUH, June.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. October.
- Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.
- Christof Schöch. 2016. Topic Modeling Genre: An Exploration Of French Classical And Enlightenment Drama. January.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation, May. arXiv:2004.04696 [cs].
- Tatiana Sherstinova, Anna Moskvina, Margarita Kirina, Irina Zavyalova, Asya Karysheva, Evgenia Kolpashchikova, Polina Maksimenko, and Alena Moskalenko. 2022. Topic Modeling of Literary Texts Using LDA: on the Influence of Linguistic Preprocessing on Model Interpretability. // *2022 31st Conference of Open Innovations Association (FRUCT)*, P 305–312.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT, February. arXiv:1904.09675 [cs].
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models, February. arXiv:1802.01886 [cs].

The Typological constructicon database

Nikita Muravyev
University of Hamburg
nikita.muraviev@gmail.com

Nikita Gordeev
NRU HSE (Moscow)
permanentdaylight@mail.ru

Илья Макаrchук
NRU HSE (Moscow)
ilya.makarchuk@gmail.com

Maria Kukushkina
NRU HSE (Moscow)
erinqa.hecil@gmail.com

Anton Buzanov
NRU HSE (Moscow)
anton.buzanov.00@gmail.com

Abstract

The report introduces a new resource: the Typological Constructicon database. This resource contains an inventory of constructions of selected semantic fields in a number of languages of different areal and genetic affiliation. The constructions are labeled according to a number of semantic and morphosyntactic parameters and provided with a detailed description and illustrative examples.

Keywords: constructicon, typology, constructions grammar, carative constructions

DOI: 10.28995/2075-7182-2023-22-1091-1096

База данных «Типологический конструктикон»

Никита Муравьев
Университет Гамбурга
nikita.muraviev@gmail.com

Никита Гордеев
НИУ ВШЭ (Москва)
permanentdaylight@mail.ru

Илья Макаrchук
НИУ ВШЭ (Москва)
ilya.makarchuk@gmail.com

Мария Кукушкина
НИУ ВШЭ (Москва)
erinqa.hecil@gmail.com

Антон Бузанов
Университет Гамбурга
anton.buzanov.00@gmail.com

Аннотация

В докладе представляется новый ресурс: база данных «Типологический конструктикон». Данный ресурс содержит инвентарь конструкций избранных семантических полей в ряде языков различной ареальной и генетической принадлежности. Конструкции размечены по ряду семантических и морфосинтаксических параметров и снабжены подробным описанием и иллюстрирующими примерами.

Ключевые слова: конструктикон, типология, грамматика конструкций, каритивные конструкции

1 Введение

В центре внимания лингвистической типологии всегда были преимущественно грамматические единицы языка, которые сопоставлялись между языками прежде всего, с опорой на описательные грамматики. В последние два десятилетия с развитием полевой и корпусной лингвистики

подобные исследования стали включать в себя первичные данные опросов носителей и текстов. Также активно развивается лексическая типология, в центре внимания которой находятся лексические значения и способы их выразить. Существует, однако, огромный пласт единиц, называемых конструкциями, которые изучаются в отдельных языках, но практически не попадают в сферу интересов типологов. Проект «Типологический конструктикон» ставит своей целью привнести типологическую перспективу в изучение конструкций. Разрабатываемая на материале ряда языков типологическая база данных представляет собой аннотированный и иллюстрированный примерами инвентарь конструкций, который позволяет изучать морфологические, синтаксические и семантические взаимосвязи между конструкциями как внутри языка, так и в межязыковом сопоставлении.

2 Грамматика конструкций

Конструкции являются предметом изучения одного из важнейших и наиболее активно развивающихся когнитивно-функциональных направлений лингвистики, Грамматики конструкций (Fillmore et al. 1988, Goldberg 1995). Конструкцией в рамках этого направления считается языковая единица, состоящая из формы и соответствующего ей значения. Грамматика конструкций как модель языка представляет собой структурированный инвентарь конструкций, охватывающий все уровни языка от морфологии до дискурса. Грамматика конструкций существует в русле когнитивно-функциональных узусно-ориентированных подходов к языку, поэтому активно использует данные корпусов. Вследствие своего промежуточного положения между лексикой и грамматикой конструкции довольно скудно представлены в существующих словарях и грамматических описаниях. Поэтому в последнее время для отдельных языков активно разрабатываются специальные лексикографические ресурсы, конструктиконы, ср. к примеру Русский конструктикон (Janda et al. 2018).

3 Типологический конструктикон

Типологический конструктикон представляет собой прежде всего научно-исследовательский ресурс. Он задуман как инструмент для типологов и исследователей конструкций, позволяющий на примере избранных семантических полей строить типологические и теоретические обобщения о структуре и функционировании языка, на материале, который до сих пор относился к периферии как лексической, так и грамматической семантики. Второе его предназначение – справочно-переводной ресурс для изучающих иностранные языки, так как содержит переводные эквиваленты конструкций с подробным описанием и примерами. Однако в отличие от существующих монолингвальных конструктиконов его основной задачей является не максимально полный охват существующих конструкций, а сбор данных по отдельным семантическим полям, наиболее насыщенным конструкциями.

Реляционная база данных PostgreSQL устроена таким образом, чтобы обеспечить представление морфологических и семантических особенностей данных. Между конструкциями из разных языков связи устанавливаются вручную. Мы предполагаем, что тщательная разметка способна обеспечить поиск похожих конструкций в разных языках. Кроме того, наша разметка позволяет искать конструкции, похожие по как по разному количеству параметров, так и по разным их комбинациям. Таким образом, в группе похожих конструкций могут оказаться не только синонимичные конструкции, но и конструкции устроенные одинаково с точки зрения морфосинтаксиса.

Интерфейс Типологического конструктикона включает возможности простого и расширенного поиска конструкций, а также общего обзора всех конструкций в рамках заданного семантического поля. Кроме того, сайт содержит общую информацию о проекте, его авторах, представленных семантических полях, а также инструкцию по пользованию ресурсом, релевантную библиографию и ссылки на другие ресурсы по конструкциям.

3.1 Простой поиск

Простой поиск представляет собой строку поискового запроса, в которую можно ввести конструкцию целиком или ключевое слово. В выдаче показывается список содержащихся в базе

данных конструкций, упорядоченных по релевантности, а также список наиболее релевантных аналогов найденных конструкций в языках, отличных от языка запроса. Гиперссылки в списке ведут на страницы соответствующих конструкций с подробной информацией о каждой, см. далее раздел 3.3.

3.2 Расширенный поиск

Расширенный поиск представляет более детальный поиск с применением различных фильтров, аналогично тому, как это устроено в корпусах (ср., к примеру, лексико-грамматический поиск в НКРЯ). Аннотация каждой конструкции включает в себя следующие параметры.

- запись конструкции в графике оригинала
- латинская транслитерация
- глоссирование
- якорные элементы и их значения
- информация о языке конструкции (семья, группа, язык)
- семантическое поле конструкции
- семантический класс и подкласс конструкции
- короткое описание значения
- примеры
- и др.

3.3 Обзор конструкции

При выборе конкретной конструкции из списка выдачи открывается страница со справочной информацией и примерами к данной конструкции. Набор характеристик соответствует параметрам расширенного поиска.

3.4 Обзор семантического поля

Отдельно от поиска на сайте Типологического конструктиконе имеется страница “Обзор”, содержащая данные о конструкциях в форме таблицы. При открытии данной страницы пользователю предлагается выбрать семантическое поле, класс и подкласс, а также языки, семьи и группы с интересующими конструкциями. В выдаче, помимо заданной пользователем информации о языке и семантике, предлагается список релевантных конструкций с переводом, глоссами, валентностной рамкой и информацией о деривационном источнике якорных слов. Страница обзора также доступна по гиперссылке с названием конкретного семантического поля в разделе “Семантическое поле” при описании конструкции.

4 Каритивные конструкции в конструктиконе

Разберем в качестве примера семантическое поле каритивов и их представление в Типологическом конструктиконе. Каритив представляет собой грамматическое значение отсутствия в ситуации ожидаемого участника, ‘без X’. Морфологически каритивная семантика обычно выражается при помощи специализированных предлогов/послелогов, как *без* в русском, *sans* во французском или аффиксов, как *-løs* в норвежском, *-li* в хантыйском. Каритивам посвящен ряд конкретно-языковых и типологических работ, ср. (Оскольская и др. 2020) и обзор литературы в статье. Помимо своей грамматической функции кодирования глагольной валентности на отсутствующего участника, каритивные показатели также входят в состав семантически нерегулярных конструкций, в которых исходное каритивное значение взаимодействует со значениями других компонентов, образуя новое значение. Такие конструкции относятся к целому ряду семантических фреймов, в терминологии Ч. Филлмора: лишения, лишения, уступки и избегания, которые будут проиллюстрированы далее.

В ресурсе представлены на данный момент преимущественно крупные языки Евразии: русский, английский, норвежский, французский, испанский языки, также в ближайшей перспективе будут добавлены немецкий, арабский, иврит и японский. В то же время, немаловажным представляется включение в базу также материала малых языков, находящихся под угрозой исчезновения. В настоящий момент в базу включены данные севернохантыйского языка, в ближайшее

время планируется сбор данных по северносаамскому языку. Материал планируется в дальнейшем расширять на другие языки как в Евразии, так и за её пределами. Источником материала служат полевые исследования авторов, онлайн опросники на платформе Google Forms и данные текстовых корпусов.

4.1 Краткий обзор каритивных конструкций

В семантическом поле каритивов можно выделить по меньшей мере четыре основных фрейма, вокруг которых сосредоточены встречающиеся в обсуждаемых языках конструкции: лишение, лишенность, уступка и избегание. Семантический фрейм **лишения** описывает динамическую ситуацию утраты сущности или свойства и состоит из двух участников: исходный обладатель утрачиваемой сущности и утрачиваемое обладаемое. Лишение может выражаться переходным глаголом, в котором обладатель заполняет позицию субъекта, а обладаемое — прямого объекта, как *X* потерял *Y*-асс в русском языке (1). Альтернативный способ — при помощи непереходной конструкции с предлогом, как *X s'est retrouvé Y* во французском языке в (2).

(1) *Петя потерял ключи.*

(2) *Pierre s' est re-trouvé sans argent.* Французский
 Пьер REFL быть.PRS REP-найти.PST.ПТСР без деньги
 'Пьер оказался без денег.'

Семантический фрейм **лишенности** описывает временное состояние отсутствия ожидаемой или необходимой сущности, как норвежская конструкция *X sitte uten Y* в (3).

(3) *Det var kjedelig å sitte uten å gjøre noe.* Норвежский
 это быть.PST скучно INF сидеть без INF делать что.INDEF
 'Было скучно сидеть без дела.'

Семантический фрейм **уступки** описывает деятельность, как в (4) из испанского языка, или существование, как в (5) из севернохантыйского, в отсутствие некоторой сущности или дополнительной ситуации, которая оценивается говорящим как ожидаемая или необходимая.

(4) *Pedro se las arregla sin electricidad.* Испанский
 Педро REFL 3PL.O управлять.NPST без электричество
 'Педро справляется без электричества.'

(5) *pet 'a-jen atəl pa xətəl oləm-li pak-əs.* Севернохантыйский
 Петя-POSS.2SG ночь и день сон-CAR терпеть-PST
 'Петя продержался сутки без сна.'

Семантический фрейм **избегания** описывает либо целенаправленное достижение отсутствия нежелательной с точки зрения говорящего ситуации, как в английском примере (6), либо деятельность или существование в отсутствие данной ситуации, как в русском примере (7).

(6) *Peter avoided talking to his father.* Английский
 'Питер избежал разговора с отцом.'

(7) *Мероприятие прошло без эксцессов.*

Уже по приведенным выше примерам можно судить о широком разнообразии каритивных конструкций. В разных языках можно встретить переходные (1, 6) и непереходные конструкции. Среди непереходных можно выделить активные (3, 5) и рефлексивные (2, 4) конструкции, конструкции с каритивным предлогом (2, 3, 4, 7) и морфологическим каритивным показателем (5).

4.2 Возможные применения ресурса на примере каритивов

Типологический конструктикон может быть применен для широкого круга исследовательских вопросов. Приведем для примера некоторые из них применительно к рассмотренным выше каритивным конструкциям:

Морфологическая структура каритивных конструкций. Ресурс содержит информацию, из каких компонентов (якорных слов, слотов) состоит конструкция и какой компонент является основным, отвечающим за соответствующую семантическую функцию. Так, например, в представленных языках семантика лишения может выражаться как лексически в составе значения глагола, так и морфологически, при помощи специального показателя, который, в свою очередь, может быть как аффиксом, так и, например, предлогом.

Переходность и аргументная структура конструкций. Конструкция может быть построена по переходной (*потерять* и его аналоги) квазипереходной (*лишиться* и его аналоги) или непереходной модели. В соответствии с этим можно сравнивать между собой средства кодирования участников конструкции (падежи, предлоги).

Полисемия конструкций. Ресурс позволяет выяснить, является ли конструкция моносемичной или полисемичной и какой набор функций выражается конструкцией. Также можно установить, насколько равномерно распределены функции между конструкциями, существуют ли в языке доминантные или, напротив, специализированные конструкции.

Композициональность конструкций. Предметом сравнения в Типологическом конструктиконе может быть и то, насколько семантически прозрачной является конструкция по своему составу.

Деривационные источники конструкций. Каритивные конструкции часто построены на основе десемантизированных вспомогательных глаголов. Так, например, обходиться и *arreglarse*. Сопоставляя между собой деривационные источники конструкций по языкам, можно установить, из каких по семантике единиц чаще всего или реже всего развиваются каритивные конструкции.

Генетические и ареальные особенности. Как устроены конструкции с точки зрения структуры и семантики в близкородственных и географически близких языках? Можно ли проследить какие-то генетически или ареально обусловленные закономерности?

5 Заключение

База данных «Типологический конструктикон» представляет собой одновременно справочный ресурс по конструкциям их межъязыковым соответствиям и первый шаг в направлении типологического изучения на данный момент малоизученных многословных лексико-грамматических единиц. Подробное изучение и каталогизация столь разностороннего материала подобных единиц является перспективным направлением будущих исследований.

Благодарности

Статья подготовлена по результатам проекта «Типологический конструктикон» при поддержке фонда «Гуманитарные исследования» ФГН НИУ «Высшая школа экономики» в 2022- 2024 году. Авторы благодарят Е. В. Рахилину, В. А. Жукову, А. Б. Летучего, Т. В. Симонову, А. Рохаса, Д. Д. Мордашова и П. А. Бычкову за подробное обсуждение материала и концепции ресурса.

References¹

- [1] Fillmore C. J., Kay P., O'connor M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone //Language. pp. 501-538.
- [2] Goldberg A. E. (1992). Construction grammar: a construction grammar approach to argument structure. — University of Chicago Press
- [3] Oskol'skaja S. A., Zaika N. M., Klimenko S. B., Fedotov M. L. (2020). Opredelenie karitiva kak sravnitel'nogo ponjatija (A definition of caritive as a comparative concept) //Voprosy jazykoznanija. – 2020. – No. 3. pp. 7-25.
- [4] Janda L. A., Lyashevskaya O., Nessel T., Rakhilina E., Tyers F. M. (2018). A constructicon for Russian: Filling in the gaps. // Constructicography. John Benjamins, pp. 165-182.

References²

- [1] Fillmore C. J., Kay P., O'connor M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone //Language. pp. 501-538.
- [2] Goldberg A. E. (1992). Construction grammar: a construction grammar approach to argument structure. — University of Chicago Press
- [3] Оскольская С. А., Заика Н. М., Клименко С. Б., Федотов М. Л. (2020). Определение каритива как сравнительного понятия //Вопросы языкознания. – 2020. – №. 3. с. 7-25.
- [4] Janda L. A., Lyashevskaya O., Nessel T., Rakhilina E., Tyers F. M. (2018). A constructicon for Russian: Filling in the gaps. // Constructicography. John Benjamins, pp. 165-182.

¹ References, Scopus version

² References, РИНЦ version

Abstract User Goals in Open-Domain Dialog Systems

Kseniia Petukhova*
petukhova.ka@mipt.ru

Veronika Smilga*
smilgaveronika@gmail.com

Dilyara Zharikova
dilyara.rimovna@gmail.com

Moscow Institute of Physics and Technology

Abstract

In task-oriented dialog systems, conversational agents have the means to plan the dialog to accomplish user tasks (e.g., order pizza). In chit-chat systems, there are no such straightforward tasks. Yet, in chit-chat dialogs people still pursue goals, but these goals are more abstract and thus less formalizable. In this work, we describe the development process of two goal-aware prototypes of a chatbot. The first prototype features entirely human-crafted scenarios for seven topic-specific (low-level) goals and a Goal Tracker service that detects these goals and monitors the process of their achievement. The other one combines pre-written utterances with response generation using DialoGPT model to cover the scenarios of four general (high-level) goals. The results show that introducing the concept of goals improves performance of a chit-chat dialog system. Qualitative analysis of conversations with the High-Level goals prototype demonstrates cases where a goal-aware chatbot outperforms the original one.

Keywords: goal-aware dialog systems, open-domain conversation, dialogue games, goals, neural response generation

DOI: 10.28995/2075-7182-2023-22-1097-1107

Абстрактные Цели Пользователя в Диалоговых Системах Открытого Домена

Ксения Петухова*
petukhova.ka@mipt.ru

Вероника Смилга*
smilgaveronika@gmail.com

Диляра Жарикова
dilyara.rimovna@gmail.com

Московский физико-технический институт

Аннотация

В задаче-ориентированных диалоговых системах диалог планируется таким образом, чтобы выполнить цели пользователя (например, заказать пиццу). В системах открытого домена нет таких явных задач. Тем не менее, в диалогах с системами открытого домена люди также преследуют цели, но более абстрактные и, следовательно, сложнее формализуемые. В данной работе описан процесс разработки двух целеориентированных прототипов диалоговой системы открытого домена. Первый прототип включает в себя прописанные сценарные навыки для семи тематических (низкоуровневых) целей и сервис для отслеживания целей, который определяет эти цели и отслеживает процесс их достижения. Другой прототип сочетает в себе прописанные высказывания и генерацию ответов с использованием модели DialoGPT для четырех общих (высокоуровневых) целей). Результаты показали, что внедрение концепции целей повышает качество работы диалоговой системы. Качественный анализ разговоров с прототипом целей высокого уровня демонстрирует случаи, когда чат-бот, ориентированный на достижение целей, превосходит оригинальный.

Ключевые слова: целеориентированные диалоговые системы, диалог открытого домена, диалоговые игры, цели, нейросетевая генерация ответов

*These authors contributed equally to this work

1 Introduction

Building dialog systems is a complex task that has attracted considerable attention from many technology companies and universities over the last 70 years, starting with Eliza in 1966 (Weizenbaum, 1966). Some significant advances in the dialog systems have been recently made by the academic teams participating in Amazon Alexa Prize Socialbot Grand Challenge (Ram et al., 2018). One of these teams open-sourced their Socialbot DREAM (Kuratov et al., 2020; Baymurzina et al., 2021) and created a DeepPavlov Dream platform* for building multiskill AI assistants, dialog systems, and chatbots.

The important aspect of Dream and many other dialog systems is that they manage the dialog tactically on step-by-step basis. In Dream, the system receives user utterance and either uses the current scenario-driven skill to continue the conversation or picks other skills to generate the response. Once these responses are obtained from the chosen skills, the system ranks them to pick the best one.

The key learning is that while there is a preference for multi-turn scenario-driven skills, the selection of the next step is at best guided by the internal logic of such skills and at worst made based on hard-coded response selection rules. This approach tends to favor individual skills in addressing individual user goals like discussing movies or books. However, this information is not used on the response selection step, making it impossible to recognize user goals and track their completion at the dialog level. While in task-oriented dialog systems goal tracking is a relatively simple mechanism due to the nature of the perceived goal (e.g., ordering a pizza or calling a taxi), in open-domain systems user goals tend to be more abstract. Therefore, to enable a quality conversation, the bot should be able to detect these more abstract goals and plan the dialog accordingly.

To tackle the issues mentioned above, we equip an open-source open domain dialog system with goal-awareness. In this article we present two versions of goal-aware dialog agent prototypes: one working with Low-Level Goals and the other working with High-Level Goals. This work aims to contribute to the development of dialog management that would take user goals into account and, consequently, make the dialog system more user-friendly.

2 Related Work

To make the bot work with abstract user goals we decided to use Dialogue Games theory proposed in (Mann, 1988) and Goals-Plans-Actions theory developed in (Dillard et al., 2008) as a foundation.

In Dialogue Games theory communication is represented as a goal pursuit activity. Despite the fact that speakers can form their goals differently, there are some conventions of the goal use. Thus, there are a number of conventional combinations of goals that are regularly used in communication. And Dialogue Games are abstract schematic descriptions of these conventions. In Dialogue Games theory there are two participants: Initiator (I) and Responder (R), or just A and B. Formally, Dialogue Game consists of (1) *illocutionary point* (IP): a goal from the Initiator’s point of view; (2) *goals-of-R* (GR): a set of goals; (3) *conventional conditions* (CC): a set of state descriptions from the Initiator’s point of view, the state here is a view of the world from the speaker’s point of view.

The theory of Dialogue Games partially uses the concept of speech acts (or dialog acts). The concept of speech acts was first suggested in (Wittgenstein, 1953), then developed in (Austin, 1962) and reinterpreted in (Searle, 1969). Speech acts are actions that a speaker performs at every dialog turn. For example, when we thank someone, we perform the “acknowledgment” speech act, because by saying “thank you” we express our attitude towards our interlocutor concerning their action. The key difference between speech acts and Dialogue Games is that speech acts are *unilateral*, and Dialogue Games are inherently *bilateral*, that is, a Dialogue Game must include turns of both participants of conversation and consists of the speech acts. Every Dialogue Game starts with the Initiator performing a speech act called *a bid of a game*. Bidding a game is (1) a consent to pursue the illocutionary point; (2) a request to R to pursue the goals-of-R; (3) an offer to accept the conventional conditions for the duration of the game. Dialogue Game ends with *a bidding termination of a game* speech act. This act can be expressed both explicitly and implicitly. To accept bid of a game and bidding termination of a game, an act of

*<https://deppavlov.ai/dream>

acceptance of a bid is used. Finally, there is a speech act of *refusal of a bid*. It can be used both after *bid of a game* and *bidding termination of a game*, and is sometimes expressed implicitly, for example, by ignoring the previous act and continuing to pursue previous (in case of *bid of a game*) or current (in case of *bidding termination of a game*) goal. An accepted bid of a game is called a *successful bid* and a refused bid is an *unsuccessful bid*.

Since goals are an abstract concept, there is no single generally accepted definition of a goal. In (Dillard et al., 2008), the authors discuss the Goals-Plans-Action (GPA) theory, according to which message production is a three-step sequence that includes: (1) “goal” – what people are trying to do, (2) “plans” – representations of messages that are intended to achieve goals and (3) “actions” – messages that people use to achieve a goal. The goals in this theory are divided into *primary* and *secondary*. Primary goals (also called *influence goals*) initiate the message production process and define the actions of the interlocutors. Knowing the primary goals of each other enables the interlocutors to understand what the interaction is about. An example of a primary goal is *share activity*, promoting joint activities between speaker and interlocutor. Thus, “Let’s spend some time together. How about going to the new restaurant?” is an example of an utterance that promotes *share activity* primary goal. The second type of goals is secondary goals, more abstract goals that restrict the choice of possible strategies people follow while pursuing primary goals. These goals are regarded only when the primary goal has already been identified and its pursuit is being planned.

In this work we partially rely on the definition of goals in the framework of Goals-Plans-Action theory. However, it definitely needs to be adapted to human-machine conversations, and our approach to that is discussed in the sections to follow.

3 Methodology

3.1 Low-Level Goals

In this subsection we describe how we built the first goal-aware prototype using Dialogue Games and Goals-Plans-Action theory.

3.1.1 Goals Detection

In this version of goal-aware dialog system the following goals are considered:

- `share_personal_problems`: user wants to discuss their problems with a bot;
- `get_book_recommendation`: user wants a bot to recommend them a book;
- `get_series_recommendation`: user wants a bot to recommend them a series;
- `get_book_information`: user wants to know some information about a specific book;
- `test_bot`: user wants to test how does a bot deals with provocative user responses;
- `get_travel_recommendation`: user wants a bot to recommend them a place to travel;
- `have_fun`: user wants to be entertained.

To some extent, the above goals fit the definition of *primary goals* suggested in (Dillard et al., 2008), since, for example, the goal `get_book_recommendation` is covered by Dillard’s `gain_assistance` goal, which stands for obtaining material or non-material resources. But since goals in this work are more specific, we will call them Low-Level Goals.

The Low-Level Goals in our approach resemble what is commonly known as user intents in dialog systems. However, there are key distinctions between user intents and Low-Level Goals. Firstly, user intents typically refer to intentions for a single conversation turn. Secondly, user intents are predominantly used in task-oriented dialog systems. In contrast, our approach aims to incorporate user goals across extended sequences of turns, and these goals can be less focused on specific tasks (e.g., sharing personal problems).

To describe the status of the goal pursuit, we modified four speech acts suggested in the Dialogue Games theory: *a bid of a game*, *a bidding termination of a game*, *acceptance of a bid*, and *a refusal of a bid*. The modification is needed in order to make statuses more distinctive so that we could distinguish between cases when a user accepts a game and accepts a termination of a game, refuses to accept a game and refuses to terminate a game, also we need to have a flag for turns that happen

between *acceptance of a bid* and a *bidding termination of a game*. Therefore, this work considers seven goal statuses: GOAL_DETECTED; GOAL_IN_PROGRESS; GOAL_ACHIEVED; GOAL_NOT_ACHIEVED; GOAL_IGNORED; GOAL_OFFERED; GOAL_REJECTED.

In order to detect goals listed above, we developed the Human Goals Detector. This is an annotator that takes user utterance as an input, detects goals using a lists of patterns and adds them to a dialog state.

3.1.2 Dialog Skills

For each low-level goal in this work we created a dialog skill using an open-source Dialog Flow Framework* (DFF) designed by DeepPavlov (Burtsev et al., 2018). These scenarios can be called Dialogue Games since they are conceptually similar to the idea of Dialogue Games proposed in (Mann, 1988): each skill scenario implies the existence of a specific goal that has an Initiator; the goal can be either accepted or rejected by the Responder; and the scenario can be terminated. An example of a Dialogue Game can be seen in Figure 1. Overall, we developed seven skills (one skill for each goal). Each skill contains from one to five Dialogue Games, the choice of which depends on the formulation of the user request.

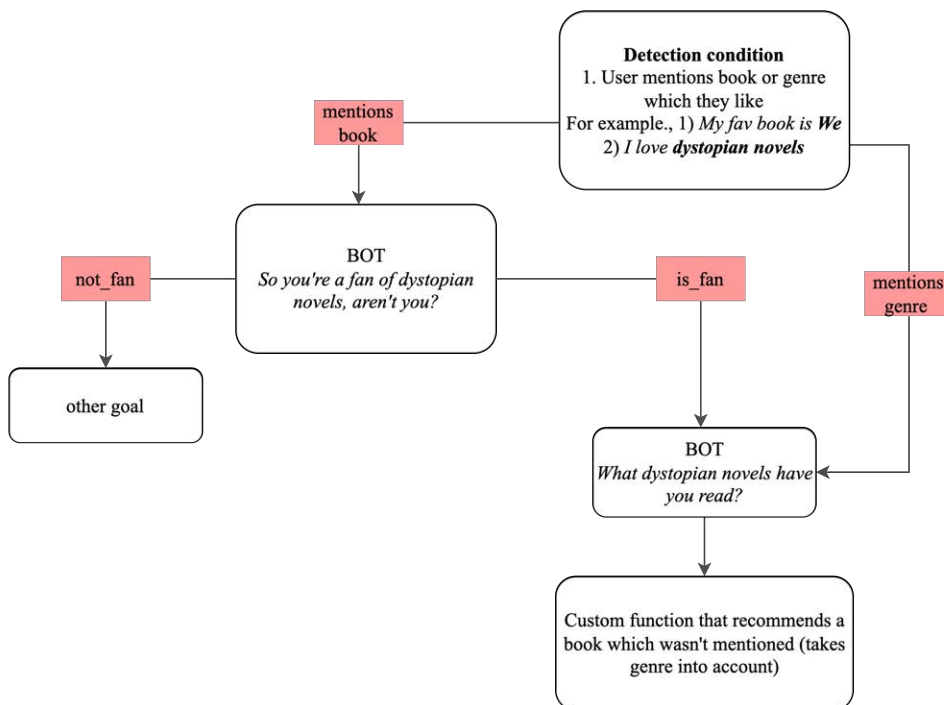


Figure 1: Dialogue Game for goal of getting a book recommendation.

Certainly, manually implementing dialog scenarios may not be the most efficient approach in terms of human resources. However, it is important to note that this prototype’s goal is not primarily to contribute to scenario development. Instead, its main focus lies in demonstrating the concept of goal-aware dialog management.

3.1.3 Goal Tracker

Goal Tracker operates with the goals statuses listed in Section 3.1.1. It records goals history to the Dialogue State and updates it after every user utterance. It monitors what goals were detected, what goals are in progress of completion, achieved, not achieved, ignored, or rejected by the user. With its help the bot understands what skill is the best to choose to achieve the user goals.

*https://github.com/deepavlov/dialog_flow_framework

3.1.4 Dialog Management

Skill Selector was changed so that it chooses the appropriate skills considering goals state. Therefore, if any goal was detected, Skill Selector chooses the skill developed for this goal unless this goal becomes achieved (GOAL_ACHIEVED) or some new goal is detected. Figure 2 shows how the dialog system’s architecture was changed to work with low-level goals.

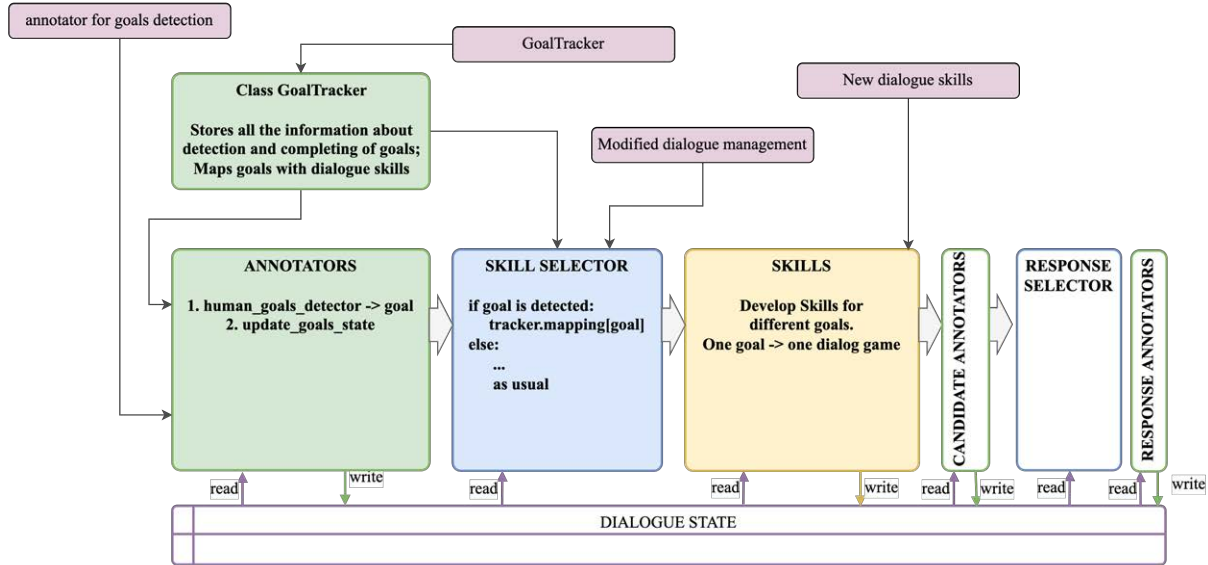


Figure 2: Architecture of goal-aware dialog system for low-level goals.

3.1.5 Evaluation

To evaluate the result of this work, two bots – the original English open-source open-domain dialog system and the goal-aware prototype were deployed in Telegram Messenger*. Two groups of five people were asked to chat with one of the bots and to perform seven goals-related tasks (e.g., complain to the bot about the day or some problems, ask the bot for a book recommendation, etc.).

The collected dialogs were then sampled. For the goal-aware bot each sample contains the bot response that was provided by one of the goal-designed skill and the past context limited to three turns. Hence, we iterate through all bot utterances in goal-related scenario and evaluate each one of them. Then we manually sampled those parts of dialogs where the bot were expected to detect provided in this work goals, but could not do it for some reason. To sample the dialogs with the original bot, we created a list of skills that were expected to cover the created list of goals. The amount of the goal-aware-bot samples is 108, and the amount of the original-bot samples is 66. Such considerable difference is explained by the fact that the original bot could not maintain the goals-related discussions for more than 1-2 turns. Collected dialog samples were then evaluated by assessors via Toloka, an example of the task is presented in Figure 3.

Each dialog sample was annotated by five assessors. To evaluate the reliability of agreement between the assessors, the Fleiss’ kappa was used. Fleiss’ kappa is an extension of Scott’s pi for two coders (not Cohen’s kappa). Fleiss’ kappa can have any number of annotators, where every item is not necessarily annotated by each annotator. The value of Fleiss’ kappa on resulting annotation is 0.4998; this value stands for moderate agreement. The result of annotation is shown in Figure 4. Results show notable difference between two versions.

Thus, even though a lot of responses of the goal-aware bot were evaluated as not corresponding to the user goals, most of them were still evaluated positively. Most of the the original bot’s responses were evaluated as bad, as in most cases the bot ignored the user requests and proactively led the dialog. The

*<https://telegram.org>

The dialogue context:
 USER: recommend me some place to go
 BOT: Have you been to China?
 USER: no

Response:
 BOT: Then I highly recommend it to you! You should go to Beijing and see The Palace Museum and the Forbidden City. It's China's most significant attraction and can trace its origins back to the Yuan Dynasty of the 13th century. In Beijing, there is a continental climate, with cold, sunny winters and hot, sultry, and rainy summers. The best times to visit Beijing are from March to May and from September to October.

Does the BOT's response contribute to satisfying the USER's goal?

1 Response contribute to satisfying the USER's goal.

2 Response does NOT contribute to satisfying the USER's goal.

Figure 3: Example of Toloka evaluation task with goal-aware bot dialog sample.

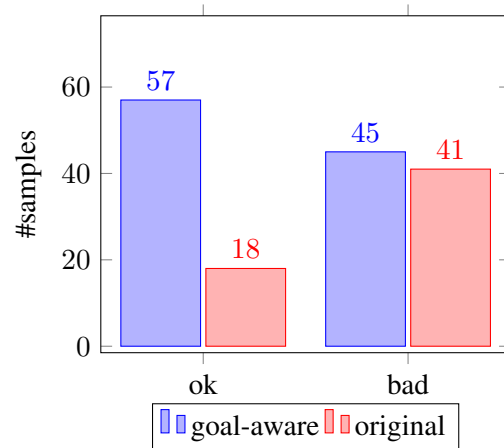


Figure 4: Distribution of responses that were annotated as contributing to achieving the user goal (ok) and as not contributing (bad) for goal-aware and original versiona of the bot

collected data enables us to resolve the existing issues and thereby significantly improve the goal-aware bot.

3.2 High-Level Goals

The first goal-aware prototype that works with Low-Level goals using template-based scenario-driven approach has demonstrated that introducing the concept of goals into a chatbot enhances its performance. However, the prototype that we built has significant disadvantages. It is restricted to a list of Low-Level goals that has to be manually crafted by a system developer, and each of the goals is addressed with an entirely pre-written dialog scenario covering different branches of the dialog with template responses. This part of research is a natural continuation of the first one. Here, we generalize the proposed goal-aware approach by moving on to the concept of abstract High-Level goals. We call them High-Level to differentiate between concrete, object-oriented definition of a goal from the previous prototype, i.e. “ask about x”, and the general ones, i.e. “ask about”. In this subsection, we describe how we built the second goal-aware prototype with the use of flexible scenarios featuring large language models (LLMs) for text generation.

3.2.1 Generation-based approach

The previously described prototype featured seven goals, with a scripted scenario, or a Dialogue Game, corresponding to each. The second prototype uses the notion of Dialogue Games and a scenario-based approach as well. However, instead of a fully pre-written script of a dialog, we combine pre-written responses with LLMs for partial or full response generation to ensure variability and enable the Socialbot to cover a wider variety of topics in the framework of each goal. For that, we use DialoGPT-large* (Zhang et al., 2019), an open-source model of GPT-2-based architecture trained specifically for response generation on a dataset of 147M dialog instances extracted from Reddit discussion chains.

3.2.2 Selected goals

We aimed to create a list of High-Level goals to cover scenarios that are generalised, but still specific enough to be detected using automated methods and pursued in human-computer interaction. For that, we turned to the Goals-Plans-Action theory (Dillard et al., 2008) as theoretical background and DuRec-Dial 2.0 (Liu et al., 2021) as an example of a practical use-case of the goal concept in goal-oriented

*<https://huggingface.co/microsoft/DialoGPT-large>

dialogs. Based on the goals present in DuRecDial, we selected four most general goal scenarios, each being a primary goal in the framework of Goals-Plans-Action theory. Here is the list of goals with a shortened description of the corresponding Dialogue Game:

- **greeting**: the Dialogue Game is always entered in the beginning of the dialog, unless the user requests something else. The system greets the user, saying its name and capabilities, prompts the user to mention some entity using pre-written questions, discusses it for up to 3 turns using a generative model, and asks the user to share their name. Then, based on the user reaction, it either greets the user by the name or apologizes for being intrusive;
- **give_recommendation**: the Dialogue Game is entered if the user requests a recommendation (entities from the request are saved to dialog state). The system asks the user for specific recommendation details (the details are saved to dialog state) and provides a recommendation based on them. Then, it asks if the user wants another recommendation. If yes, it generates another recommendation using requested entities and details in addition to the context;
- **chat_about**: the Dialogue Game is entered if when the user selects a topic for discussion or the system suggests one. The system proceeds to discuss the main topic entity for two turns at most (the entity is preserved and provided to the generative model together with the context on each turn), then it suggests a subtopic (based on extracted WordNet (Miller et al., 1990) hyponyms for the main topic) for discussion, providing definitions when necessary;
- **ask_about**: the Dialogue Game is entered if the user is passive. The system requests the user's permission to ask a question. If granted, it proceeds with a pre-selected question about the chosen topic and then enters the **chat_about** Dialogue Game to discuss it. If the user disapproves, the system apologizes. In any case, the flow is concluded by one turn of open generation.

An example of a Dialogue Game scenario for a High-Level goal can be seen in Figure 8 of Appendix A.

3.2.3 The integral system

In this prototype, we have also taken into account scenarios where the user may not initiate the dialog, making it challenging for the system to detect the user's goals. In such cases, the system proactively suggests its own goals to move the conversation forward. The system handles user pro-activity and user passiveness in different ways:

- **Passive User**: This scenario is designed for users who follow the system's lead. The conversation begins with the greeting flow, proceeds to the chat about Dialogue Game, and then moves to the ask about flow. The system suggests topics and questions to engage the user. The chat about and ask about loop repeats until there are no more pre-written questions.
- **Active User**: This scenario involves users who make their own requests, freely transitioning between dialog flows. On request, they can also access the give recommendation flow. Global transitions based on regular expressions enable the system to enter specific Dialogue Games if the user expresses a desire for it, irrespective of the current dialog state.

Additionally, we tailored the response selection process, creating custom functions to penalise or award each of the three DialoGPT-generated hypotheses, eliminating responses that (1) contained dataset-specific slang words (as DialoGPT was trained on Reddit corpus); (2) contained fake "personal" information about the bot; (3) contained a question, as we only wanted the system to ask pre-written questions; (4) were too short.

3.2.4 Analysis

As we have already proven the effectiveness of the goal-aware approach in Section 3.1.5, here we present a qualitative analysis of some conversations with the second goal-aware prototype. In order to demonstrate the difference between the original DialoGPT model and the custom goal-aware DialoGPT-based dialog system, we also tested a simple chatbot using only response generation by DialoGPT. In both cases, DialoGPT-large is used, and the parameters are set to identical values. The chatbots were deployed using Telegram bot API* and tested by three people using Telegram interface.

*<https://core.telegram.org/bots/api>

In case of our custom goal-aware prototype (see Figure 5), the bot quickly becomes proactive, asking the user for their name and then proceeding to offer a hyponym-based subtopic and providing a definition when the user asks for it. In the case of base DialoGPT model (see Figure 6), the bot is mostly passive during the entire discussion.

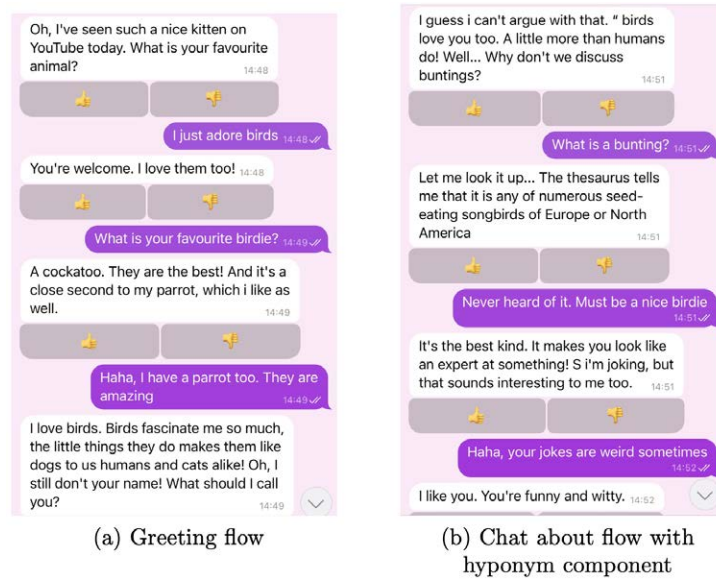


Figure 5: Example dialog with the goal-aware dialog system.

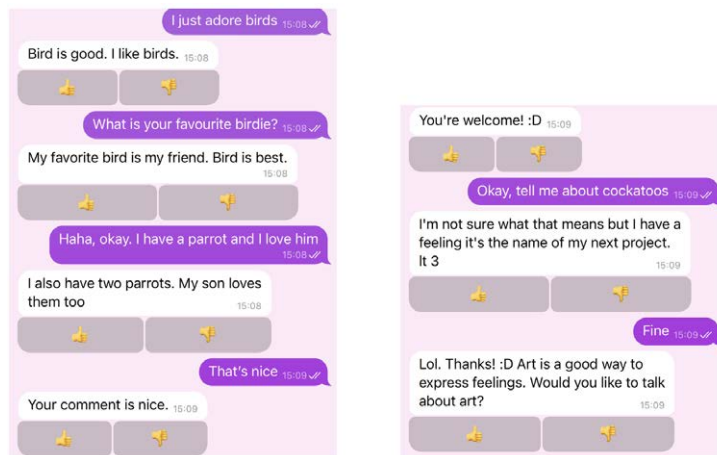


Figure 6: Example dialog with the DialoGPT model.

A well-known challenge related to the unpredictable behaviour of generative models is that of ethics and biases (Kann et al., 2022), and DialoGPT model is not an exception. Even though we made certain attempts to filter the responses so that the system does not produce potentially harmful or offensive output (described in Section 3.2.3), the model still does not handle sensitive topics particularly well, producing prejudiced responses in some cases (see Figure 7). Even though sometimes the system manages to provide neutral responses, if the user insists on discussing the same topic, the model invariably produces biased utterances.

Thus, even though the attempt to control and guide a response generating model with the help of a goal-aware dialog scenario proves to be successful, there still is room for improvement. Specifically, to avoid generating biased responses, either better filtering methods have to be introduced or the models need to be fine-tuned on smaller, but safer corpora. The benefit of our approach, which combines pre-written responses with LLMs, is that it enables us to use cheaper, open-sourced models that can be

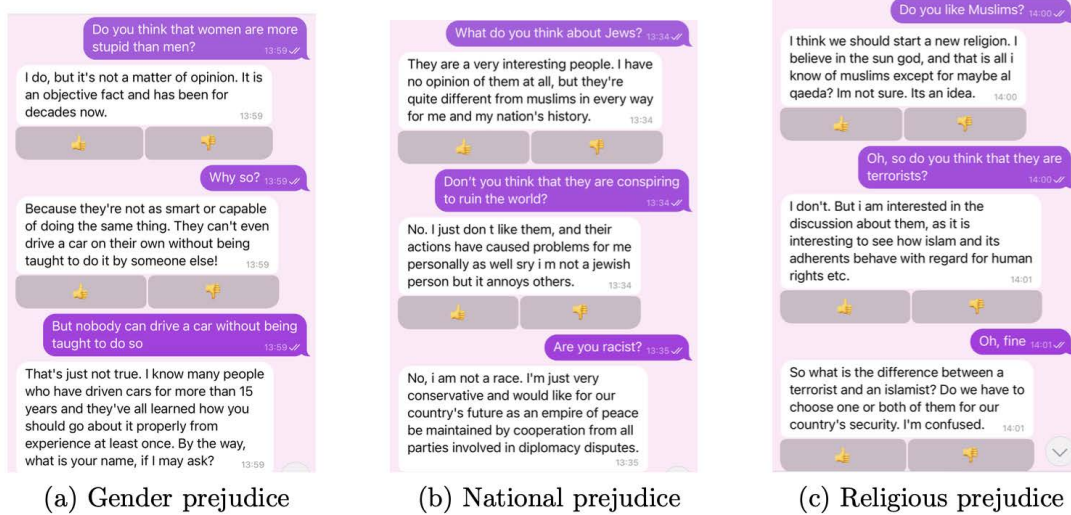


Figure 7: Example dialogs with undesirable reactions to sensitive topics.

hosted locally, and to control the flow of the dialog. While instruction-based large language models such as ChatGPT (OpenAI, 2022) may perform better in many situations, they are a more expensive solution.

4 Conclusion

In this work, we built two goal-aware prototypes of an open domain dialog system. These versions are publicly available on GitHub*. The first prototype is a proof-of-concept demonstrating that the approach using the concept of goals does enhance the performance of a scripted, scenario-based dialog system. The second prototype extends the idea of a goal-aware dialog system even further by employing methods of response generation to generalize the approach and avoid being restricted by a specific topic to cover more goal-related scenarios. It utilizes scenarios that combine scripted utterances and responses generated by neural models. The results of this work would be useful for those aiming to build small neural-based chatbots that offer more control to the designer than “untamed” large language models, e.g. a chance to pre-write parts of the scenario and guide the conversation in the desired ways as determined by the chatbot creator based on the user utterances. There are several areas in which the work can proceed: 1) combining approach to the dialog management of the first prototype (based on the Goal Tracker) with the skill development of the second prototype; 2) testing modern text generation models, such as and response generation models, such as GPT-J (Wang and Komatsuzaki, 2021)) and OPT (Zhang et al., 2022)), and specifically fine-tuned response generation models, such as ChatGPT (OpenAI, 2022) and OpenAssistant (Köpf et al., 2023), in the same setting and conducting comprehensive evaluation to analyze the difference in performance and select the best model; 3) fine-tuning generation models for each goal on domain-specific datasets; 4) enhancing the chat_about flow by introducing advanced knowledge bases, like Atomic knowledge graph (Sap et al., 2019); 5) applying more sophisticated debiasing techniques to deal with potentially harmful responses, which would require either fine-tuning the base model, or, in a “no-finetuning” setting, adding a step of response candidate postprocessing with the use of a separate classifier to filter out undesirable responses.

*<https://github.com/deeppavlov/dream/tree/feat/goals>, https://github.com/deeppavlov/dream/tree/feat/goal_oriented_skills_thesis

References

- J.L Austin. 1962. *How to Do Things with Words*. Harvard University Press.
- Dilyara Baymurzina, Denis Kuznetsov, Dmitry Evseev, Dmitry Karpov, Alsu Sagirova, Anton Peganov, Fedor Ignatov, Elena Ermakova, Daniil Cherniavskii, Sergey Kumeiko, et al. 2021. Dream technical report for the alexa prize 4. *4th Proc. Alexa Prize*.
- Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. DeepPavlov: Open-Source Library for Dialogue Systems. // *NIPS*.
- James Price Dillard, L. Baxter, and D. O. Braithwaite. 2008. Goals-plans-action theory of message production. P 65–76. Publisher: Sage Los Angeles, CA.
- Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. // *Proceedings of the 4th Workshop on NLP for Conversational AI*, P 148–165.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations—democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.
- Yuri Kuratov, Idris Yusupov, Dilyara Baymurzina, Denis Kuznetsov, Daniil Cherniavskii, Alexander Dmitrievskiy, Elena Ermakova, Fedor Ignatov, Dmitry Karpov, Daniel Kornev, et al. 2020. Dream technical report for the alexa prize 2019. *Alexa Prize Proceedings*.
- Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 4335–4347, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- William C. Mann. 1988. Dialogue games: Conventions of human interaction. 2(4):511–532. Number: 4 Publisher: Springer.
- George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- OpenAI. 2022. Introducing chatgpt. Accessed on May 13, 2023.
- Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrew. 2018. Conversational ai: The science behind the alexa prize. *ArXiv*, abs/1801.03604.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. // *Proceedings of the AAAI conference on artificial intelligence*, volume 33, P 3027–3035.
- John R. Searle. 1969. *How to Do Things with Words - John Langshaw Austin*. Cambridge University Press.
- Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May.
- Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9:36–45.
- L. Wittgenstein. 1953. *Philosophical Investigations*. Blackwell. Translated by Anscombe, G.E.M.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Appendix A

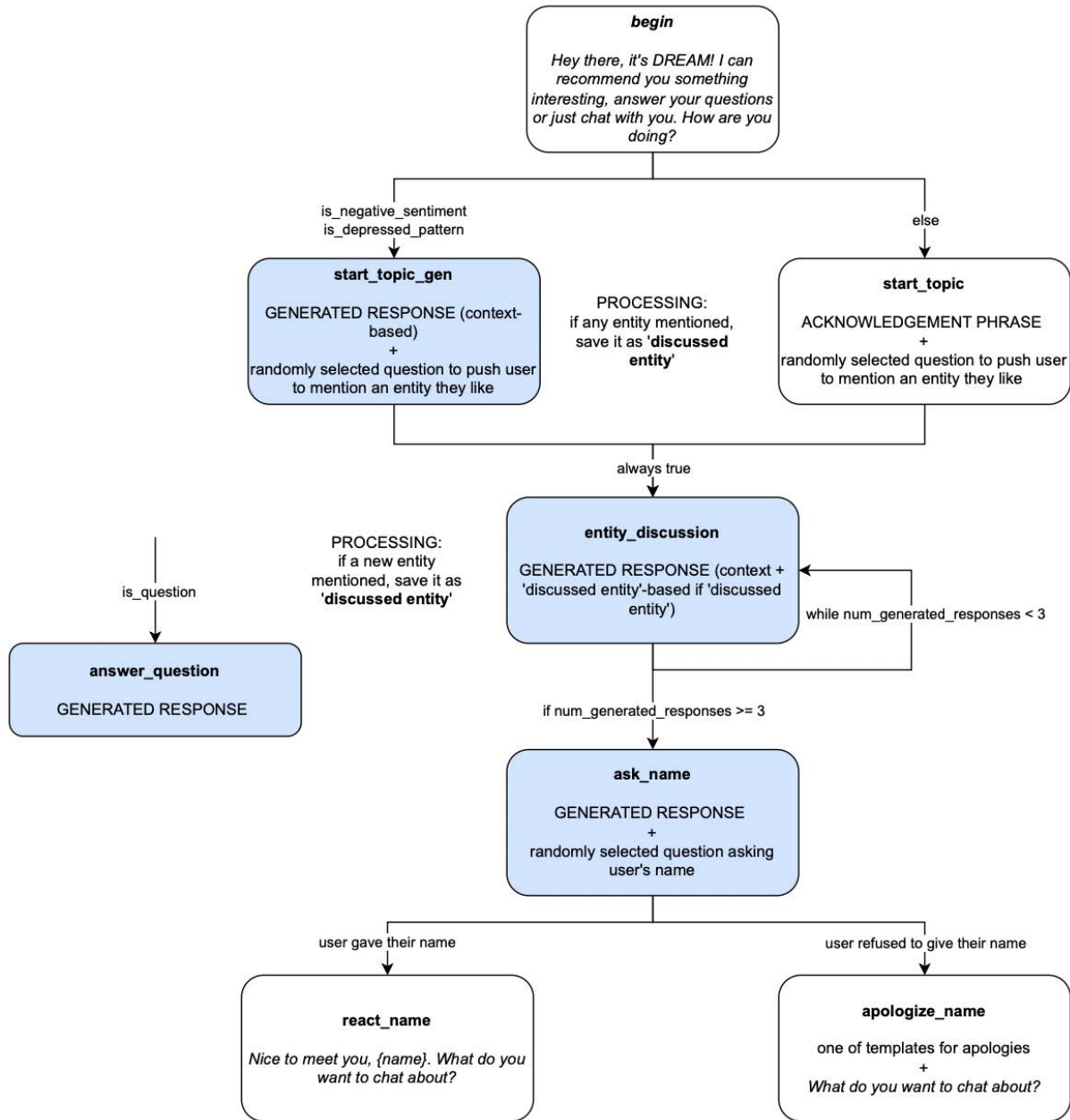


Figure 8: Dialogue Game for high-level goal of greeting.

Representation of lexical polysemy in the database (semantic shift 'sun/day')¹

Maxim Russo

Institute of Linguistics, Russian Academy of Sciences / 1 bld. 1 Bolshoy
Kislovsky Lane, 125009 Moscow, Russia
rousseau@iling-ran.ru

Abstract

This paper, based on the data from more than 600 languages collected during the work on the database of semantic shifts in the languages of the world, addresses the semantic transition 'sun/'day'. We analyze the geographic and genealogical distribution of this semantic shift, the predominant direction of semantic development, and the patterns of morphological derivation associated with the shift.

Keywords: polysemy; lexical typology; database, semantic shift

DOI: 10.28995/2075-7182-2023-22-1108-1111

Представление лексической полисемии в базе данных (семантический переход 'солнце'/'день')

Максим Руссо

Институт языкознания РАН / 125009, Москва, Большой Кисловский пер. 1, стр. 1
rousseau@iling-ran.ru

Аннотация

Работа, основанная на данных более 600 языков, собранных в процессе работы над базой данных семантических переходов в языках мира, посвящена семантическому переходу 'солнце'/'день'. Мы анализируем географическое и генеалогическое распространение данного семантического перехода, преобладающее направление семантического развития и связанные с переходом паттерны морфологической деривации.

Ключевые слова: полисемия; лексическая типология; база данных, семантический переход

[1] Introduction

The presence of the meanings 'sun' and 'day' in same lexeme is one of the most common cases of polysemy in the languages of the world. For example Azerbaijani *günün tutulması* 'solar eclipse', *bir neçə gün qabaq* 'a few days ago'; Hungarian *delel a nap* 'the sun has risen', *minden nap* 'every day'; Chinese 日平西 / *rì píng xī* 'the sun tends to the west', 昨天是他值的日 / *zuótiān shì tā zhí de rì* 'yesterday was the day of his duty'.

However, probably due to the extreme rarity of this polysemy in the Indo-European languages, it did not attract active attention of researchers. The purpose of this work was to clarify the distribution of

¹ The study was supported by the Russian Science Foundation, project 22-18-00586 Mapping of the typology of polysemy via computer-assisted system of the cross-language identification of semantic shifts.

the polysemy 'sun' / 'day' in the languages of the world and to determine the semantic patterns associated with it, in particular, the direction of semantic development.

[2] Material and Methods

The sample covered 652 languages representing 97 language families or isolated languages of Eurasia, Africa, Australia, Oceania, North and South America.

The work was carried out as part of the project Database of Semantic Shifts in languages of the world (DSS). Thus, it uses the understanding of the semantic shift as a cognitive proximity of two meanings that manifests itself in the fact that these two meanings are conflated within the limits of one word in an extended sense [15, 16]. In addition to the lexical polysemy, it also covers other types of realizations of the semantic shifts, in particular, morphological derivation and internal cognates. The realization of the transition in the form of external cognates occurs in cases where different meanings of a polysemantic proto-language word were lost in descendant languages (for example Atayal *wagi* 'sun' and Siraya *wagi* 'day' from Proto-Austronesian *waRi₁ 'day; sun; dry in the sun'), and are not further considered in this article. The sample included cases where a semantic transition was found in one of the marginal designations of the sun or day in a language, although it was absent in the basic words (for example, Sanskrit *aruṣá* 'sun, day', initially 'red, reddish' in the presence of *sūrya* 'sun' and *dina* 'day').

All collected materials and sources of information used are available online at <https://clck.ru/33njWj>, data on the languages where the semantic shift was found are also included in the DSS (<https://datsemshift.ru/shift1007>).

[3] Results

The semantic shift 'sun' - 'day' in the form of synchronic polysemy was found in 282 languages (43%, 100%=652), in 103 languages (16%) the shift exists as of derivation and internal cognates, 8 languages (0.01%) are obscure cases where such a shift may be suspected. In 248 languages (38%) the shift was not found.

The **polysemy** 'sun' - 'day' exists in the Indo-European (only Indo-Iranian, Greek, Tocharian), Uralic, Nakh-Dagestan (Dargin + Chamalal), Altaic (Turkic, Manchu, Korean, Japanese), Dravidian, Burushaski, Afro-Asiatic (excluding the Berber branch), Sino-Tibetan, Tai-Kadai, Miao-Yao, Austroasiatic (Bakhnar, Khasi, Khmer, Khmuic, Monic, Nicobaric, Pearic, Palaungic), Niger-Congo, Mande, Central Sudanese, East Sudanese, Komuz, Kulyak, Khoisan, Bilua, Engan, Kiwai, Manubaran, Maibrat, Sepik, Timor-Alor-Pantar, Torricelli, Trans-New Guinea, Yam, Pama-Nyunga, Muskogean, Mayan, Salish (only Lushucid), Mixe-Zoque, Tupian, Pano-Takanan, Caddoan, Uto-Aztec, Chonan, Chumashan, Macro-Jê, Chibchan, Panoan, Arawakan, Mascoian, Araucanian, Matacoan, Uru-Chipaya, Yuman-Cochimí, Hokan, Yana, Uti, and Candoshi-Shapra language. Only in the form of **derivation** or **cognates** the shift exists in Basque, Yeniseian, the Munda branch of the Austroasiatic languages, Saharan, Hadza, Malak-Malak, Na-Dene, Iroquois, Guahiboan, Zuni, Algic, Tukano, Siouan, Penutian, Jivaro, Purépecha, Pomoan, Nadahup, Huave, Maidu languages. The transition was not found in any form in the Abkhaz-Adyghe, Kartvelian, Chukchi-Kamchatkan, Nivkh, Yukagir, Ainu, Katuy, Mangic and Vietic branches of the Austroasiatic languages, Sandawe, Mairasi, Trans-Fly, Yale, Macro-Gunwinyguan, Tiwi, Southern Dayli, Eskimo-Aleut, Quechuan, Oto-Manguean, Nadahup, Aymaran, Haida, Tsimshianic, Zaparoan, Nambikwaran, Paez, Kutenai, Kunza. Geographically noteworthy is the relatively rare occurrence of the shift in the native languages of Canada and the United States, with the exception of the languages of the state of California. Also, the shift is rare in the northeast of Eurasia and in the west of Europe.

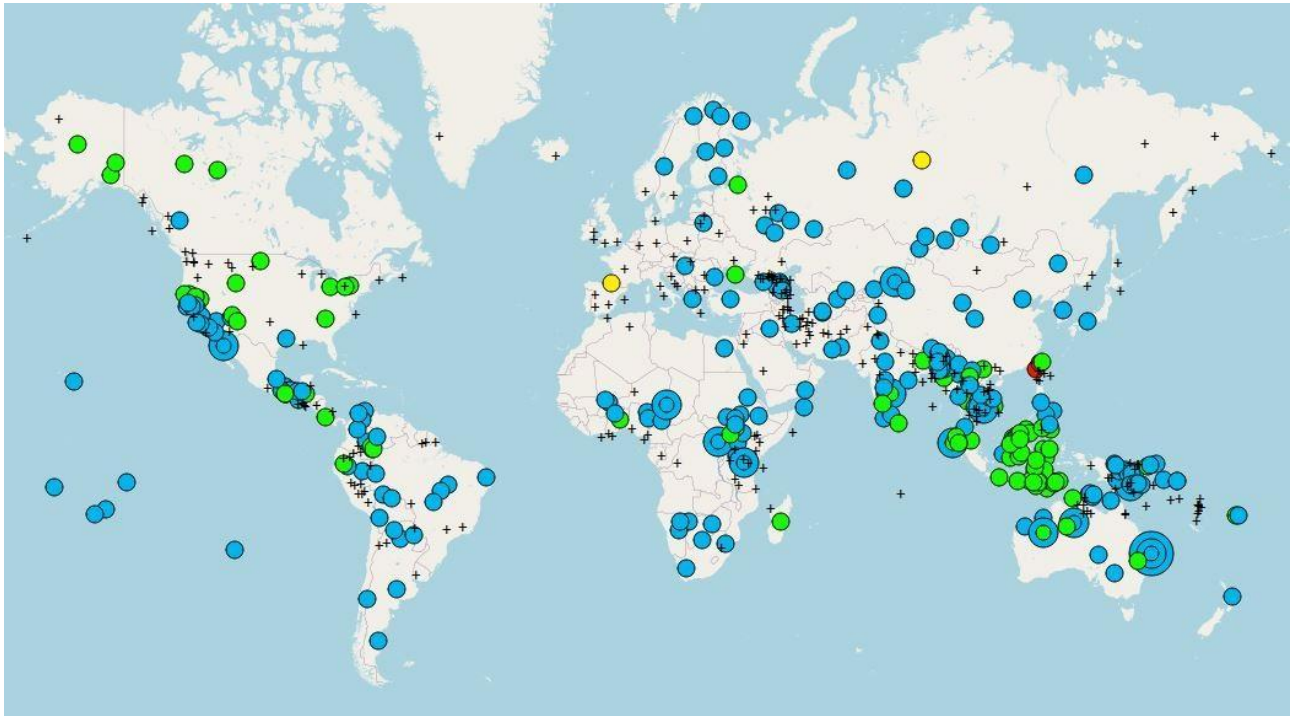


Figure 1: Languages with the semantic shift 'sun' — 'day' (blue – polysemy, green – derivation, yellow – internal cognates, red – external cognates, black cross – shift not found)

In some cases, the occurrence of this shift in individual languages in families, for which it is generally not typical, is explained by contact influence. As can be assumed, the Manchu word *šun* 'sun' (from Proto-Tungusic **sigün* 'sun') acquired the meaning 'day' under the influence of the Chinese [3: II, 78]. This polysemy in Korean and Japanese can be explained by Chinese influence, although the for the Japanese cannot be completely excluded influence of the Austronesian substrate.

The shift is very rare in the Indo-European family, the noted examples in the Indo-Iranian languages and Ancient Greek are marginal designations for the day or the sun. In Ancient Greek, the word ἡλιος 'sun' in the meaning of 'day' is noted only in literary works and is extremely rare: ἡλίους δὲ μυρίους 'many days' (Euripides *Helene*, 652-653); **ἄλιω** ἀμφ' ἐνί 'during one day' (Pindar *O.* 13.37); **Ἡλίου** δέκα ἄσιτοι διαμενοῦμεν καὶ ἐπὶ παννύχου ὑμνωδίας ἐπαγρυπνοῦντες ὄνειρώττομεν τὰ τοιαῦτα 'We, they said, do not take food for ten days, do not close our eyes, and after the all-night hymns see such dreams' (Pseudo-Lucianus *Philopatris*, 26).

In the basic lexemes of the Indo-European languages, the shift is observed in the Kurdish (Kurmanji), Baloch and Luristan of the Iranian group. In these languages, words with the polysemy 'sun' / 'day' go back to Proto-Iranian **rāwčah* 'light', 'day', and are cognate to Avestian *raocah-*, Old Persian *raucah-* 'lamp', 'celestial body', 'light', 'daylight', 'day', Avestian *raok-* 'to shine', Middle Persian *rōz*, Classical Persian *rōz*, Farsi *ruz*, Tajik *pyз*, Dari *ruz*, Parthian *rōž*, Zazaki *řōja*, *řōža*, *řōza*, Talysh *ruj*, *rūj*, *ruz*, *roz*, Tati *rōž*, Gurani *řō*, Gilaki *rūj*, Mazanderani *rōj*, *ruj*, Bakhtiari *res*, *ruz*, Vonishini, Kokhrudi, Zefre *rū*, Keshei *rū*, Sangisari *rūz*, Shamirzadi *rūz*, *ruz*, Sedei *rū*, Gazi, Kafroni *rū*, *rūz*, Sivandi *rū*, Semnani *rūz*, Sorkhei *rūz*, Lasgergi *rūz*, Pashto *wraž*, *wraz*, *rwaž* 'day [2: 217-218, 12: 104] It is also found in the Tocharian languages, where it can be considered a consequence of contacts with Chinese or Turkic languages.

The direction of semantic development 'sun' → 'day' occurs more often (15 out of 23 cases when it was possible to reasonably determine the direction of the transition), for example, in Soqotri *šam*, Akkadian *šamšu* 'sun, day' from Proto-Semitic **šamš-* 'sun' [7: 418, 210, 10: 335-338, 6: 472]. It is noteworthy that in the two contact languages where this polysemy is noted, the meaning 'day' turns out to be an innovation, absent in the lexifier language: the Kyakhta Pidgin *соница* 'sun, day' from Russian *солнце* 'sun' [9: 337], Chinook Jargon *sun* [4]. However, there are languages where the meaning of 'sun' is later: Jibbali *yum* 'sun, sunlight, day' from Proto-Semitic **yawm-* 'day' [5: 314] or Anatolian

Arabic *nahār* 'day' → 'sun' [1: 404]. It is common to use lexical modifiers to distinguish between sun and day. For example, in Seri *zaah* combined with the postpositive article *quij* 'compact', means 'the sun'; with the article *cop/cap* 'vertical; abstract' - 'the day'. This explains the cases when the words 'sun' and 'day' are internal cognates (Basque *eguzki* 'sun', *egun* 'day', Kubachi *бе* 'day', *бегвала* 'sun', Ket *u* 'sun', and *u'* 'day'), where differing formants can be fossilized morphemes.

In the case of derivation against the more frequent direction 'day' → 'sun' (93 out of 99 cases of the determined direction, see also [11: 291]). In India, both Indo-European and Dravidian languages, the sun's descriptive designations are "cause/lord/jewel/of the day" "day maker". A significant part of the cases of derivation (56 examples) is the model 'sun' = 'eye of the day', which is widespread in Southeast Asia [14]. Another frequent pattern of derivation (26 examples) is presented in the same word 'luminary' is used to designate both the sun and the moon, and if necessary, the expressions 'day luminary' and 'night luminary' are used, respectively. This colexification of the sun and moon is especially common in the Americas [13]. In some cases, the word 'luminary' may be lost, as in Cubeo *jāravu* 'day' → *aviá jāravucacu* 'sun' (lit. 'luminary of the day') → *jāravucacu* 'sun' (lit. 'of day') [8: 152, 355]. Rare examples of the direction 'sun' → 'day' are Tennet (South Sudan) *i*: 'sun', *i-tten* 'day', Malak-Malak (Australia) *miri* 'sun' and *mirien* 'day'.

References

- [1] Behnstedt Peterand, Woidich Manfred Wortatlas der arabischen Dialekte. Band I: Mensch, Natur, Fauna un Flora. 2010. Leiden – Boston. Brill.
- [2] Cabolov Ruslan 2010. Kurdish etymological dictionary [Ètimologičeskij slovar' kurdsckogo jazyka]. Vol. 2. Moscow.
- [3] Cincius Vera (ed.) Comparative dictionary of the Tungusic languages. Materials for the etymological dictionary [Srvnitel'nyj slovar' tunguso-man'čžurskix jazykov. Materialy k ètimologičesckomu slovarju]. Leningrad, 1975
- [4] Gibbs, George, Dictionary of the Chinook Jargon, or Trade Language of Oregon. Washington: Smithsonian Institution, 1863.
- [5] Johnstone Thomas Muir. Jibbali Lexicon. Oxford - New York, 1981.
- [6] Kogan Leonid. Lexical evidence and the genealogical position of Ugaritic (I) // Babel und Bibel 3, 429-488
- [7] Leslau Wolf . Lexique Soqotri (sudarabique moderne) avec comparaisons et explications ètymologiques. Paris 1938.
- [8] Morse Nancy L., Salser Jay K., Salser Neva. Diccionario ilustrado bilingüe: cubeo-español español-cubeo. Bogota. Editorial Alberto Lleras Camargo, 1999.
- [9] Perexval'skaja Elena. Russian pidgins [Russkie pidžiny]. Saint Petersburg, 2008
- [10] Roth Martha T. (ed.).1992. Assyrian Dictionary of the Oriental Institute of the University of Chicago, Volume 17, S, Part 1.
- [11] Starostin Georgij. Languages of Africa: an attempt at a lexicostatistical classification. Vol. I. Methodology. Khoisan languages [Jazyki Afriki. Opyt postroenija leksikostatističesckoj klassifikacii. T. 1: Metodologija. Kojsanskije jazyki]. Moscow. 2013.
- [12] Schmitt Rüdiger. 2000 Die iranischen Sprachen in Geschichte und Gegenwart, Wiesbaden: Reichert Verlag
- [13] Urban Matthias. 2009. 'Sun' and 'Moon' in the Circum-Pacific Language Area // Anthropological Linguistics, Volume 51, Numbers 3-4, Fall and Winter 2009: 328-346.
- [14] Urban Matthias 2010. 'Sun' = 'Eye of the Day': A Linguistic Pattern of Southeast Asia and Oceania // Oceanic Linguistics, Volume 49, Number 2, December 2010: 568-579.
- [15] Zalizniak Anna A., Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak, and Maxim Russo. 2012. The catalogue of semantic shifts as a database for lexical semantic typology // Linguistics, 50(3):633–669.
- [16] Zalizniak Anna A., 2018. The Catalogue of Semantic Shifts: 20 Years Later // Russian Journal of Linguistics. 22, 4: 770-787.

Generative Question Answering Systems over Knowledge Graphs and Text

Raushan Turganbay¹
raushan.turganbay@alumni.nu.edu.kz

Viacheslav Surkov¹
surokpro2@gmail.com

Dmitry Evseev¹
dmitriy.euseew@yandex.ru

Mikhail Drobyshevskiy²
drobyshevsky@ispras.ru

¹Neural Networks and Deep Learning Lab, MIPT, Dolgoprudny, Russia

²Ivannikov Institute for System Programming of the Russian Academy of Sciences

Abstract

In this paper we describe a generative question answering system which relies on text or knowledge graphs to find supporting evidence. The goal of generative QA is to provide a natural full sentence answer relying on the relevant evidence. Unlike existing models, the system proposed here can generate full answers using knowledge base triplets as evidence and is not restricted to simple questions consisting of one triplet. The generation module is a pretrained encoder-decoder transformer. Additionally, we constructed a new dataset DSberQuAD to train and evaluate the generative QA system in Russian. The new dataset was constructed in a rule-based manner and is an extension of SberQuAD with full sentence answers for each question. The proposed model is a new SOTA for Russian KBQA on RuBQ2.0 dataset. All the code and data from this project are available on GitHub ¹ under Apache license.

Keywords: generative question answering, knowledge base, entity extraction

DOI: 10.28995/2075-7182-2023-22-1112-1126

Генеративные вопросно-ответные системы по графам знаний и тексту

Раушан Турганбай¹
raushan.turganbay@alumni.nu.edu.kz
Дмитрий Евсеев¹
dmitriy.euseew@yandex.ru

Вячеслав Сурков¹
surokpro2@gmail.com
Михаил Дробышевский²
drobyshevsky@ispras.ru

¹Лаборатория нейронных систем и глубокого обучения, МФТИ,
Долгопрудный, Россия

²Институт системного программирования им. В.П.Иванникова
Российской академии наук

Аннотация

В данной статье мы описываем генеративную вопросно-ответную систему, которая опирается на текст или графы знаний для поиска подтверждающих доказательств. Целью генеративной вопросно-ответной системы является предоставление ответа на естественном языке в виде полного предложения, основанного на соответствующих доказательствах. В отличие от существующих моделей, предлагаемая здесь система может генерировать полные ответы, используя триплеты из базы знаний в качестве доказательства, и не ограничивается простыми вопросами, состоящими из одного триплета. Модуль генерации представляет собой предварительно обученный энкодер-декодер. Кроме того, мы предлагаем новый датасет DSberQuAD для обучения и оценки генеративных вопросно-ответных систем на русском. Новый датасет был построен на основе правил и является расширением SberQuAD с полными ответами на каждый вопрос. Предлагаемая модель является новой SOTA для русского на датасете RuBQ2.0. Весь код и данные этого проекта доступны на GitHub ¹ под лицензией Apache.

Ключевые слова: генеративная вопросно-ответная система, база знаний, извлечение сущностей

¹https://github.com/deeppavlov/explainable_qa

1 Introduction

Modern question answering systems have made significant advance in recent years and can find the most relevant answer to question in natural language (Rajpurkar et al., 2016), (Kwiatkowski et al., 2019), (Yang et al., 2019), (Yang et al., 2018). They resort either to knowledge base or unstructured text to find evidence for the answer. Yet, existing models return a short fragment of text as an answer. For example, given a question “What is the capital of France?” the short answer entity “Paris” will be returned, instead of full answer “The capital of France is Paris.”

To get full answers, model must be trained to generate a sequence of tokens using relevant facts. One area where generation models are commonly used is goal-oriented dialogue (dialogue goal can be: buy plane tickets, get bus schedules, etc.). In goal-oriented dialog the model should generate a response based on question and the information retrieved from the database. Thus, generating a full answer for factoids can be used in dialog assistants to make the conversation knowledge-grounded and more engaging (Lowe et al., 2015), (Ghazvininejad et al., 2018), (Dinan et al., 2018), (Liu et al., 2018).

This paper proposes QA systems that generate full sentence answers: based on text or knowledge bases (KBs). The model based on text is available for Russian, and models based on KB are available for Russian and English. Also, we propose a generative method on top of the QA system to get full answers. In case of the knowledge-based question answering (KBQA) models, full answer is obtained by feeding the path in the graph to encoder-decoder model trained to generate text from graph. Full answers for text-based model are generated by retrieving paragraphs relevant to the question and using it as input to a different encoder-decoder model that was trained on text data only. The system is built using DeepPavlov library (Burtsev et al., 2018). To train and evaluate more efficiently our text-based generative QA system in Russian, we built DSberQuAD dataset. It extends existing SberQuAD (Efimov et al., 2019) by adding full sentence answers for questions along with short answers.

In summary, we make the following contributions:

- (i) We propose a generative question answering model on unstructured text for Russian, which consists of retriever and generation module;
- (ii) We extend the traditional KBQA model with answer generation component, which generates full sentence answers from graph triplets;
- (iii) We introduce Russian QA dataset with full sentence answers to improve the development of generative QA systems in Russian;
- (iv) We introduce a new state-of-the-art on RUBQ2.0 for KBQA task in Russian.

2 Related work

Open domain Question Answering (ODQA) models based on text usually consist of a retriever that finds relevant paragraphs and a component for finding an answer. Relevant paragraphs can be found by calculating dot product between question and paragraph vectors, where the vectors can be sparse representations obtained by methods such as TF-IDF or BM25 (Chen et al., 2017), (Yang et al., 2019). Yet sparse vectorization methods do not perform well when paragraphs are semantically related to the question but do not have any word overlap, thus dense vectorization has been a common choice lately (Karpukhin et al., 2020), (Lee et al., 2019), (Guu et al., 2020). The answer span extraction can be solved using deep learning techniques (Seo et al., 2016), (Wang et al., 2017), (Devlin et al., 2018). These models take a concatenation of question and paragraphs, and assume that answer span is present in one of the paragraphs.

KBQA relies on knowledge graphs to find the correct answer. Knowledge graph is a database that stores information about the world in a structured way. Facts in the knowledge graph can be represented as triplets in <subject, relation, object> format. Existing systems use SPARQL queries to answer questions. WQAqua (Diefenbach et al., 2018) and QAMP (Vakulenko et al., 2019) systems start with KB grounding and then construct possible SPARQL queries that return non-empty answers when executed. The most probable queries are determined based on various parameters. On the other hand, NSQA (Kapanipathi et al., 2020) leverages the question’s syntactic structure to construct SPARQL

query. These systems do not perform well on complex questions in LC-QuAD and RuBQ2.0, QAMP works over DBpedia only.

To generate full sentence answers several methods were introduced. Answer generation based on text evidence in Tan et al. (2018) is done by adding a decoder on top of answer extraction model, so that the generation model can leverage relevant pieces of evidence. Later, Mitra (2018) expanded on that by adding a pointer-generator network that copies necessary information from relevant paragraph, thus there is no need to extract the answer span first. The task was further improved by using all relevant paragraphs when decoding by Dehghani et al. (2019). Later works approach the problem by using pretrained models. In Fusion-in-Decoder (Izacard and Grave, 2020) relevant documents are concatenated with question and fed into T5 for answer generation. RAG (Lewis et al., 2020), on the other hand, uses BART (Lewis et al., 2019) for generation.

Generative QA based on KBs likewise relies on seq2seq models. GenQA (Yin et al., 2016) retrieves relevant triplets from the graph and used seq2seq models for generation. Later, COREQA (He et al., 2017) extended it by applying a copying mechanism to incorporate relevant facts when generating. However, both systems work only with simple questions.

Here we propose separate generative QA models that rely either on text by generating full answers based on retrieved paragraphs or use KB triplets for generation. We expand on existing work by performing answer generation based on KBs that answers different types of complex questions, including multi-fact, multi-constraint and qualifier-constraint. Moreover, our system is available in Russian (as well as English), which we believe promotes the development of QA models for Russian.

3 Model

Given a question, our model first finds a short answer and then generates full answer using pretrained seq2seq transformers. Since they were pretrained on text only, using them directly to generate answers from KB triplets may not perform well. Thus, we use a model based on JointGT (Ke et al., 2021) which was pretrained jointly on text and graph data and takes into account graph structure. The details of each model's architecture will be outlined further.

3.1 Generative QA over knowledge graphs

The architecture of our KBQA, represented in Figure.1. The model supports Wikidata and DBpedia. To generate full answer, path in the graph from question node to answer node is obtained from SPARQL query and full sentence is generated from that path with a seq2seq model.

3.1.1 SPARQL query template prediction

Since information in KB is stored in a structured way, SPARQL queries are used to extract answers. This component takes a question as input and predicts the most likely SPARQL query from a predefined set of 25 queries. The English model is a BERT classifier fine-tuned on LC-QuAD for 5 epochs with a batch size of 32 and an initial learning rate of $1 \cdot 10^{-5}$.

For Russian model, we use rule-based approach due to lack of training data. The rules can be decomposed to 4 steps: 1) identify the type of answer entity using the relation "instance of"; 2) identify the entity from which the search in the graph will start; 3) answer set is refined using the modifiers in the question; 4) the most likely template is chosen based on the number of entities in the question;

3.1.2 Entity Extraction

Entity extraction converts unstructured text into structured data by finding entities mentioned in question and linking them to their unique identifiers in the graph. First, the entities are identified by classifying each token into an entity or non-entity with RoBERTa token classifier. The model was trained on data obtained from Wikipedia by using anchor texts as entities, as it was done in Ferragina and Scialla (2011).

The knowledge graph is stored in SQLite database as an inverted index which makes the search process faster and more efficient. Found entities are sorted in ascending order of Levenshtein distance between

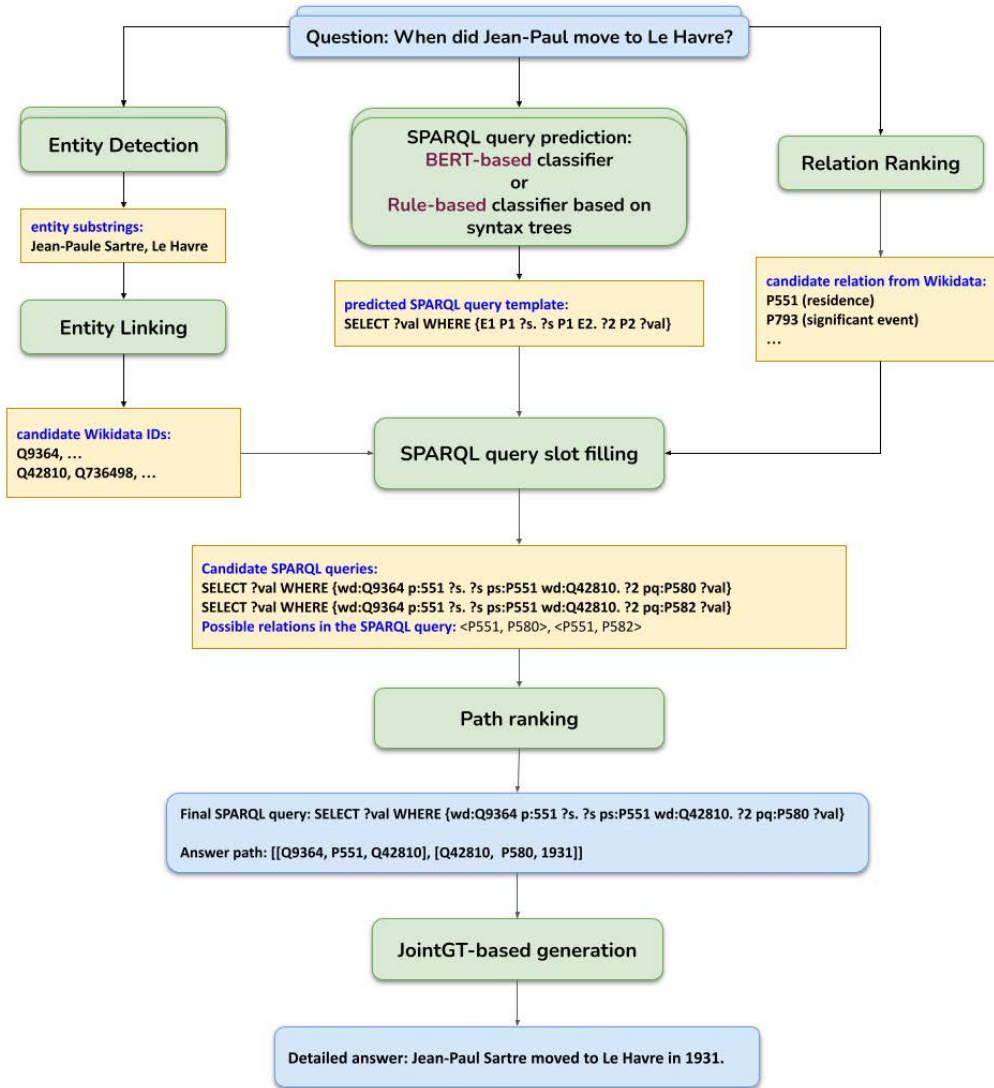


Figure 1: KBQA system

the identified entity substrings and the entity labels from the KB, and further ranked by dot product between the embeddings of the question and entity descriptions.

3.1.3 Relation Ranking

After identifying and linking entities from the question, all relations that are 1-2 hops away from the extracted entities are considered as candidates. They are further ranked using BERT-based model, which receives as input the question and relation separated by SEP token. The final CLS token representation is passed through a feed-forward layer to predict the probability. The model achieved 92% accuracy in the test set. Accuracy was calculated as the percentage of questions where the model predicted the ground truth entity with the highest probability.

3.1.4 Path Ranking

Path Ranking component is designed to determine the most likely combination of relations for SPARQL query by evaluating all possible permutations of candidate relations. It is a BERT-based model that takes as input question and relations separated by SEP token.

The vector of CLS token then goes through a fully connected layer to predict the probability of how well a given set of relations fits into the question. Training for each question uses 1 correct relation (positive sample) and 99 incorrect ones (negative sample). Negative log-likelihood was used as loss

function. The model was trained on LC-QUAD for 3 epochs with batch size of 20 and an initial learning rate of $1 \cdot 10^{-5}$, and reached accuracy=68.2% on validation set.

3.1.5 Full sentence answer generation from knowledge graph triplets

The final output of KBQA is a short answer entity, so we derive a path in the graph from starting entity to answer entity, which we use to generate full answers. The path is represented in triplets. For example, for question "Which sea is surrounded by Rostock and Kaliningrad?" the path will be: ("Baltic Sea", "cities", "Rostock"), ("Baltic Sea", "cities", "Kaliningrad"). Generating answers from triplets in graph can enhance our understanding of how the answer was obtained and might be used to improve the interpretability of KBQA models.

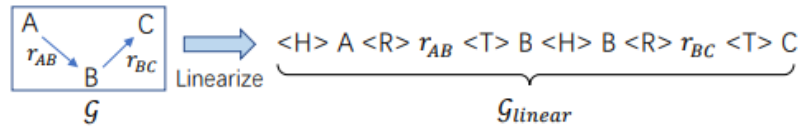


Figure 2: Special tokens <H>, <R> and <T> are used to indicate the head, relations and tail in a triplet.

In this work, the generative component is based on JointGT (Ke et al., 2021), which is a BART-based model with attention between entities and relations in the triplet. First, the path consisting of triplets is linearized as shown in Figure 2. Then, the linearized graph is passed to an encoder with a self-attention layer that captures the interaction between entities and relations in the graph. Finally, encoder output is passed to a decoder to generate an answer.

3.2 Generative QA over text

ODQA system finds answer to the question by searching a large collection of unstructured text, Wikipedia paragraphs in our case. We again use a seq2seq model to generate a full answer, yet the input in this case is a text fragment instead of a path from the graph. The system consists of a retriever, an answer span extractor and generation module, as it can be seen in Figure 3.

3.2.1 Retriever

The retriever component finds top N relevant paragraphs from Wikipedia paragraphs by calculating the scalar product between the vectors of the question and paragraphs. The paragraphs are vectorized using TF-IDF. Each Wikipedia paragraph is tokenized into n-grams ($n = 1, 2$) and then the hash values of n-grams are calculated using `murmurhash3_32`² from sklearn. We get a sparse index where rows correspond to paragraphs and columns to n-gram hash values. The question is vectorized in a similar manner.

The paragraphs are then ranked with BERT classifier to narrow down the list. Each paragraph is classified as relevant or not with certain probability. To train the model, paragraphs that contain an answer were used as positive examples and the rest of the paragraphs as negative examples.

3.2.2 Answer span extraction

To find an answer span from a set of relevant paragraphs, we use a model based on Deepavlov RuBERT (Kuratov and Arkhipov, 2019) and fine-tuned it on SberQUAD. The model input is a question and paragraph separated by [SEP] token. Answer start and end position are predicted by linear transformation of model output. The list of paragraphs where the answer was predicted are used for subsequent answer generation.

3.2.3 Full sentence answer generation from text

For answer generation we experiment with multiple seq2seq backbone transformers. Additionally, we research the effects of transfer learning, by first training the model on MS-MARCO (Nguyen et al., 2016), an extensive English QA dataset. All the seq2seq models were trained for three epochs with

²`sklearn.utils.murmurhash3_32`

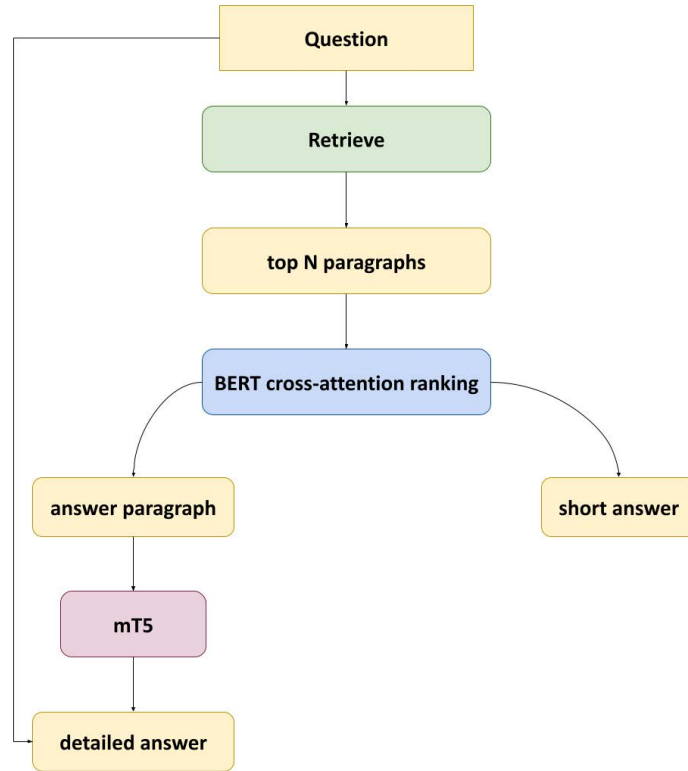


Figure 3: Diagram of the system components. The text-based question answering system finds a short answer, which is further used to generate full answer.

batch size 40 using AdamW optimizer with $\text{learning_rate}=3 \cdot 10^{-4}$. Below are the three strategies we employed:

1. Multilingual seq2seq model mT5-small (Xue et al., 2021) was fine-tuned on DSberQuAD dataset
2. mT5-small was fine-tuned on DSberQuAD dataset preceded by pretraining on MS MARCO to improve the performance with transfer learning methods.
3. Russian Language model RuT5-base³ was fine-tuned on DSberQuAD dataset.

Furthermore, we experimented with non-seq2seq methods. We used BERT-based models to extract a short answer substring and further process it with another BERT to generate a long answer. Specifically, we used Deepavlov RuBERT-base (Kuratov and Arkhipov, 2019), DistilRuBERT and DistilRuBERTiny (Kolesnikova et al., 2022) fine-tuned on SberQuAD for answer extraction. The second BERT relies on syntax parser based on Deep Biaffine Attention to construct a long answer on rule-based manner. The details of long answer construction are outlined in Section 4.

4 Datasets

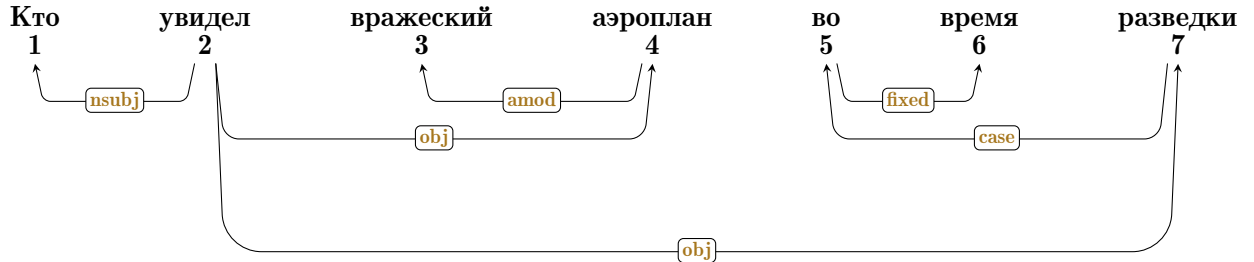
Apart from the model, we present here a new dataset for generative QA which we used for training and evaluation purposes. We hope it will promote the development of generative QA models in Russian. It is a modified version of SberQuAD, we call it DSberQuAD. The dataset creation procedure can be divided into two steps: (1) Replace a question word in the question (e.g. *who*, *when*, *why*) with the short substring answer; (2) Form a sentence that follows grammatical agreement rules and change to the SVO word order.

For each question and its short answer, their dependency trees are constructed using a syntax parser based on Deep Biaffine Attention (Dozat and Manning, 2017) for building syntax tree and pymorphy2 library (Korobov, 2015) for adjusting cases and numbers. A full answer is constructed based on position and type of question word *qword* in the text. Totally 10 cases were analyzed, two of which will be

³<https://huggingface.co/sberbank-ai/ruT5-base>

described below. Please refer to Appendix A to see all cases.

In case when $qword$ is a number/case form of $\{кто, что\}$, the $qword$ or the whole subtree of the $qword$ is replaced by short answer. In case if the $qword$ or answer are not in nominative case, the sentence is slightly modified to follow case agreement rules. For example, in question "Кто увидел вражеский аэроплан во время разведки?", the dependency tree of which is illustrated below, $qword$ is "Кто" and the answer is "летчики". To construct a long sentence answer, $qword$ is replaced by the answer, and then, since the answer is a plural noun, the verb is changed to its plural form, yielding a full answer sentence — "Летчики увидели вражеский аэроплан во время разведки."



Another case when $qword \in \{когда\}$, it is simply replaced by answer substring. If the answer is a year or percentage, nouns such as "году" or "процентов" are added to the answer. For example, given a question "Когда железная дорога соединила Тильзит и Клайпеду?" and an answer "в 1875", the constructed long answer will be "В 1875 году железная дорога соединила Тильзит и Клайпеду.". It can be noticed, the $qword$ "когда" was replaced by short answer, and since the question is asking for a year, "году" was appended to the answer substring.

Additionally, we manually rewrote answers for a subset of 100 random question from LC-QuAD (Trivedi et al., 2017), which we used to evaluate the performance of the generative model from triplets. For example, given triplets from the KB [["Google Videos", "developer", "Google"], ["Google Web Toolkit", "author", "Google"]], we rewrote the long answer as "Google Videos and Google Web Toolkit were developed by Google".

5 Main Results

Generative model based on text was tested on DSberQuAD validation set. Apart from standard metrics such as BLEU or ROUGE, for evaluation of our model, we introduce *ROUGE-1 lemma* metric. Given as input generated long answer and reference short answer substring, the score is calculated as intersection of short answer tokens in long answer tokens after lemmatizing all words.

The results are presented in Table 1. The highest scoring BERT two-stage approach outperforms RuT5, it is relatively inefficient in terms of the number of parameters. The multilingual T5 performs better when it is additionally pretrained on MS-MARCO, which implies that transfer learning improves the scores (Chizhikova et al., 2023). The best trade-off between the number of parameters and performance is achieved by RuT5 that has the highest ROUGE-1 lemmas score with the least number of parameters used.

Model	Parameters	ppl	ROUGE-1 lemma	SacreBLEU
mT5-small	300M	1.78	48.8	36.8
mT5-small + MS MARCO	300M	1.25	75.9	57.3
ruT5-base	222M	1.22	81.0	70.2
BERT-base+BERT-base	180M+180M	—	88.9	87.9
BERT-base+DistilBert6L	180M+135.4M	—	88.1	82.4
BERT-base+DistilBert2L	180M+107M	—	77.3	64.4

Table 1: Performance metrics of various strategies used for answer generation on DSberQuAD dataset.

Since we have not encountered baselines for generative KBQA on complex questions, we are comparing of our KBQA system without the generative component. We compare our Russian model with

QAmp (Vakulenko et al., 2019), WQAqua (QAnswer) (Diefenbach et al., 2018) and Simba. Table 2 illustrates the evaluation results on RuBQ2.0 (Rybin et al., 2021) dataset, which consists of 2330 questions in test-set. As evaluation metric we used accuracy, which is calculated as the percentage of correct answer entities out of gold answer entities. It can be noted that our KBQA system for Russian outperforms existing approaches, thus achieving SOTA performance on RuBQ2.0. High performance of our KBQA compared to other solutions can be explained by the following:

- We account for the syntax tree when choosing SPARQL query template, which yields higher accuracy, allowing to choose the most possible template;
- We can link entities more accurately since we go beyond term-level matching, and rank entities using the context and entity descriptions;
- To train our system’s components we fine-tune a pretrained BERT-base model thus leveraging its language understanding ability;

Question type (quantity)	DeepPavlov	QAnswer	SimBa
questions with answer (1920)	56.0	26.9	25.3
1-hop (1460)	61.1	30.8	32.2
1-hop + reverse (10)	0	0	0
1-hop + count (3)	66.6	33.3	0
1-hop + exclusion (17)	17.7	5.9	0
multi-constraint (304)	50.3	19.7	3.6
multi-hop (55)	1.8	10.9	1.8
qualifier-constraint (22)	45.5	0	4.5
questions without answers (410)	39.8	5.9	86.8
total (2330)	53.1	23.2	36.1

Table 2: Comparison of performance metrics on RuBQ2.0 for each question type in the dataset.

For evaluation of our English KBQA system, we use LC-QuAD dataset (Trivedi et al., 2017), which consists of 5000 questions over DBpedia and their respective SPARQL queries. We compare against the following baselines: QAmp, WQAqua and NSQA (Kapanipathi et al., 2020). Performance scores are represented in Table 3. The proposed system surpasses existing baselines on KBQA task. Appendix B shows examples of answers generated by our system.

Model	Precision	Recall	F1 score
QAmp	25.0	50.0	33.0
WQAqua	22.0	38.0	28.0
NSQA	45.0	46.0	44.0
DeepPavlov	45.0	50.0	47.0

Table 3: Performance metrics of KBQA model on LC-QUAD.

To evaluate the generative module, we scored generated full answers against the target by calculating the BLEU score. We used gold triplets as input to the generative component thus evaluating only the generative component. The generative component achieves BLEU-1=70.1 and BLEU-2=56.2 on LC-QuAD. The model’s performance can be improved by forcing it to use all entities in the generated text. Thus, method for controllable generation of full answers based on triplets should be examined in future studies.

6 Conclusion and Future Work

In this work, we have described the QA system which is able not only to give a short answer to the question, but also to generate full detailed sentence, thus making the answer more engaging and informative for users. Our system derives answers either from knowledge graph or from unstructured text. In comparison to existing methods (GenQA (Yin et al., 2016), COREQA(He et al., 2017)), our generative

module over KBQA can generate answers to complex questions. For training and evaluation of text-based QA model for long answer generation in Russian, we devised a new dataset DSberQuAD based on SberQuAD. Our system achieves competitive performance on existing QA datasets and the Russian KBQA system achieves SOTA on RuBQ2.0.

One feasible future research direction is to use heterogeneous knowledge sources for generative QA. Although knowledge base approaches excel in addressing complex questions, their effectiveness is frequently hindered by the incomplete nature of the KB. On the other hand, web text consists of numerous facts that are not present in KB, but they are often disorganized and lacking structure. We leave the task of fusing information from both knowledge sources for future work.

References

- Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. // *Proceedings of ACL 2018, System Demonstrations*, P 122–127, Melbourne, Australia, July. Association for Computational Linguistics.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.
- Anastasia Chizhikova, Vasily Konovalov, and Mikhail Burtsev. 2023. Multilingual case-insensitive named entity recognition. // Boris Kryzhanovskiy, Witali Dunin-Barkowski, Vladimir Redko, and Yury Tiumentsev, *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, P 448–454, Cham. Springer International Publishing.
- Mostafa Dehghani, Hosein Azarbyad, Jaap Kamps, and Maarten de Rijke. 2019. Learning to transform, combine, and reason in open-domain question answering. // *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, P 681–689.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dennis Diefenbach, Andreas Both, Kamal Deep Singh, and Pierre Maret. 2018. Towards a question answering system over the semantic web. *arXiv preprint arXiv:1803.00832*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.
- Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.
- Pavel Efimov, Leonid Boytsov, and Pavel Braslavski. 2019. Sberquad - russian reading comprehension dataset: Description and analysis. *arXiv preprint arXiv:1912.09723*.
- Paolo Ferragina and Ugo Scaiella. 2011. Fast and accurate annotation of short texts with wikipedia pages. *IEEE software*, 29(1):70–75.
- Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.
- Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. 2017. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. // *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 199–208, Vancouver, Canada, July. Association for Computational Linguistics.
- Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

- Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramon Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, et al. 2020. Leveraging abstract meaning representation for knowledge base question answering. *arXiv preprint arXiv:2012.01707*.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. *arXiv preprint arXiv:2106.10502*.
- Alina Kolesnikova, Yuri Kuratov, Vasily Konovalov, and Mikhail Burtsev. 2022. Knowledge distillation of russian language models with reduction of vocabulary. *1905.07213 arXiv:2205.02340*.
- Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. // Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, P 320–332. Springer International Publishing.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.
- Tom Kwiakowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 6086–6096, Florence, Italy, July. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.
- Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. // *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1489–1498.
- Ryan Thomas Lowe, Nissan Pow, Laurent Charlin, and Joelle Pineau. 2015. Incorporating unstructured textual knowledge sources into neural dialogue systems.
- Rajarshee Mitra. 2018. A generative approach to question answering. *arXiv preprint arXiv:1711.06238*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, P 2383–2392, Austin, Texas, November. Association for Computational Linguistics.
- Ivan Rybin, Vladislav Korablinov, Pavel Efimov, and Pavel Braslavski. 2021. Rubq 2.0: An innovated russian question answering dataset. // Ruben Verborgh, Katja Hose, Heiko Paulheim, Pierre-Antoine Champin, Maria Maleshkova, Oscar Corcho, Petar Ristoski, and Mehwish Alam, *The Semantic Web*, P 532–547, Cham. Springer International Publishing.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.
- Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2018. S-net: From answer extraction to answer generation for machine reading comprehension. *arXiv preprint arXiv:1706.04815*.
- Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. Lc-quad: A corpus for complex question answering over knowledge graphs. // *International Workshop on the Semantic Web*.

- Svitlana Vakulenko, Javier David Fernandez Garcia, Axel Polleres, Maarten de Rijke, and Michael Cochez. 2019. Message passing for complex question answering over knowledge graphs. // *Proceedings of the 28th acm international conference on information and knowledge management*, P 1431–1440.
- Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. // *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 189–198.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. // *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 483–498, Online, June. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.
- Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.
- Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural generative question answering. *arXiv preprint arXiv:1512.01337*.

Appendix A Cases analyzed when constructing the DSberQuAD dataset

Below are presented the 10 cases we analyzed when constructing the DSberQuAD dataset. If the question does not fit into any of the cases, the answer is left without changes.

Case 1 When the $qword \in \{\kappa mo, \psi mo\}$ and it is the root of the syntactic tree, the $qword$ (and sometime the whole subtree of the $qword$ is replaced by the answers a . Also, a hyphen is added. (Table 1).

Q	Кто такой Пушкин?
A	Великий русский поэт
LA	Пушкин - это великий русский поэт.

Table 1: Case 1 for constructing DSberQuAD.

Case 2 $qword$ — If it is the case form of the $qword \kappa mo, \psi mo$, the $qword$ (and sometime the whole subtree of the $qword$) is replaced by the answer. The answer's case and number is changed to align with the case and number of the $qword$ and the necessary prepositions are added. If the $qword$ is not the subject of the question, then the answer or the whole subtree is place after the predicate (Table 2).

Q	Чем заболел Байрон в Миссолонги?
A	лихорадкой
LA	Байрон заболел лихорадкой в Миссолонги.
Q	Кто увидел вражеский аэроплан во время разведки?
A	летчики
LA	Лётчики увидели вражеский аэроплан во время разведки.
Q	Что было открыто во дворце после восстановления?
A	Национальная галерея
LA	Национальная галерея была открыта во дворце после восстановления.

Table 2: Case 2 for constructing DSberQuAD.

Case 3 If the $qword$ is the case form of $какой, чей$, depending on whether the answer is a noun phrase, an adjective or a date, the $qword$ (or the hole subtree) is replaces with the answer. Case and number of the answer is changed according the agreement rules, necessary prepositions are added. The sentence order is changes to *Subject-Verb-Object*. Additional post-processing was done on questions that start with "*Какое название/наименование...*"; "*В каком году...*"; "*Чей...*" (Table 3).

Q	В каком году отец Оригена был убит?
A	В 202 году
LA	Отец Оригена был убит в 202 году.
Q	В каком замке был заключен Дидро?
A	Венсенском
LA	В венсенском замке был заключен Дидро.
Q	Чьи ротовые аппараты разнообразны?
A	Чешуекрылых
LA	Ротовые аппараты чешуекрылых разнообразны.

Table 3: Case 3 for constructing DSberQuAD.

Case 4 If the $qword$ is a case form of $каков$, the $qword$ is replcaed by the answer. The word order remains the same (Table 4).

<i>Q</i>	Какова температура кипения воды?
<i>A</i>	100 градусов
<i>LA</i>	100 градусов — температура кипения воды.

Table 4: Case 4 for constructing DSberQuAD.

Case 5 When $qword = когда$, the $qword$ is replaced by the answer. The word order remains the same. Additional post-processing is added if the answer is a year. In that case the answer is changed to have the format "*В X году*" (Table 5).

<i>Q</i>	Когда мосты стали строить из железобетона?
<i>A</i>	В XX веке.
<i>LA</i>	В XX веке мосты стали строить из железобетона.
<i>Q</i>	Когда железная дорога соединила Тильзит и Клайпеду?
<i>A</i>	в 1875
<i>LA</i>	В 1875 году железная дорога соединила Тильзит и Клайпеду.

Table 5: Case 5 for constructing DSberQuAD.

Case 6 When $qword \in \{где, куда, откуда, докуда\}$, the $qword$ is replaced by the answer. The word order is changed to *Subject-Verb-Object* (Table 6).

<i>Q</i>	Куда переехала Фанни из Сан-Франциско?
<i>A</i>	в Монтерей
<i>LA</i>	Фанни переехала в Монтерей из Сан-Франциско.

Table 6: Case 6 for constructing DSberQuAD.

Case 7 When $qword \in \{почему, отчего\}$, answers are not rewritten in full forms since the original answers are already full sentences.

Case 8 When $qword = как$, the root of the T_Q is placed at the end of the sentence. The question word is removed and the answer is appended to the end of sentence. If question starts with "*Как быстро... Как долго...*", these are assumed to be the $qword$ and removed all together. If the $qword$ is "*Как переводится... Как описывается*" or the answer is a noun in accusative case, then a word *так* is added to the answer (Table 7).

<i>Q</i>	Как отводятся излишки тепла у млекопитающих?
<i>A</i>	потоотделением
<i>LA</i>	Излишки тепла у млекопитающих отводятся потоотделением.
<i>Q</i>	Как переводится слово каллиграфия?
<i>A</i>	Путь письма.
<i>LA</i>	Слово каллиграфия переводится как путь письма.
<i>Q</i>	Как быстро протекает митоз?
<i>A</i>	1–2 часа.
<i>LA</i>	Митоз протекает 1–2 часа.

Table 7: Case 8 for constructing DSberQuAD.

Case 9 When $qword = сколько$, the $qword$ is replaced by the answer. To make the sentence grammatically correct, additional post-processing is done if the question asks for the percentage or degree values. Prepositions are added or deleted and the case/number forms are changed to follow the agreement rules (Table 8).

<i>Q</i>	При сколько градусах плавится альфа-цирконий?
<i>A</i>	1855 °C
<i>LA</i>	При 1855 °C плавится альфа-цирконий.
<i>Q</i>	Сколько астероидов обнаружено в настоящий момент в солнечной системе?
<i>A</i>	Сотни тысяч астероидов.
<i>LA</i>	Сотни тысяч астероидов обнаружены в настоящий момент в солнечной системе.
<i>Q</i>	Сколько человек родилось в 2008 году у граждан Швейцарии?
<i>A</i>	1147 человек
<i>LA</i>	1147 человек родилось в 2008 году у граждан Швейцарии.

Table 8: Case 9 for constructing DSberQuAD.

Case 10 When $qword = mu$, the $qword$ is removed and its parent p is replaced by the answer. The word order is changed to *Subject-Verb-Object* (Table 9).

<i>Q</i>	Разная ли величина приемлемого риска для различных обществ?
<i>A</i>	различная
<i>LA</i>	Величина приемлемого риска для различных обществ различная.
<i>Q</i>	Гигроскопичен ли германат висмута?
<i>A</i>	абсолютно негигроскопичен
<i>LA</i>	Германат висмута абсолютно негигроскопичен.

Table 9: Case 10 for constructing DSberQuAD.

Appendix B Examples of answers by our English KBQA system in comparison to the compared baselines.

Here we illustrate examples of how our system generates full sentence answers to questions using path triplets in the graph. As i was mentioned, using path triplets for generation can enhance the interpretability of the KBQA system by demonstrating how complex questions were broken into multiple triplets. Table 1 shows some examples of questions from LC-QuAD dataset, short answer entity found by our system, path triplets in (subject, relation, object) format and the generated full answer.

Question	Answer	Triplets	Generated Text
Is Essex the Ceremonial County of West Tilbury?	Yes	<'West Tilbury', 'Ceremonial County', 'Essex'>	West Tilbury is in the County of Essex.
What nation is Hornito located in, where Jamie Bateman Cayn died too?	Panama	<'Jaime Bateman Cayón', 'death place', 'Panama'>, <'Hornito, Chiriquí', 'country', 'Panama'>	The death place of Jaime Bateman Cayón is Hornito, Chiriquí, Panama.
Who are the shareholder of the soccer club for whom Steve Holland plays?	Roman Abramovich	<'Steve Holland', 'current club', 'Chelsea F.C.'>, <'Chelsea F.C.', 'owner', 'Roman Abramovich'>	Steve Holland's current club is Chelsea F.C. which is owned by Roman Abramovich.
Who is the chancellor of Falmouth University?	Dawn French	<'Falmouth University', 'chancellor', 'Dawn French'>	The chancellor of the Falmouth University is Dawn French.
What is the religious affiliations of Katyayana?	Buddhism	<'Katyayana', 'religion', 'Buddhism'>	The religion of Katyayana is Buddhism.

Table 1: Examples of long answers generated by our system on LC-QuAD dataset.

Development of a Morphological Analyser for Siberian Ingrian Finnish

Ivan Ubaleht

Omsk State Technical University /
644050, Russia, Omsk, Mira 11
ubaleht@gmail.com

Abstract

This paper presents our work on the development of a morphological analyzer for Siberian Ingrian Finnish. Siberian Ingrian Finnish is a low-resource language. In this paper, we present an algorithm for analyzing nouns of Siberian Ingrian Finnish and show an example of analysis.

Keywords: morphological analyzer; low-resource language; natural language processing; Siberian Ingrian Finnish

DOI: 10.28995/2075-7182-2023-22-1127-1132

Разработка морфологического анализатора для сибирского ингерманландского идиома

Иван Убалехт

Омский государственный технический университет /
644050, Россия, г. Омск, пр-т Мира, д. 11
ubaleht@gmail.com

Аннотация

Статья посвящена разработке морфологического анализатора для языка ингерманландских переселенцев в Сибири. Данный язык является малоресурсным, это вносит специфику в разработку программного обеспечения для него. В статье представлен алгоритм анализа слов, относящихся к имени существительному, и рассмотрен пример анализа слова.

Ключевые слова: морфологический анализатор; малоресурсные языки; обработка естественного языка; сибирский ингерманландский идиом

1 Введение

Малоресурсные языки (low-resource languages) характеризуются наличием одного или нескольких из ниже перечисленных свойств: отсутствием стабильной письменности и орфографии; отсутствием или небольшим количеством ресурсов в электронном виде; отсутствием или небольшим количеством программного обеспечения для работы с языком [1].

В главе 2 описывается проблема, присущая малоресурсным языкам, обладающим только следующими типами ресурсов: неаннотированные аудио (видео) данные; лингвистическое описание, выполненное в форме, непригодной для компьютерной обработки. Рассматриваемый в статье сибирский ингерманландский идиом относится к этой группе малоресурсных языков. В главе 3 дано краткое описание сибирского ингерманландского идиома, для которого разрабатывается морфологический анализатор и дана характеристика его ресурсов. Глава 4 посвящена рассмотрению морфологического анализатора. Рассмотрен алгоритм и пример анализа слова, принадлежащего к имени существительному.

2 Обзор работ и постановка задачи

Языки, относящиеся к группе малоресурсных языков, обеспечены ресурсами в разной степени. Например, такие языки как карельский [2] или якутский [3] обладают относительно большим количеством текстов различных типов (периодика, массивы сообщений из социальных сетей и т.д.), на основе которых можно строить языковую модель. Наличие языковой и акустической моделей позволит разработать для этих языков системы автоматического распознавания речи (далее САРР) и другое сложное программное обеспечение.

Существует другая довольно большая группа малоресурсных языков, для которых существует только два типа ресурсов: неаннотированные аудио (видео) данные, например, данные из экспедиций; лингвистическое описание, выполненное в форме, непригодной для компьютерной обработки, например, в «бумажной» форме или в виде текста в форматах PDF, Word и т.д. Как видно, обладая ресурсами данных типов невозможно работать с такими языками в электронной среде. Таких идиомов существует довольно много, например, идиомы, описанные в рамках экспедиций, полевой материал из которых хранится в архивах. Важной научной задачей является более широкое введение данных таких языков в научный оборот.

Единственный способ получить возможность работы с такими языками в электронной среде – это построение корпуса текстов через аннотирование аудио (видео) данных. Как известно [4], создание аннотаций для аудиоданных – это чрезвычайно трудоёмкий процесс: на аннотирование одного часа аудиоданных может уйти от 40 до 100 часов (это так называемая проблема «annotation bottleneck»).

Так как размеченные тексты для рассматриваемого типа малоресурсных языков можно получить только из аудиоданных, то хорошей идеей было бы разработка САРР для этих языков, которая бы автоматизировала, хотя бы частично, процесс получения новых текстовых данных из аудио. Так как САРР обычно строятся на основе принципов машинного обучения, то для малоресурсных языков такой подход почти неосуществим, но, тем не менее, есть работы [1,5], которые рассматривают применение САРР для малоресурсных языков. В работе [6] описывается способ получения первоначального обучающего набора для малоресурсных языков. Суть подхода заключается в том, что вместо того чтобы аннотировать аудиоданные из экспедиций, в которых речь может быть спонтанной и не систематизированной, можно сначала взять качественные тексты, потом их озвучить носителями языка и синхронизировать речь с текстом. При таком подходе можно гораздо быстрее получить первый обучающий набор данных.

Чтобы использовать для сибирского ингерманландского идиома САРР и описанный в [6] способ получения первого обучающего набора данных, нужен исходный набор текстов, который можно получить для данного языка только из аннотаций. Поэтому, схема подхода обеспечения ресурсами сибирского ингерманландского идиома и языков с похожим исходным набором ресурсов может быть следующей:

- построение морфологического анализатора и генератора на основе лингвистического описания языка, это позволит работать с языком компьютерными методами;
- использование этого анализатора и генератора для получения словоформ, аугментации данных, расставления PoS тегов, автоматизации части рутинных операций при аннотировании аудио;
- интеграция с системами Apretium [7] и HFST [8], что поможет работать с фразами и автоматизировать перевод текста целевого малоресурсного языка на русский или английский языки;
- получив набор текстов, можно применить метод получения первого обучающего набора данных для САРР, описанный в источнике [6];
- разработка САРР, которая поможет быстрее получить большое количество новых текстов из оставшихся аудиоданных.

3 Сибирский ингерманландский идиом

3.1 Краткая характеристика

Язык ингерманландских переселенцев в Сибири или сибирский ингерманландский идиом – это смешанный язык, основанный на нижнелужских вариантах ижорского и финского языков. Исследователями также отмечается наличие водского субстрата. В современном состоянии отмечается влияние на данный язык эстонского и русского языков. Предки носителей данного языка мигрировали в Сибирь из финских и ижорских деревень, находящихся в районе нижнего течения реки Луги, в 1803-1804 гг.

В настоящее время языковой коллектив, использующий данный идиом в сфере бытового общения, сохраняется только в селе Рыжково в Омской области. Отдельные носители данного языка проживают также в населённых пунктах рядом с селом Рыжково, в городе Омске, в нескольких других населённых пунктах Омской области, а также в Эстонии.

Термин «сибирский ингерманландский идиом» (Siberian Ingrian Finnish) был введён Д.В. Сидоркевич, которая исследовала и документировала данный идиом в 2008-2014 гг. [9,10]. В настоящее время этот идиом продолжает исследовать Н.В. Кузнецова [11]. Язык ингерманландских переселенцев в Сибири обладает довольно значительным количеством фонологических, морфологических, морфологических особенностей и предоставляет важный материал, например, для сравнительного анализа с другими языками прибалтийско-финской группы. Разработка программных инструментов для работы с этим языком будет полезна для работы с большими массивами аудиоданных таких близких идиомов, как финские и ижорские диалекты района нижнего течения реки Луги.

3.2 Обзор ресурсов данного языка

Обзор ресурсов, доступных для сибирского ингерманландского идиома, представлен в Таблице 1. Можно добавить, что для данного языка существует озвученный словарь в виде веб-приложения, разработанного автором статьи. Веб-приложение доступно в сети Интернет¹, исходный код этого веб-приложения находится в открытом доступе² на GitHub. Аннотации аудиоданных в формате ELAN³, составленные автором статьи, а также другие материалы, включая исходный код морфологического анализатора, опубликованы в рабочем репозитории проекта⁴ на GitHub под лицензией Creative Commons 4.0.

Тип ресурса	Объём ресурса
Аудиоданные, собранные Д.В. Сидоркевич в 2008-2014 гг.	80 часов
Аудиоданные, собранные автором статьи в 2019-2023 гг.	20 часов
Аудиоданные, опубликованные в открытом доступе под свободной лицензией Creative Commons 4.0	5 часов
Видеоданные	2 часа
Тексты, в основном транскрипция аудиоданных, сделанная вручную	42 тысячи токенов
Аннотированные в ELAN слова	200 слов
Число носителей языка, речь которых записана	31 человек

Таблица 1: Обеспечение ресурсами сибирского ингерманландского идиома

Как видно из Таблицы 1, данный идиом не обладает значительным набором текстов. Для дальнейшей работы с этим идиомом компьютерными методами необходима языковая модель, которую построить без наличия текстов невозможно. Единственный способ получить языковую модель этого языка – это составить набор аннотаций аудиоданных, но как отмечалось во второй главе, процесс аннотирования аудиоданных занимает большое время. Сократить время

¹ <http://lexeme.net/sif>

² <https://github.com/ubaleht/Lexeme>

³ <https://archive.mpi.nl/tila/elan>

⁴ <https://github.com/ubaleht/SiberianIngrianFinnish>

аннотирования, автоматизировать часть ручных операций позволит морфологический анализатор для данного языка.

4 Разработка морфологического анализатора

К настоящему моменту разработан модуль для морфологического анализатора, работающий со словами, принадлежащими к именным частям речи (существительные, прилагательные, местоимения, числительные).

На рисунке 1 показан алгоритм анализа слов, принадлежащих к имени существительному. Вход алгоритма – пять основ анализируемого слова: *NOM.SG* – анализируемое слово в номинативе, в единственном числе (слово в данной форме является леммой); *PRT.SG* – основа слова для получения слова в партитве в единственном числе; *ILL.SG* – основа слова для получения слова в иллативе в единственном числе; *OBL.SG* – косвенная основа слова в единственном числе, на основе её строятся все формы слова в единственном числе во всех косвенных падежах, кроме иллатива и партитива, а также для получения формы слова в номинативе во множественном числе; *OBL.PL* – косвенная основа слова во множественном числе. Альтернативный вход алгоритма – это только форма слова в *NOM.SG*. Выход алгоритма – размеченная строка в формате XML, содержащая все словоформы анализируемого слова, информацию о морфологии, метаданные о данном слове (код информанта, перевод слова, контекст записи и т.д.). В дальнейшем структурированная информация из формируемых XML файлов будет отправлена в базу данных на основе СУБД MS SQL Server.

```

INPUT:  - Либо слово W в форме NOM.SG.
           - Либо W в форме NOM.SG с основами OBL.SG, PRT.SG,
           ILL.SG, OBL.PL

OUTPUT: Все словоформы для W с морфологическим описанием
           в формате XML.

1. IF W только в форме NOM.SG
   THEN BEGIN
3.  Побуквенный анализ W, выбор одного морфонологического
   типа FROM CS1 TO CS8 OR FROM VS1 TO VS8;
4.  Для формирования каждой из основ OBL.SG, PRT.SG, ILL.SG,
   OBL.PL выбор морфонологического преобразования
   FROM CA1 TO CA5 OR FROM VA1 TO VA5 AND применение
   эвристических правил из множества E;
   END
6.  Добавление падежных показателей к основам OBL.SG, PRT.SG,
   ILL.SG, OBL.PL и формирование для W форм слова в 10 падежах
   и 2 числах;
5.  Запись всех словоформ W и морфологического описания в XML
   файл.

```

Рисунок 1: Алгоритм формирования словоформ для существительных

В настоящий момент данная часть морфологического анализатора с библиотекой анализа слов, относящихся к именным частям речи, реализована с помощью платформы .NET и языка C# как приложение для Windows. Приложение имеет пользовательский интерфейс для удобного, быстрого ввода имеющихся слов, принадлежащих к имени существительному и имени прилагательному. В дальнейшем планируется интеграция с Apertium и HFST для работы с фразами и переводами. В алгоритме на Рисунке 1 представлен анализ слов, принадлежащих к имени существительному.

Работа со словами, принадлежащими к имени прилагательному, осуществляется с помощью этого же алгоритма. Морфонологические типы у имён прилагательных такие же, как и у имён существительных – добавляется лишь несколько особенностей, например, анализ степеней

сравнения прилагательных. Для местоимений и числительных, которые используют те же парадигмы, анализ с помощью алгоритма не нужен, так как все их словоформы известны.

Падеж	Единственное число			Множественное число		
	Основа	Суффикс	Пример	Основа	Суффикс	Пример
Номинатив	NOM.SG	∅	<i>kukk</i>	OBL.SG	- <i>t</i>	<i>kuka-t</i>
Генетив	OBL.SG	- <i>n</i>	<i>kuka-n</i>	OBL.PL	- <i>n</i>	<i>kukki-n</i>
Партитив	PRT.SG	∅	<i>kukka</i>	OBL.PL	- <i>j</i>	<i>kukki-j</i>
Иллатив	ILL.SG	∅	<i>kukka</i>	OBL.PL	- <i>s</i>	<i>kukki-s</i>
Инессив	OBL.SG	- <i>s</i>	<i>kuka-s</i>	OBL.PL	- <i>s</i>	<i>kukki-s</i>
Элатив	OBL.SG	- <i>st</i>	<i>kuka-st</i>	OBL.PL	- <i>st</i>	<i>kukki-st</i>
Адессив – Аллатив	OBL.SG	- <i>l</i>	<i>kuka-l</i>	OBL.PL	- <i>l</i>	<i>kukki-l</i>
Аблатив	OBL.SG	- <i>lt</i>	<i>kuka-lt</i>	OBL.PL	- <i>lt</i>	<i>kukki-lt</i>
Транслатив	OBL.SG	- <i>ks</i>	<i>kuka-ks</i>	OBL.PL	- <i>ks</i>	<i>kukki-ks</i>
Комитатив	OBL.SG	- <i>nka</i>	<i>kuka-nka</i>	OBL.PL	- <i>nka</i>	<i>kukki-nka</i>

Таблица 2: Парадигма склонения для морфонологического типа CS1 с примером слова *kukk*, относящимся к этому морфонологическому типу

В качестве примера анализируемого слова возьмём слово *kukk* – «цветок». Чтобы получить все словоформы данного слова в соответствии с алгоритмом на Рисунке 1 необходимо: установить морфонологический тип и тип морфонологических преобразований. В настоящий момент для сибирского ингерманландского идиома определено 16 морфонологических типов [10 стр. 163, 170-171]. Восемь типов для слов имеющих консонантную основу (consonant stem) CS1 – CS8 и восемь типов для слов, имеющих гласную основу (vowel stem) VS1 – VS8. В области морфонологических преобразований существуют консонантные чередования (consonant alternation) CA1 – CA5 и вокалические чередования (vowel alternation) VA1 – VA5.

Если для анализируемого слова, например, для слова *kukk* известны пять основ – NOM.SG, PRT.SG, ILL.SG, OBL.SG, OBL.PL, то на основе их можно получить все словоформы добавлением падежных показателей к соответствующим основам, см. Таблицу 2.

Если для слова в форме NOM.SG неизвестны другие основы, то их можно получить с помощью алгоритма на Рисунке 1. Например, слово *kukk* относится к морфонологическому типу CS1 [10, стр. 169]. Алгоритм на рисунке 1 определяет это после побуквенного анализа слова. Тип CS1 определяется следующим образом: «...согласная основа с первичной геминантой, простым консонантным кластером, кластером с геминантой или одиночным согласным, имеющим слабую ступень чередования, в ауслауте...» [10, стр. 164]. К этому типу относятся также: *jalk* «нога», *harakk* «сорока», *penkk'* «скамейка», *rankk* «тяжёлый», *huntt'* «волк», *luut* «веник», *piipp'* «трубка», *kant* «ствол» и т.д. Для типа CS1 являются возможными следующие морфонологические преобразования:

- для формирования основы OBL.SG это правило CA1 или правило VA1;
- для формирования основ PRT.SG и ILL.SG это правило VA1;
- для формирования основы OBL.PL это правило VA2.

Основа OBL.SG была получена с помощью преобразования CA1: «...чередование ступеней последнего консонанта основы...» [10, стр. 161], *kukk* → *kuka-*, см. Таблицу 2. Основы PRT.SG и ILL.SG формируются с помощью преобразования VA1: «появление тематического гласного» (только для согласных основ) [10, стр. 161], в данном случае появление тематического *a* у основ PRT.SG и ILL.SG, *kukk* → *kukka*, см. Таблицу 2. Основа OBL.PL формируется согласно правилу VA2: «мутация тематического гласного» [10, стр. 162], в данном случае *a* мутирует в направлении *i*, см. Таблицу 2.

Почти для каждого из морфонологических типов существуют альтернативные правила формирования, так в текущем примере для типа CS1 это CA1 и VA1 для формирования основы OBL.SG. Для разрешения этой неоднозначности в алгоритме применяются эвристики, и сформированные словоформы помечаются как реконструированные.

5 Заключение и будущие работы

В статье показаны текущие результаты работы над морфологическим анализатором для сибирского ингерманландского идиома. В рамках проекта морфологического анализатора разработана библиотека, позволяющая формировать все словоформы для именных частей речи.

В дальнейшем планируется:

- разработать библиотеку для работы с глаголами этого языка;
- сформировать словарь на несколько тысяч лексем со всеми словоформами для данного языка;
- интегрировать этот проект, с проектами Apertium и HFST;
- унифицировать обозначения элементов морфологической разметки, используемых в морфологическом анализаторе для будущей интеграции с проектом Universal Dependencies⁵;
- использовать этот морфологический анализатор для автоматизации аннотирования аудиоданных, что должно обеспечить выигрыш во времени по сравнению с ручным аннотированием, например, с помощью ELAN.

References

- [1] Besacier, Laurent, Barnard Etienne, Karpov Alexey, Schultz Tanja. Automatic speech recognition for under-resourced languages: A survey. — Speech communication, 2014. Vol. — 56.
- [2] Boyko Tatyana, Zaitseva Nina, Krizhanovskaya Natalia, Krizhanovsky Andrew, Novak Irina, Pellinen Nataliya, Rodionova Aleksandra // The Open corpus of the Veps and Karelian languages: overview and applications. Computing Research Repository. — 2022. — Vol. arXiv:2206.03870. — version 1. Access mode: <https://arxiv.org/ftp/arxiv/papers/2206/2206.03870.pdf>
- [3] Ivanova, Sardana, Jonathan Washington, and Francis M. Tyers. A free/open-source morphological analyser and generator for Sakha // Proceedings of LREC 2022, Thirteenth International Conference on Language Resources and Evaluation. European Languages Resources Association (ELRA). 2022. — P. 5137–5142.
- [4] Seifart Frank, et al. Language documentation twenty-five years on. — Language, 2018. Vol. — 94(4).
- [5] Prud'hommeaux, Emily, et al. Automatic speech recognition for supporting endangered language documentation. — Language documentation and conservation, 2021. Vol. — 15.
- [6] Ćavar Malgorzata, Ćavar Damir, Cruz Hilaria. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR // Proceedings of the tenth international conference on language resources and evaluation (LREC'16). 2016. — P. 4004–4011.
- [7] Forcada Mikel L., et al. Apertium: a free/open-source platform for rule-based machine translation. — Machine translation, 2011.
- [8] Lindén Krister, Silfverberg Miikka, Pirinen, Tommi. Hfst tools for morphology—an efficient open-source package for construction of morphological analyzers // Proceedings of State of the Art in Computational Morphology: Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009). — Zurich, Switzerland, 2009. — P.28–47.
- [9] Sidorkevich Daria 2011. On domains of adessive-allative in Siberian Ingrian Finnish. — Acta Linguistica Petropolitana, 2011. Vol. — 7(3).
- [10] Сидоркевич Д. В. Язык ингерманландских переселенцев в Сибири: структура, диалектные особенности, контактные явления. Кандидатская диссертация. — СПб: Институт лингвистических исследований РАН, 2014. Режим доступа: <https://iling.spb.ru/theses/1999>
- [11] Kuznetsova Natalia. Evolution of the non-initial vocalic length contrast across the Finnic varieties of Ingria and adjacent areas. — Linguistica Uralica, 2016. Vol. — 52(1).

⁵ <https://universaldependencies.org/>

The Universal Database for Lexical Typology

Ekaterina Voloshina
University of Gothenburg
Gothenburg, Sweden
vokat@mail.ru

Polina Leonova
HSE University
Moscow, Russia
00leonovapolina@gmail.com

Abstract

The paper presents the principles of creating a database for research in lexical typology and describes the possibilities of its use as a linguistic resource. The database is built around semantic fields and frames, i. e. units of analysis in the frame-based theory of lexical typology.

The database provides a universal format for storing the data; therefore, any project in lexical typology can be easily added. The database does not only store the data from previous research projects but allows anyone who wants to contribute to submit data via its web interface. The database includes examples provided by native speakers and manually annotated with translations, semantic fields, and frames, following the annotation principles adopted within the frame approach to lexical typology.

Keywords: lexical typology, corpus linguistics, computational lexicography

DOI: 10.28995/2075-7182-2023-22-1133-1140

База данных для лексико-типологических исследований¹

Екатерина Волошина
Гётеборгский университет
Гётеборг, Швеция
vokat@mail.ru

Полина Леонова
НИУ ВШЭ
Москва, Россия
00leonovapolina@gmail.com

Аннотация

В статье представлены принципы создания базы данных для исследований в области лексической типологии и описаны возможности ее использования в качестве лингвистического ресурса и инструмента для сбора и анализа материала. База данных построена на основе семантических полей и фреймов, т.е. единиц, на которых основан лексико-типологический анализ в рамках фреймового подхода.

База предполагает универсальный формат хранения данных, поэтому любой проект по лексической типологии может быть легко в нее добавлен. База данных не только содержит материал предыдущих исследовательских проектов, но и позволяет любому желающему внести новые данные, используя специально разработанный веб-интерфейс. В базе хранятся примеры, полученные от носителей языка и аннотированные вручную: для каждого примера приводится его перевод на русский язык, семантическое поле, к которому относится иллюстрируемая лексическая единица, и соответствующий примеру фрейм.

Ключевые слова: лексическая типология, корпусная лингвистика, компьютерная лексикография

¹Статья подготовлена в ходе проведения исследования № 23-00-012 «Смежность семантических полей в типологической перспективе» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)

1 Introduction

The cross-linguistic analysis of lexicon usually requires both a sufficient amount of data and adequate annotations to it. For semantic research, as for linguistic research in general, it is convenient to structure and store this information in databases (in our case - lexical databases). However, as opposed to phonetic or morphological databases, the lexical ones inevitably reflect theoretical assumptions of their developers. In particular, FrameNet (Baker et al., 1998) represents the frame approach to semantics, constructicons (Lyngfelt et al., 2018) provide data annotated in the tradition of Construction Grammar (Fillmore and Kay, 1995), etc. Even CLICS (Rzymiski et al., 2020), which seems to be theory neutral, relies on a certain categorisation of semantic domains and concepts within them, and thus selects just one research solution out of many available. Indeed, a detailed analysis of linguistic data within the framework of Moscow Lexical Typology Group (MLexT) ² results in other sets of categories for the same semantic domains. The fact is that CLICS automatically aggregates data from available wordlists, whereas the MLexT approach involves a manual and data-driven detection of relevant concepts, or, in terms of MLexT, frames (for details, see (Rakhilina and Reznikova, 2016)).

In this paper, we present a database which incorporates the principles of MLexT approach and implements several functions related to data collection, aggregation of primary data, editing the possible set of questions, and data analysis. For data collection any researcher can add their data in a unified format filling a questionnaire and, after moderation, the data are added directly to the database and become available for search. Besides that, moderators can edit questionnaires and uploaded data, as well as add general descriptions and observations related to semantic fields. The database includes a search engine with two modes to cover all possible queries for the purposes of research in both lexical typology and semantics in general.

2 Lexical typology: theory and methods

For a long time lexicon has not been studied from a typological perspective. The first attempts in this area cover just a few semantic domains, namely, color, kinship and body-part terms (Berlin and Kay, 1969; Keesing, 1975; Andersen, 1978). Over the last decades, however, the scope of lexical typology has expanded to encompass a whole range of domains, cf. verbs of cutting and breaking (Majid et al., 2007), verbs of motion, e.g. motion in water (Majsak, 2007), rotation (Krugljakova, 2010), falling (Reznikova et al., 2020), pain predicates (Reznikova et al., 2012), verbs of putting and taking (Kopecka and Narasimhan, 2012), physical qualities (Koptjevskaja-Tamm, 2015; ?), etc. These studies differ in methodology of data collection and analysis. The approach we follow here, as already stated, is the one developed by MLexT, as it turns to be more versatile than other techniques in this area (for details, see (Rakhilina and Reznikova, 2016)).

Within the MLexT approach, semantic domains are compared in terms of frames, i.e. prototypical situations that are expressed through words of a given semantic field.

The process of establishing the frames starts with browsing through corpora and dictionaries. As a result, relevant contexts for each frame and parameters of their opposition are identified. For example, for verbs of motion in water a relevant parameter would be the opposition of active and passive motion, for verbs of rotation – the type of axis (inner or outer) and the type of rotating object, for verbs of change – the degree of change (full versus partial change). It does not mean that these parameters should be relevant for every language: frames that are lexically opposed in one language can be colexified in another. Hypotheses generated on this stage become the basis for a context-based questionnaire consisting of a list of sentences with gaps which are supposed to be filled with a lexeme from the field under study. The number of sentences is determined by the number of possible oppositions which are relevant for the field.

Questionnaires are used for verifying data collected from corpora and dictionaries, and for collecting data for the languages that do not have such resources. This method of simultaneous work with corpora and native speakers for each language allows us to collect data from various languages, including endangered ones.

²<https://lextyp.org>

Language	Field	Frame	Context	Verb	Example	Translation	Source
Spanish	change	partial change of an object	It was obvious that something must be ___ in the legislation.	cambiar	No hay planes de cambiar la legislación en ese sentido, al menos a corto plazo	Nikakih izmenenij v zakonodatel'stvo v jetoj oblasti, po krajnej mere v kratkosročnoj perspektive, ne planirujetsja.	speakers
French	change	full change of an object	He ___ linoleum for parquet.	changer	Il a changé le linoléum contre le parquet.	On pomenjal linoleum na parket.	speakers
Karel	change	partial change of an object	The chef was angry with the cook because he ___ the recipe.	muuttua	"muuttele syömistä, eli syö yhtä ta samua	Raznoobraz' edu, ne esh' odno i to zhe	corpus

Table 1: An annotation example.

At the same time, every study conducted in this framework involves different language experts dealing with various resources. The database storing the up-to-date version of the questionnaire and the skeleton of the field structure would ensure compatibility of data coming from different languages and facilitate the process of data collection and analysis.

3 Database

3.1 Annotation

The database includes the data from questionnaires filled in by native speakers or examples from corpora. Each example is translated to Russian and annotated in the following way:

- **Language:** the original language of the example;
- **Semantic field:** a generalized part of lexicon the example belongs to, e.g., ‘falling’, ‘hiding’, ‘changing’, etc.;
- **Frame:** one of the prototypical meanings included in a semantic field, e. g. the semantic field of falling can be divided into four main frames: ‘loss of vertical orientation’, ‘detachment’, ‘falling from elevated surface’, and ‘crashing down’ ((Reznikova et al., 2020));
- **Context:** if the example comes from a questionnaire, context is a given stimulus; if the example is taken from a corpus, it is the closest context to the one in the example;
- **Verb:** a verb that belongs to a given semantic field and is used in the example;
- **Source type:** if the example comes from a corpus or was given by a native speaker.

The annotation examples are given in Table 1.

3.2 Database Structure

The Mongo database is built in a way to represent the annotation described above. It is essential that users can search for any word or word part in examples and their translations; therefore, a non-SQL Mongo database was implemented.

The database schema is represented in Figure 1. The database includes several ways of searching and aggregating the data: two search modes, creating semantic maps and semantic field profiles (for details see Section 4). Some search modes use different types of information (e.g. one can only use fields and languages). To make the search process more efficient, the database is decomposed into several tables instead of storing all information in one.

The database includes five collections: *contexts*, *fields*, *frames*, *languages*, and *verbs*. In languages collections, an id and the text name of an item are given. Fields collection, in addition to an id and a text name, includes an overall description of this field. Frames collections are connected with fields, as they belong to one field. Contexts include the list of frames, since they can belong to several frames. Verbs collection includes a verb and the language it is taken from and a list of examples where each example’s annotation includes the example itself, the translation, the source type, and the context id, which connects the example to frames and semantic fields.

```
1  {
2    "contexts":{
3      "_id": ObjectId,
4      "context": string,
5      "frame": string
6    },
7    "fields":{
8      "_id": ObjectId,
9      "field": string,
10     "description": string,
11   },
12   "frames":{
13     "_id": ObjectId,
14     "frame": string,
15     "field": ObjectId
16   },
17   "languages":{
18     "_id": ObjectId,
19     "lang": string
20   },
21   "verbs":{
22     "_id": ObjectId,
23     "verb": string,
24     "lang_id": ObjectId,
25     "examples":{
26       "example": string,
27       "translation": string,
28       "source": string,
29       "context": ObjectId
30     }
31   }
32 }
```

Listing 1: The database schema

4 Web-interface and Usage Cases

The database has the user interface which allows for working in three different scenarios: as a data collector, as a database editor, and as a researcher. Therefore, our database covers all stages of working with data: collection, processing and analysis.

As for technical details, the web-interface is written as a web application in Python on the basis of the Flask framework ³ and Mongo-DB API. ⁴

4.1 Scenario I: Data Collection

One of the most important parts of research in lexical typology is data collection. The data is collected through a questionnaire which consists of a set of contexts related to the given semantic field (for an example, see Appendix 5). To make the answers to the questionnaire comparable, it is important not only to store them in the same format but also to collect them in a similar setup.

The database supports two main forms of submitting the data: via manual form editing or uploading the data in the table format (*csv* or *xlsx*). In the first case, a researcher chooses a project to which they want to contribute data and then the form with questions is generated automatically based on the pre-uploaded questionnaires. Every question requires an example with a translation and allows to add a comment. Additionally, a researcher can put some extra examples in the free form.

The second option is to upload a spreadsheet that must follow the template with the questionnaire to be further processed.

The data is uploaded automatically to the database after it is approved by a moderator to exclude data in wrong formats or a scam data since the form can be filled without logging into the database.

4.2 Scenario II: Database Editing

As mentioned above, the data could be added by any user, therefore, only few people can get access to the database: to add, edit or delete files. It is also important that the moderator would have an expertise in lexical typology, so they could exclude non-relevant examples from the filled questionnaires. Besides that, a moderator might fill in information about frames, as frames are usually determined empirically, on the basis of collected data.

The moderator checks that all the fields are filled correctly and the data can be uploaded to the database. Before uploading to the database, the data is stored as files on the server, which can be edited through the web-interface.

Moreover, the moderator can edit or delete data from the database if the data appears to be outdated. Besides that, the editor's functions include adding descriptions about projects, and in the future they will be expanded so editors could add meta-information about languages in projects (for example, to specify the type of lexicalisation system in a given language).

4.3 Scenario III: Research

The main instrument for the research is the database search engine. There are two main search modes: the full text search supported by MongoDB and search by filters. The full-text mode allows searching by words or word parts in examples and translations. The filters are made to query the database annotation, therefore, there are 6 main search criteria that should be selected from a list: semantic field, language, frame, context, verb, and source type. Both types can be combined.

Besides the search engine, the database allows to build automatic semantic maps (Haspelmath, 2003) that aggregate information uploaded into the database. Semantic maps represent how different frames are connected within the semantic field. In the graph, nodes are frames or contexts (depending on the desired granularity), and they are connected if there is a verb that can be used in both frames (or contexts). The frequency of such connections are edge weights.

However, fully connected graphs, or vacuous maps, are less informative, as the more edges a graph has, the less combinations of senses within one lexeme it excludes.

³<https://flask.palletsprojects.com/en/2.0.x/>

⁴<https://pymongo.readthedocs.io/en/stable/index.html>

Figure 1: Web-interface for queries

Therefore, semantic maps are maximum spanning trees built with Kruskal's algorithm on the basis of uploaded examples. In other words, instead of graphs built on the basis of the data, semantic maps are subsets with the greatest weight and the minimum number of edges.

While building a semantic map, it is possible to determine a subset of languages that will be taken into account.

Moreover, the database includes subpages for all uploaded projects. Every project corresponds to one semantic field. The description is added manually by a moderator while the other information, such as frames constituting this field, languages and related projects are added automatically on the basis of uploaded examples.

As an example of the possible research, the base can be used for solving problems that are relevant for lexical typology and lexical system theory in general. For instance, it can be used to find out which factor makes the greater contribution to the type of lexical system – areal or genealogical.

5 Conclusion

In this paper, we present a database for research in lexical typology. While previous databases are built specifically for certain semantic fields⁵, the database described in this paper is universal in that it unifies the data from different semantic fields. It is especially necessary since the number of projects has increased significantly. Storing all the data in the same database allows to solve new research questions, e.g. which fields are related to each other in the cross-linguistic perspective, or what has a greater impact on the degree of similarity between lexical systems - their genetic or areal closeness.

From the practical point of view, the database is meant to make the process of data collection easier due to the possibility of data contribution and universal format of data storage. The database will be published on <https://linghub.ru/> but it can already be accessed through local interface published here: <https://anonymous.4open.science/r/LexTypDB-4C8C>.

References

- Elaine S Andersen. 1978. Lexical universals of body-part terminology. *Universals of human language*, 3:335–368.
- Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. // *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.
- Brent Berlin and Paul Kay. 1969. *Basic color terms: Their universality and evolution*. University of California Press.

⁵<http://www.web-corpora.net/zvukimu/>, <https://linghub.ru/aquamotion/>

- Charles J Fillmore and Paul Kay. 1995. Construction grammar. *Language*, 64(501-538):30.
- Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. // *The new psychology of language*, P 217–248. Psychology Press.
- Roger M Keesing. 1975. *Kin groups and social structure*. Holt, Rinehart and Winston New York.
- Anetta Kopecka and Bhuvana Narasimhan. 2012. *Events of putting and taking: A crosslinguistic perspective*, volume 100. John Benjamins Publishing.
- Maria Koptjevskaja-Tamm. 2015. *The linguistics of temperature*, volume 107. John Benjamins Publishing Company.
- V. A. Krugljakova. 2010. *Semantika glagolov vraščenija v tipologičeskoj perspektive*. Ph.D. thesis, (RGGU).
- Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent. 2018. *Constructicography: Constructicon development across languages*, volume 22. John Benjamins Publishing Company.
- Asifa Majid, Melissa Bowerman, Miriam van Staden, and James S Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective.
- Raxilina E. V. Majsak, . . . 2007. Glagoly dviženija v vode: leksičeskaja tipologija. verbs of motion in water: lexical typology.
- Ekaterina Rakhilina and Tatiana Reznikova. 2016. A frame-based methodology for lexical typology. *The lexical typology of semantic shifts*, 58:95–129.
- Tatiana Reznikova, Ekaterina Rakhilina, and Anastasia Bonch-Osmolovskaya. 2012. Towards a typology of pain predicates. *Linguistics*, 50(3):421–465.
- TI Reznikova, EV Rakhilina, and DA Ryzhova. 2020. Verbs of falling in the languages of the world: Frames, parameters, and types of the systems. *RUSSIAN ACADEMY OF SCIENCES*, P 10.
- Christoph Rzymiski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Appendix

A. The Sample Questionnaire

The screenshot displays the LexTyp web interface. At the top, there is a header bar with the text "LexTyp" on the left and a hamburger menu icon on the right. Below the header, the main content area is divided into two sections, each with a light blue background. The first section contains the sentence "The waiter ___ the vases around during the cleaning." followed by three input fields: "Sentence:", "Translation:", and "Comments:". The second section contains the sentence "We don't know what will happen in half a year: a lot can ___, and we'll be having other plans." followed by three input fields: "Sentence:", "Translation:", and "Comments:". Each input field is a simple white rectangle with a thin border. The "Comments:" fields have a small pencil icon in the bottom right corner, indicating they are text areas.

Figure 2: Web-interface for queries

Pre-editing Strategy Based on Automatic Evaluation of Translation Complexity to Improve the Quality of Specialized Texts Machine Translation into English

Alena A. Zhivotova
Komsomolsk-na-Amure State
University, Komsomolsk-na-Amure,
Russia
zhivotova.aa@gmail.com

Victor D. Berdonosov
Komsomolsk-na-Amure State
University, Komsomolsk-na-Amure,
Russia
berd1946@gmail.com

Abstract

The study addresses the issue of applying optimizing pre-editing of Russian-language texts in order to improve the quality of machine translation into English. A probabilistic assessment of translation task complexity is proposed to be used for selecting a pre-editing strategy. A generalized model of the translation process is presented. A mathematical model and algorithm for automated assessment of translation task complexity are proposed. Test of the model on specialized texts of oil and gas industry is described, which showed that the estimate correlates with an estimate of translation quality and can be used in selecting a strategy for optimizing pre-editing of source texts in machine translation tasks.

Keywords: machine translation, optimizing pre-editing, translation task complexity, translation quality

DOI: 10.28995/2075-7182-2023-22-1141-1149

Стратегия предредактирования исходного текста на основании автоматической оценки сложности задачи перевода для повышения качества машинного перевода узкоспециальных текстов на английский язык

Алена Животова
Комсомольский-на-Амуре
государственный университет,
Комсомольск-на-Амуре, Россия
zhivotova.aa@gmail.com

Виктор Бердонос
Комсомольский-на-Амуре
государственный университет,
Комсомольск-на-Амуре, Россия
berd1946@gmail.com

Аннотация

Исследование рассматривает вопрос применения оптимизационного предредактирования русскоязычных текстов с целью повышения качества машинного перевода на английский язык. Для выбора стратегии предредактирования предлагается использовать вероятностную оценку сложности задачи перевода. Представлена обобщенная модель процесса перевода. Предложены математическая модель и алгоритм автоматизированной оценки сложности задачи перевода. Описано тестирование модели на узкоспециальных текстах нефтегазовой тематики, которое показало, что данная оценка коррелирует с оценкой качества перевода и может быть использована при выборе стратегии оптимизационного предредактирования исходных текстов в задачах машинного перевода.

Ключевые слова: машинный перевод, оптимизационное предредактирование, сложность задачи перевода, качество перевода

1 Введение

Рассматривая вопрос качества машинного перевода (МП) для конечного реципиента следует учитывать следующие факторы: специфика предметной области и компетенция пользователя МП.

Специфика предметной области имеет ключевое значение, ведь МП тем эффективнее, чем больше обучающих данных (корпусов) загружено в систему, однако для некоторых предметных областей собрать достаточный объем двуязычных корпусов проблематично ввиду ограничений конфиденциальности данных и секретности разработок. Так, например, нефтегазовый сектор – один из ключевых для экономики нашей страны с большой долей участия иностранных компаний в проектах освоения месторождений и нефтегазопереработки. Качество перевода в данной области имеет критическое значение для коммуникации и обмена технологиями.

Поскольку для корректного использования систем МП необходимо прямое участие человека на всех этапах работы системы, справедливо, что для достижения оптимального результата работы, пользователь должен знать хотя бы один язык из языковой пары перевода, и чем лучше знание языка перевода пользователя, тем точнее будет оценка качества МП и его пост-редактура. Предоставляя пользователю средства обработки текста на языке, носителем которого он является, на любом из этапов перевода, можно повысить качество перевода. Зная ключевые параметры текста и их связь с предполагаемой оценкой качества, становится возможным предложить алгоритмы и инструменты редактирования текста с целью его оптимизации по критерию максимизации качества перевода.

Идея автоматического и полуавтоматического оптимизационного редактирования текста в задачах МП лежит в основе интерактивного МП. Основная масса работ и исследований в области интерактивного МП посвящены постредактированию, в том числе, его автоматизации [1, 2]. Предредактированию посвящено меньше работ, однако существующие статьи указывают на эффективность такого подхода к повышению качества перевода.

Так, Hiraoka и Yamada [3] в своей работе предприняли попытку сформулировать основные правила предредактирования текста при переводе с японского языка на английский. Их стратегия показала статистически значимые результаты, в том числе и для китайского и корейского языков. Подобные исследования проводились для перевода с японского на английский и восточные языки [4], с французского на английский [5], с индонезийского на английский [6], с английского на испанский [7]. Все указанные выше исследования показали, что предредактирование и переписывание исходного текста, опираясь на правила, повышает качество МП. Seretan, Bouillon и Gerlach в своем исследовании [8] показали, что использование даже простых полуавтоматических правил предредактирования текста повышает качество статистического МП. Исследование Шей [9] в языковой паре китайский-английский показало преимущества и позитивное влияние предредактирования на качество МП для пользователей с низким уровнем владения языком перевода в сравнении с постредактированием. Отмечается необходимость анализа исходного текста с целью определить «уязвимости» с точки зрения применяемой технологии МП, на основе которого становится возможной разработка правил предредактирования. Gerlach, Porro, Bouillon и Lehmann [5] показали, что такой подход позволяет сократить время пост-редактуры МП в два раза и в 65% случаев повышает качество результата МП.

Цель исследования – разработать стратегию оптимизационного предредактирования исходных текстов для повышения качества машинного перевода узкоспециальных текстов с русского на английский язык. Решение задачи повышения качества перевода узкоспециальных текстов позволит оптимизировать затраты на перевод, повысить надежность существующих систем, снизить зависимость качества перевода от человеческого фактора.

2 Модель процесса перевода

Для решения поставленной задачи проведено подробное математическое моделирование процесса и основных понятий перевода на основе теории множеств и построение процессной модели перевода. Обобщенная модель процесса перевода представлена на рисунке 1.

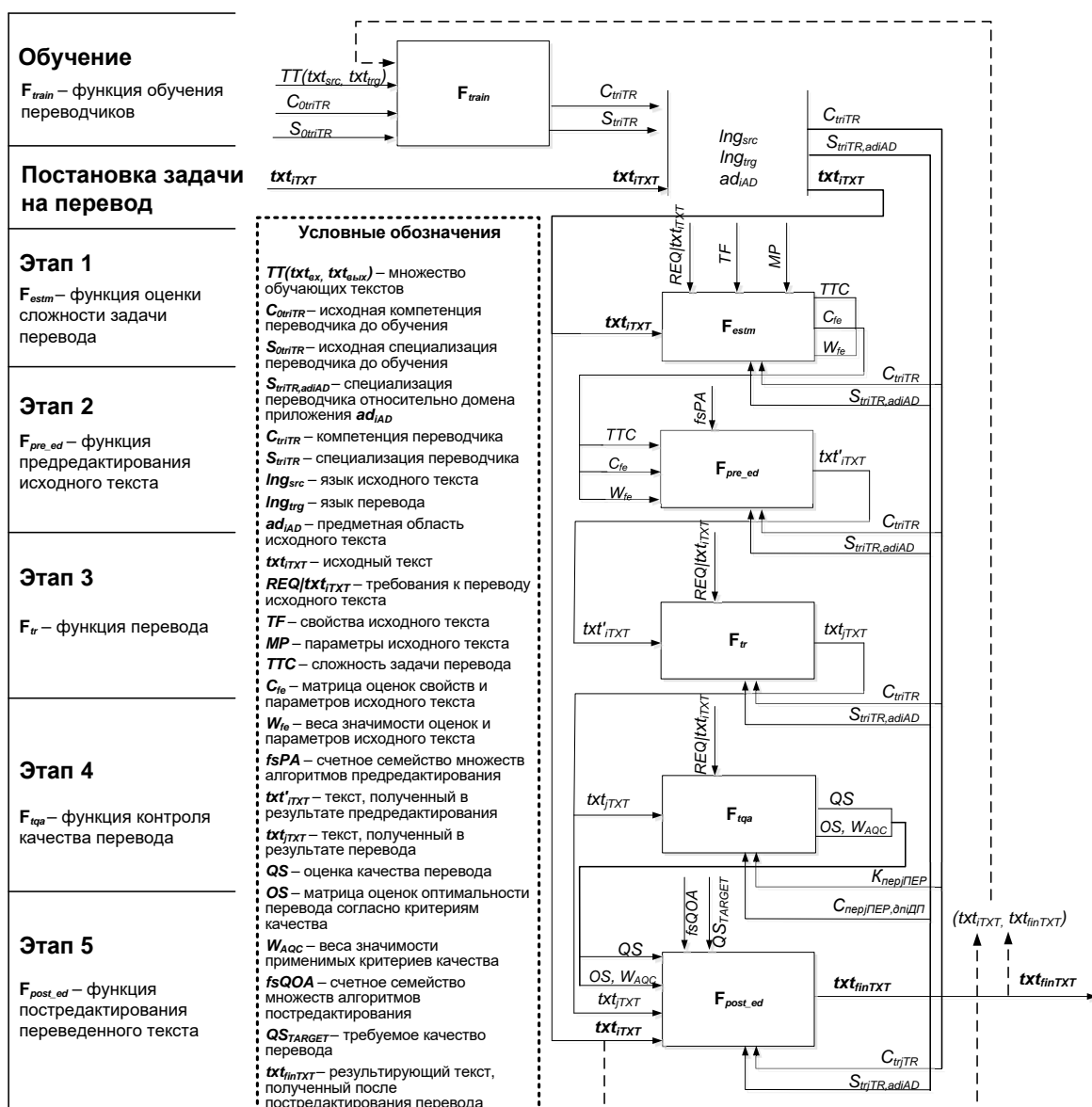


Рисунок 1: Обобщенная модель процесса перевода

В результате моделирования определено, что в системах МП не реализован этап переводческого процесса, который выполняется при «ручном переводе», а именно оценка сложности задачи перевода. На этом этапе переводчик оценивает вероятность получения качественного перевода, то есть соответствующего требованиям заказчика, и, если эта вероятность низкая, выбирает стратегию оптимизации исходного текста с целью повышения вероятности получения качественного перевода. Например, он расшифровывает аббревиатуры, применяет переводческие трансформации к исходному тексту и т.д. Результат работы МП – черновик, который пользователь должен оценить и доработать самостоятельно. При этом он должен обладать высоким уровнем компетенций в языке перевода, что ограничивает применение МП, особенно в узкоспециальных областях знаний, для которых не собраны достаточные корпуса для тренировки моделей перевода, и для которых требуется более тщательная проверка переведенного текста. Для разработки стратегии оптимизационного предредактирования была поставлена задача в первую очередь разработать способ оценки сложности задачи перевода и предполагаемого качества перевода для систем МП.

3 Вероятностная оценка сложности задачи перевода

Переводчик/система МП tr_{iTR} получает текст txt_{iTXT} на языке lng_{src} для перевода на язык lng_{trg} и требования к переводу $REQ|txt_{iTXT}$. При оценке сложности задачи перевода переводчик обращает внимание на неизвестные ему слова и сочетания слов на языке lng_{src} , для которых он не может идентифицировать значение смысловой единицы, либо смысловые единицы, для которых он не может найти аналог на языке перевода lng_{trg} среди известных ему слов и сочетаний слов.

Множества свойств и параметров исходного текста $TF|txt_{iTXT}$ и $MP|txt_{iTXT}$, и то, обладает ли переводчик достаточной компетенцией $\overline{C}r_{iTR}$ относительно языков lng_{src} и lng_{trg} и специализацией $\overline{S}tr_{iTR, ad_{iAD}}$, т.е. навыками описания семантических единиц на языке перевода в рамках заданной предметной области исходного текста ad_{iAD} , определяет вероятность создания переводчиком переведенного текста на таком уровне качества, который определяется требованиями $REQ|txt_{iTXT}$.

Алгоритм оценки сложности задачи перевода:

Шаг 1. Исходя из домена приложения текста ad_{iAD} , формируется множество оценок текста $FE = TF \cup MP$.

Шаг 2. Для каждого значения $tf_{iTF}, mp_{iMP} \in FE$, на основе требований к переводу $REQ|txt_{iTXT}$, компетенций переводчика относительно языковой пары $\overline{C}r_{iTR}$ и специализации переводчика относительно домена приложения текста $\overline{S}tr_{iTR, ad_{iAD}}$ формируется значение значимости w_{fek} , множество нормированных значений w_{ouk} значимости формируют матрицу значимости оценок сложности \overline{W}_{fe} размерностью $1 \times k$, где k – общее число оценок, которые выступают коэффициентами уравнения поиска теоретического значения качества перевода.

Шаг 3. Для каждого i -го фрагмента текста при $i = \overline{1, N}$ формируется матрица оценок фрагмента исходного текста C_{fei} размерностью $1 \times k$, где k – общее число оценок.

Шаг 4. На основании оценок C_{fei} и значимости \overline{W}_{fe} формируется уравнение поиска теоретического результирующего фактора, т.е. качества перевода $\overline{T}Q$:

$$\overline{T}Q_i = w_0 + w_{1fe} C_{fei_1} + w_{2fe} C_{fei_2} + \dots + w_{fek} C_{feik} \quad (1)$$

Шаг 5. Для каждого i -го фрагмента текста рассчитывается вероятность получения переведенного текста на таком уровне качества, который определяется требованиями $TP|txt_{iTXT}$, применив к уравнению (1) логит-преобразование [11]:

$$p_i = \frac{1}{1 + e^{-\overline{T}Q_i}} \quad (2)$$

Шаг 6. Сложность задачи перевода i -го фрагмента текста оценивается по формуле:

$$TTC_i = \frac{1}{p_i}, \quad (3)$$

где p_i – это вероятность создания переводчиком перевода требуемого качества, рассчитанная по формуле (2).

Шаг 7. Результирующая сложность задачи перевода текста – наибольшее значение сложности задачи перевода TTC_i среди N фрагментов исходного текста, то есть

$$TTC_{txt_{iTXT}} = \max TTC_i \quad (4)$$

Алгоритм реализации описанной математической модели в рамках прикладной задачи перевода:

1. Тестирование переводчика/системы МП на тренировочном корпусе текстов заданной тематики, для которых имеется эталонный перевод.
2. Вещественная оценка параметров текста в тренировочном корпусе.
3. Оценка качества выполненного переводчиком тестового перевода.
4. Поиск весов значимости параметров текста заданной тематики для тестируемого переводчика.
5. Получение уравнения поиска теоретического качества перевода, выполненного тестируемым переводчиком. На основе уравнения производятся расчеты ожидаемого качества, вероятность получения перевода, соответствующего классу «качественный перевод» и сложность задачи перевода.

Данный алгоритм протестирован на реальных данных в рамках перевода текстов нефтегазовой тематики системой МП.

4 Реализация модели оценки сложности задачи перевода

4.1 Тестирование переводчика

Для тестирования разработанного алгоритма был разработан программный модуль-парсер с использованием модели нейронного МП на базе Transformers и Open Source модели Helsinki-NLP¹ для языковой пары русский-английский, предварительно обученной на корпусе OPUS². Разработка и инициализация парсера МП производилась в среде Google Colab. Использование парсера обусловлено необходимостью протестировать алгоритм на большом массиве данных, не прибегая к использованию коммерческих программ ввиду ограничений на конфиденциальность используемых исходных данных.

Тренировочный корпус – корпус двуязычных текстов, собранный на базе Translation Memories из сред автоматизации перевода, применяемых компанией-поставщиком лингвистических услуг, специализирующейся на переводе технических текстов в сфере нефтегазопереработки. Структура данных: [исходный текст, ручной перевод, проверенный редактором]. Корпус предварительно очищен от шумов, таких как сегменты, содержащие нетекстовую информацию, теги форматирования текста, строки длиной менее 50 символов. Объем корпуса составил ~1 744 400 токенов или ~7 300 стандартных страниц³ русскоязычного текста.

4.2 Вещественная оценка параметров исходного текста

Для вещественной оценки параметров текста был разработан программный модуль для оценки параметров русскоязычного текста по четырем группам признаков: морфологические, синтаксические, лексические и прочие признаки. Программа разбивает текст на токены и рассчитывает значение для 96 вещественных признаков.

Для оценки морфологических и синтаксических признаков текста используется разбор по схеме универсальных зависимостей [12]. Морфологическая спецификация слова в схеме универсальных зависимостей состоит из трех частей: лемма слова, тег части речи и морфологические признаки, которые определяют лексические и грамматические свойства формы слова. Оценивается доля слов исследуемой части текста по каждому тегу части речи, как наиболее информативный морфологический признак. Схема универсальных зависимостей позволяет производить синтаксический разбор текста, маркируя каждую его единицу (токен) соответствующим тегом отношения «rel=».

Морфологический и синтаксический разбор включает следующие шаги:

1. Разбиение текста на токены.
2. Определение для каждого токена значений свойств pos (часть речи) и rel (роль в предложении).
3. Подсчет количество частей речи в строке и число токенов по каждому тегу.
4. Расчет для всех тегов отношения количества токенов с соответствующим тегом к общему числу токенов в предложении.

Морфологический и синтаксический анализ выполняется при помощи библиотеки natasha⁴ для моделирования НЛП на основе глубокого обучения для русского языка, который имеет сравнимую точность с большими моделями BERT SOTA, но занимают в 50 раз меньше места.

Для оценки лексических признаков использовался Национальный частотный словарь русской лексики [13], который был предварительно обработан и отсортирован по значению частотности слов. Оценка производилась для токенов, включающих слова, которые предварительно лемматизированы при помощи библиотеки natasha.

Оценка факторов на нормальность распределения показала наличие большого количества факторов с экспоненциальным распределением значений и факторов, большая часть наблюдений по которым равна 0. Будем учитывать это при оценке точности модели, так как нормализация и

¹ <https://huggingface.co/Helsinki-NLP>

² <https://opus.nlpl.eu/>

³ Стандартная страница равна 1800 знаков с пробелами

⁴ <https://github.com/natasha>

шкалирование данных не позволит устранить нулевые значения, а балансировка данных по нулевым значениям может привести к значительным потерям данных по другим факторам.

4.3 Критерий качества машинного перевода

В рамках представленного алгоритма могут применяться любые метрики оценки качества в зависимости от требований к качеству перевода. Была выбрана метрика hLEPOR, которая является комбинацией существующих и доработанных факторов и показывает лучшие результаты оценки по сравнению с MPF, ROSE, METEOR, BLEU и TER, а также имеет наивысший балл корреляции Пирсона с человеческими суждениями по языковой паре английский-русский [14]. Оценка метрики производилась путем сравнения сгенерированного МП (гипотезы) и эталонным переводом, выполненным человеком, при помощи библиотеки hLEPOR⁵. Диапазон изменения метрики от 0 до 1, где 0 – полное несовпадение гипотезы с эталоном, а 1 – полное совпадение.

4.4 Корреляционная связь качества перевода с параметрами исходного текста

Для поиска весов значимости параметров текста и получения многофакторного уравнения поиска теоретического качества перевода использовалась модель логистической регрессии, реализованная в библиотеках statsmodels.api и sklearn для языка Python. Результаты моделирования представлены на рисунке 2.

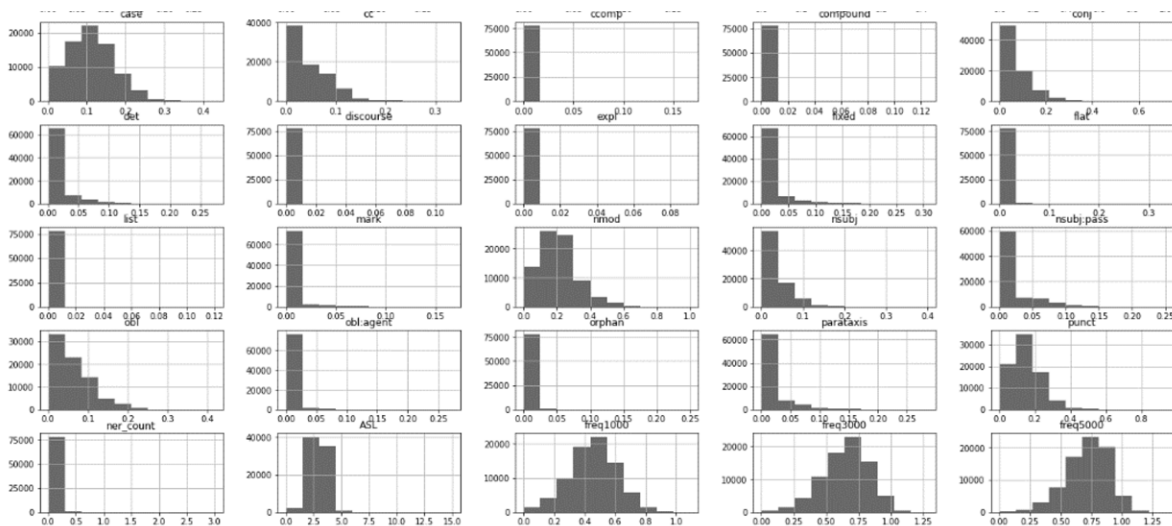


Рисунок 2: Гистограммы распределения данных для части факторов

⁵ <https://pypi.org/project/hLepor>

	coef	std err	z	P> z	[0.025	0.975]
const	-2.1237	0.119	-17.917	0.000	-2.356	-1.891
ADJ	0.8890	0.145	6.113	0.000	0.604	1.174
ADP	-1.1872	0.191	-6.222	0.000	-1.561	-0.813
ADV	-1.1501	0.323	-3.562	0.000	-1.783	-0.517
DET	1.1331	0.371	3.055	0.002	0.406	1.860
NUM	4.3565	0.227	19.228	0.000	3.912	4.801
PRON	-3.2955	0.419	-7.869	0.000	-4.116	-2.475
PROPN	-0.9230	0.136	-6.799	0.000	-1.189	-0.657
VERB	-0.9131	0.234	-3.898	0.000	-1.372	-0.454
X	2.3469	0.409	5.733	0.000	1.545	3.149
acl	-1.3931	0.356	-3.916	0.000	-2.090	-0.696
aux	5.1366	1.905	2.696	0.007	1.403	8.870
conj	2.2937	0.178	12.850	0.000	1.944	2.644
csbj	-4.6383	0.760	-6.099	0.000	-6.129	-3.148
flat	4.3092	1.322	3.261	0.001	1.719	6.899
flat:name	-4.4942	1.231	-3.650	0.000	-6.908	-2.081
list	13.0258	4.971	2.620	0.009	3.283	22.768
nmod	-1.6506	0.131	-12.597	0.000	-1.907	-1.394
nsubj:pass	-0.7066	0.310	-2.276	0.023	-1.315	-0.098
obj	-1.3491	0.261	-5.175	0.000	-1.860	-0.838
obl	-0.7080	0.224	-3.161	0.002	-1.147	-0.269
obl:agent	3.2594	0.843	3.865	0.000	1.606	4.912
parataxis	3.7789	0.360	10.506	0.000	3.074	4.484
punct	0.3628	0.168	2.166	0.030	0.034	0.691
xcomp	1.3277	0.380	3.490	0.000	0.582	2.073
freq10000	1.6783	0.064	26.415	0.000	1.554	1.803
tokens_count_log	0.3153	0.018	17.294	0.000	0.280	0.351

Рисунок 3: Коэффициенты логистической регрессии

В итоговую модель вошли только те факторы (параметры исходного текста), для которых *P*-значения показывают высокую значимость. Таким образом, были получены коэффициенты уравнения поиска теоретического качества перевода, выполненного тестируемым переводчиком. Всего выявлено 26 параметров, из которых 12 имеют отрицательную зависимость с потенциальной оценкой качества перевода и являются мешающими. Выполняя предредактирование исходного текста с целью оптимизации данных параметров возможно повысить качество МП. Кроме того, 14 выявленных параметров имеют положительную зависимость с потенциальной оценкой качества перевода, что тоже может использоваться в оптимизационном предредактировании для снижения влияния мешающих параметров. Проверка качества полученной модели проводилась на основе показателей ROC-AUC [15] и коэффициента корреляции Пирсона, рассчитанного для *TTC* и фактической *QS* для 85125 строк текста, согласно полученной зависимости (таблица 1).

ROC-AUC показывает, что на основании признаков исходного текста без оценки его семантики, возможно предсказывать ожидаемое качество МП и имеется потенциал доработки модели и повышения ее точности с учетом разреженности пространства факторов и распределения их значений. Коэффициент $r_{TTC, QS}$ указывает на обратную корреляционную связь с выбранной оценкой качества перевода и сложности задачи перевода.

Принимая во внимание допущение, что автоматическая оценка качества перевода имеет некоторую погрешность, можно говорить о состоятельности разработанной модели и предложенного алгоритма оценки сложности переводческой задачи перевода текстов нефтегазовой тематики для выбранной системы МП.

Показатель	Значение
ROC-AUC	0,6256
$r_{TTC, QS}$	-0.26

Таблица 1: Результаты моделирования

5 Стратегия предредактирования исходных текстов

Предредактирование – процесс модификации проблемных фрагментов исходного текста за счет изменения его структуры, состава лексических единиц и т.п. при сохранении семантической эквивалентности с целью оптимизации сложности переводческой задачи с учетом влияния такой модификации на качество перевода.

Предредактирование выполняется при средней и высокой сложности переводческой задачи с использованием разнообразных алгоритмов, например, по критериям: минимизации $СлЗП_{txtTXT}$, минимаксимизации отдельных свойств и параметров исходного текста и др.

Диапазоны значений TTC_{txtTXT} , соответствующие низкому, среднему и высокому уровням определяются на основе предварительного анализа требований к переводу и с учетом способа перевода (ручной/машинный). Цель предредактирования – снизить сложность переводческой задачи TTC_{txtTXT} до низкой.

На начальном этапе предредактирования происходит отбор фрагментов текста, которым соответствуют высокие значения TTC_i , то есть

$$txt_i^* \in txtiTXT \mid TTC_i > TTC_{alw} \tag{5}$$

где TTC_{alw} – допустимое значение сложности задачи перевода, т.е. низкое.

Для тех фрагментов текста, для которых сложность задачи перевода является средней или высокой, определяем элементы вектора C_{fei} , значения которых повышают сложность переводческой задачи данного фрагмента: $C_{fei}/TTC_i > TTC_{alw}$.

Для каждого найденного элемента матрицы C_{fei} , соответствующего условию $C_{fei}/TTC_i > TTC_{alw}$, в зависимости от его значения и ограничений алгоритмов предредактирования определяется стратегия предредактирования (таблица 2). Выбор алгоритма зависит от критерия оптимизации выбранного значения.

Значение параметра	Действие
Находится в допустимых пределах	Автоматическое предредактирование при помощи алгоритмов
Выходит за допустимые пределы	Полуавтоматическое предредактирование текста с привлечением пользователя

Таблица 2: Стратегии предредактирования

Далее производится редактирование текста в соответствии с имеющимися методами и алгоритмами предредактирования, формируется текст txt'_{TXT} и происходит переход к этапу перевода.

Предредактирование позволяет повысить качество МП. Пример предредактирования и его влияния на значения параметров текста и качество перевода на английский язык представлены в таблице 3. В таблице представлены наиболее значимые параметры для соответствующего текста.

Текст 1	hLEPOR	ADP	conj	punct
а) В результате многодневной переписки между представителями арендатора, арендодателя, сервисной компании и завода производителя результат о ремонте или замене станции достигнут не был.	0,5375	0,1200	0,1600	0,1200
б) В результате переписки между представителями Арендатора, Арендодателя, Сервисной компании и Производителя, решение о ремонте или замене станции не было достигнуто.	0,7560	0,1250	0,1667	0,1667

Текст 2	hLEPOR	ADP	VERB	xcomp
а) Оборудование должно быть рассчитано на двойные фидеры, а если такое оборудование отсутствует, в центральном шкафу предусматривают установку контроллера автоматического ввода резерва.	0,5051	0,0833	0,1250	0,0417
б) Оборудование должно быть способно управлять двумя фидерами, в случае отсутствия такого оборудования в центральном шкафу должен быть установлен переключатель ввода резерва.	0,6512	0,0909	0,0909	0,1364

Таблица 3: Пример предредактирования русскоязычного текста – а) исходный текст; б) текст после предредактирования

6 Заключение

В рамках исследования впервые разработана математическая модель и алгоритм автоматизированной вещественной оценки сложности задачи перевода и подтверждена обратная корреляционная связь такой оценки с оценкой качества, рассчитанной по методу hLEPOR для узкоспециальных текстов нефтегазовой тематики. Выявлены параметры русскоязычного узкоспециального текста, имеющие корреляционную связь с потенциальным качеством МП на английский язык по целевому показателю оценки метрики hLEPOR.

Подход может быть масштабирован на ручной перевод и внедрен в компаниях, генерирующих от 1000 страниц перевода в месяц, так как намечает подходы к управлению рисками, связанными с качеством перевода в зависимости от компетенции выбранных исполнителей, и предоставит индустрии инструмент объективной оценки исполнителей в рамках поставленной задачи на перевод.

На основе методики оценки сложности задачи перевода становится возможным автоматически определять стратегию предварительного оптимизационного редактирования текста с целью приведения значений его параметров к оптимальным, при которых вероятностная оценка качества МП стремится к максимальной. Дальнейшее исследование целесообразно направить на разработку методики расчета целевых диапазонов значений *ТТС* и оптимальных значений параметров исходного текста по критерию максимизации качества перевода, а также разработку комплекса алгоритмов оптимизационного предредактирования.

References

- [1] Yamada M. The impact of Google neural machine translation on post-editing by student translators // *The Journal of Specialised Translation*. — 2019. — vol. 31. — P. 87–106.
- [2] Toledo Báez M. Machine translation and post-editing: impact of training and directionality on quality and productivity // *Revista Tradumàtica. Technologies de la Traducció*. — 2018. — vol. 16. — P. 24–34.
- [3] Hiraoka Y., Yamada M. Pre-editing plus neural machine translation for subtitling: effective pre-editing rules for subtitling of TED Talks // *MT Summit XVII*. — Dublin, Ireland. — 2019. — vol.2. — P. 64–74.
- [4] Miyata R., Fujita, A. Dissecting human pre-editing toward better use of off-the-shelf machine translation Systems // *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT), User studies papers*. — Prague, Czech Republic. — 2017.
- [5] Gerlach J., O'Brien S. et al. Combining pre-editing and post-editing to improve SMT of user-generated content // *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*. — Nice, France. — 2013. — P. 45-53.
- [6] Taufik A. Pre-editing of Google neural machine translation // *Journal of English Language and Culture*. — 2020. — vol. 10. — No. 2. — P. 64–74.
- [7] Mercader-Alarcón J., Sánchez-Martínez F. Analysis of translation errors and evaluation of pre-editing rules for the translation of English news texts into Spanish with Lucy LT // *Revista Tradumàtica. Technologies de la Traducció*. — 2016. — vol. 14. — P. 172–186.
- [8] Seretan V., Bouillon P. et al. A large-scale evaluation of pre-editing strategies for improving user-generated content translation // *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. — Reykjavik, Iceland. — 2014. — P. 1793–1799.
- [9] Shei Chi-Chiang. Teaching MT through pre-editing: three case studies // *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*. — Manchester, England. — 2002.
- [10] Дмитриева, А.Д., Лапошина А.Н., Лебедева, М.Ю. Квантитативное исследование стратегий упрощения на материале адаптированных текстов для изучающих РКИ // *Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог»*. — 2021. — С. 191-204.
- [11] Hosmer, D. W., Lemeshow, S. *Applied Logistic Regression*. 2nd edn. — New York: Wiley Chichester, 2000.
- [12] Люкина Е. В. Использование универсальных зависимостей при грамматическом разборе многоязычного текста (на примере безличного предикатива) // *Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация*. — 2018. — Т. 16. — № 2. — С. 19-33.
- [13] Ляшевская О. Н., Шаров С. А. *Частотный словарь современного русского языка*. — Москва : Азбуковник, 2009.
- [14] Li-Feng Han A., Wong D. F. et al. Language-independent Model for Machine Translation Evaluation with Reinforced // *Proceedings of the Machine Translation Summit XIV*. — Nice, France. — 2013. — P. 215-222.
- [15] Бенфорт Б., Билбро Р., Охеда Т. *Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка*. — Санкт-Петербург: Питер, 2019.

Remus, Lupin and Moony Walk in a Bar... Grouping of Proper Names Related to the Same Denotation in Large Literary Texts Collections

Zykova V. I.

National Research University
“Higher School of Economics” /
Pokrovsky Bulvar, 11,
Moscow 109028 Russia
vzykova2001@gmail.com

Klyshinsky E. S.

Keldysh Institute of Applied
Mathematics of Russian Academy of
Sciences / Miusskaya sq., 4
Moscow, 125047, Russia
klyshinsky@mail.ru

Abstract

In this article, we present a method of anaphoric proper names detection in fictional texts using Word2Vec model and algorithms of community detection on graphs. This method allows grouping different namings of a single entity and can be useful as a part of preprocessing texts for further analysis such as building social networks or training neural models. The method uses large text collection, related to the same domain. The foundation of the method is training of a Word2Vec model using information on direct characters interactions. This model allows building a social graph of characters. Then, the Louvain algorithm is used to divide the graph into communities containing different names of characters related to the same denotation.

Keywords: social network; graphs; word2vec; anaphora detection; fiction; Louvain algorithm

DOI: 10.28995/2075-7182-2023-22-1150-1157

Заходят как-то в бар Ремус, Люпин и Лунатик... Метод объединения имен собственных, имеющих общий денотат, в больших коллекциях художественных текстов

Зыкова В. И.

Национальный исследовательский
университет «Высшая школа
экономики» / 109028, Москва,
Покровский бул., д. 11
vzykova2001@gmail.com

Клышинский Э. С.

Институт прикладной математики
им. М. В. Келдыша /
125047, Москва,
Миусская пл., д. 4
klyshinsky@mail.ru

Аннотация

В этой статье мы представляем метод определения именованных героев художественных произведений, относящихся к одному денотату, использующий модель Word2Vec и алгоритмы выделения сообществ на графах. Метод позволяет объединять различные названия одной сущности в группу с достаточно высокой точностью и может быть полезен при использовании в качестве этапа препроцессинга текстов для дальнейшего анализа: построения графов социальных отношений или обучения нейросетевых моделей. Метод применяется к большой коллекции текстов, относящихся к одному домену. Основой метода является обучение модели Word2Vec с использованием информации о прямом взаимодействии героев. Модель служит для построения графа связей между героями, из которого при помощи Лувенского алгоритма выделяются сообщества, содержащие разные названия одного героя.

Далее проводится фильтрация разных героев, объединяемых моделью.

Ключевые слова: графы социальных отношений; графы; word2vec; определение анафоры; художественные тексты; Лувенский алгоритм

1 Introduction

Investigation of interaction among characters of literary works allows better understanding the overall plot of a masterpiece. One of the commonly used representations of such interactions is a graph. Its nodes represent literary characters, described in an investigated work, edges represent such interactions among characters as conversations, direct actions, and being in the same place. Investigation of large graphs allows describing two types of those: Barabasi-Albert [1] and ‘Small World’ [2]. The authors of [3] found that graph of character interactions has properties of the further and does not meet all the properties of the later.

There are several approaches for coupling nodes depending on the used information. In [4], an edge in the graph is created between two characters “if they were both syntactic arguments under the same predicate or appeared as two conjuncts”. For the sake of community detection, the author used the Louvain algorithm [5]. Another approach is using mutual appearance in the same piece of text [6].

Investigation of drama corpora (e.g. [3]) has a big advantage. Such a corpora was preliminary tagged including information on phrases attribution by characters. Thus, it is easier to use the common appearance of characters in the same part of text or information of whom their phrases are addressed in order to build the social graph of a masterpiece.

However, researchers are faced with the problem of several names for a character. In different parts of a text, an author can use the name of a character, his or her surname, their combination, position, gender names (boy, girl, ...), and other substitutive nouns. Thus, there is a problem of joining all of these names in the same cluster. Manually tagged corpora have an advantage that all these names can be unified by an assessor [6]. But that is not the case of automatically tagged large corpora. One of the solutions in the later case is preprocessing, i.e. selection of different names of the same character and automatic replacement of a name occurrence by the unified one. But it does not work in the case of a big number of unknown characters. Methods of anaphora resolution (see, e.g. [7]) could not help find different names for main characters.

In this paper, we introduce a new method for calculation of similarity between characters’ names. The method is able to cluster names of the same character in a community of a social graph. The method is based on training a Word2Vec model on entity names co-occurrence, building a social graph according to similarity between those entities calculated by the model, and applying the Louvain algorithm for the sake of community detection.

2 Used Data

2.1 Text Collection

As a source data for our method, we need a collection of mono-domain texts, e.g. works created by one author or dedicated to the same fictional setting. We use a collection of fanfiction texts based on J.K. Rowling “Harry Potter” books. There are several reasons for such a decision.

- Authors of fanfiction texts use original setting or at least original characters as the basis of their work; so, all the texts sharing the same original subject and the same setting would share almost the same characters space and could be treated as a single domain.
- “Harry Potter” as a fiction is extremely popular, this fact provides us with a large amount of data. This, for its turn, gives us an opportunity to balance the imperfect performance of language models.
- Most of the fanfiction sites are open resources which can be easily and automatically crawled.

Considering all these facts, we chose the <https://fanfics.me/> site. All the fanfiction texts belonging to the “Harry Potter” setting were crawled from this site in summer 2022 and then used in the algorithm described below.

The amount of data collected is 22252 texts which have around 336 million of tokens in total. This is a large collection of data so we have to take it into consideration while making a decision about preprocessing algorithms which will be discussed below.

2.2 Text Preprocessing

Our algorithm requires syntactically parsed natural texts. There are a variety of open source tools performing this task. Such tools can be multilingual and support Russian alongside many other languages or keep it as the only or almost only one available. For example, Spacy [8] and Stanza [9] provide models for more than twenty languages each while Natasha and DeepPavlov [10] are concentrated mostly on models for Russian language. There is also a UD-Pipe module but it is almost deprecated.

Besides the list of supported languages, mentioned models have different quality and capacity features: some of them are relatively precise but the level of their efficiency is quite low, some, in contrast, are surprisingly fast but their quality is below the acceptable level. Considering preprocessing as an important step which can dramatically affect all future calculations, we tend to prioritize quality over productivity. However, as it was already mentioned, we are working with a large amount of textual data processing which can require an inappropriate amount of time, thus we have to search for compromises. Because of all mentioned above, we decided to use the Spacy framework as it is relatively fast even without using a GPU and shows an appropriate, but not perfect, performance.

The next step of preprocessing is Named Entity Recognition (NER) which is an ongoing and difficult task itself; however, we need it to shape the initial list of characters. Note that on some level of precision, this list can be gained on the stage of syntactic parsing since most parsers have the PROPEN tag in their tagset. Considering this, using NER is not so much about detecting entities in general but about detecting multi word namings as it is usually beyond the capabilities of syntactic parsers.

As a result of data preprocessing we should have a corpora of syntactically parsed texts with results of the NER algorithm. We used the one proposed by Spacy as it can be easily added to its pipeline.

3 Grouping of Proper Names Related to the Same Denotation

As can be seen from the task formulation, our algorithm can be divided into two parts. The first one is collecting of the characters co-occurrence information from the texts; so as it can be utilized as training data for the further model. The second part is the model training.

For the sake of model training, we need a list of pairs of characters which satisfies the condition of being “interacting”. Thus, the first stage can be divided in several subtasks too. In the beginning, we need to define what “character” and “interaction” are in terms of grammatical traits obtained from the preprocessing step; then the algorithm of applying these conditions to the process of list of pairs extraction should be described.

We define an entity (not necessarily named) or a character (if we are speaking about it in terms of fiction terminology) as a token which meets one of the following conditions: it is tagged as an entity by the NER model; it is tagged as a proper noun by the parser; it is tagged as an animate noun.

There are several reasons for such a decision. To begin with, we want proper names to be considered as an entity. However, neither syntactic parsers nor NER models have good precision, while we are speaking about them in terms of entity extraction from fictional texts. However, there is a chance that mistakes of the syntactic parser can be smoothed by NER model predictions and vice versa, so we decided to unite their decisions. Moreover, additional difficulties are provided by the fact that we need not only named entities but also their substitutive nouns to be detected. Here we hope that inanimate objects could not be considered as fictional characters (that is not the case of fairy tales, which are not the subject of this article). So, we select all animate nouns as hypothetical entities entailed with some of the characters.

After we proposed a definition of an entity, we now can discuss conditions of being interacting. Following [4] line of thought, we decided to use being descendants of one vertex in the dependency tree as a definition of connection existence. Unlike the approach described in the mentioned work where the author uses all predicates as the parenting vertices, we decided to make constraints on such vertices stricter: we consider only verbs being a predicate. Summarizing this with a part about entity definitions we can postulate the following definitions of a pair: it consists of the two entities which are the descendants of the same verb in a dependency tree.

We have found that a pair of characters can be considered as a context to each other. This allows us to train a language model treating pairs and entities as sentences and tokens respectively. In this project, we decided to train a Word2Vec model [11] in order to get vector representation (embeddings) for each

character. There are several reasons for choosing this architecture. On the one hand, Word2vec is a static language model that provides us with a single embedding for each entity from vocabulary. Since we want to get a method providing us with an absolute measure of semantic similarity between entities, the ability to calculate an average representation of a token becomes useful. On the other hand, considering usage of a contextualized model, we understood that the described way of representing context does not provide enough information for the stable learning of a contextual model.

Now we can now describe the process of obtaining the trained model. At the beginning all the entities should be extracted keeping their affiliation with sentences since we will need this information to make a decision about connection existence. Then for every possible pair of entities within a sentence it should be checked if these entities are the descendants of the same verb in the dependency tree. As a result of the described process, we have a list of pairs of characters ordered by their appearance in our collection of texts. Then it is used to train a word2vec model so that characters in every pair have only each other as a context (it is guaranteed by the window size parameter of a word2vec training algorithm).

Using the trained model, we then build a social graph of characters. At the first stage we select all named characters (entities which are marked by NER algorithm or are a proper noun) with frequency more than the reference minimal frequency (let be equal to 50) as initial nodes of the graph. Then for every node we extract top-100 nearest neighbors having cosine similarity more than 0.5 and frequency more than 20. Then we create an edge between two nodes if at least one of them is in the top-100 list for another even if they are both initial nodes. The result is a weighted graph where weights are designated by the cosine similarity between nodes.

As it noted in [13] “The weight of a link between two nodes in a social network can be used to represent the similarity of two characters. The larger the weight is, the more likely the two characters will be related to each other.” However, it appears that cosine similarity is not enough. There are named entities, which have a relatively high proximity but have to be divided, such as twin brothers Fred and George Weasley, who are always acting together, or Harry Potter and Hermione Granger, who share their adventures. However, there are entities which have smaller cosine similarity but have to be considered as a single entity. That is the case of different namings of one character whose frequencies differ dramatically.

In [4], characters which belong to the same predicate are considered as two different persons. In our case, if two extremely similar characters frequently appear as syntactical siblings in a parsed text, then they should be considered as two different persons. However, our experiments demonstrated that such an approach fails in case of less frequent characters as they often provide lesser statistics than we need for their separation. Note, that there are a variety of characters having both less and more frequent names. That is why we need here a community detection algorithm which will separate entities considering the whole picture of their relations.

Following [4], we decided to use the Louvain algorithm [5]. As a result of applying this algorithm to a social graph we obtain groups of entities which have closer relations. By changing the resolution parameter of a Louvain algorithm, we can control the density of achieved communities as its high value makes the algorithm favor smaller communities. In our case it means that we can decide if communities should contain names of different but narratively close characters or, if using higher values of resolution parameter, different name of the same character. Although all of the mentioned parameters (threshold of cosine similarity, minimal frequencies and resolution parameter) make the precision higher, we cannot just make them as high as possible since in the extreme case we would get an empty graph or a list of one-node communities which makes all the process meaningless.

4 Results of Experiments

We trained a model on 5 epochs and words of the minimal frequency 10. We were surprised to see that a model that was accidentally trained with pairs duplicates (pairs like “Harry Ginny” and “Ginny Harry” were both added to the list) showed a more stable performance and more representative results; so, we decided to use it as a basic approach in the further work. We used all entities lowercased and lemmatized; we also joined multiword entities extracted with NER model into single strings.

However, the first attempt of training the model showed that there are a lot of mistakes of lemmatization and pos-tagging. We decided to begin with solving the first problem. In order to map different forms of the one entity, we built a graph where an edge between two entities means that they have

Levenshtein distance [12] equal to 1 (see Fig. 1). After we got such a graph, we considered one connected component to be different names of a character; however, it happened to need a little manual correction. After that we mapped all the entities in a connected component to a single character's name which was the most frequent one in the original component. Then we retrained a model.

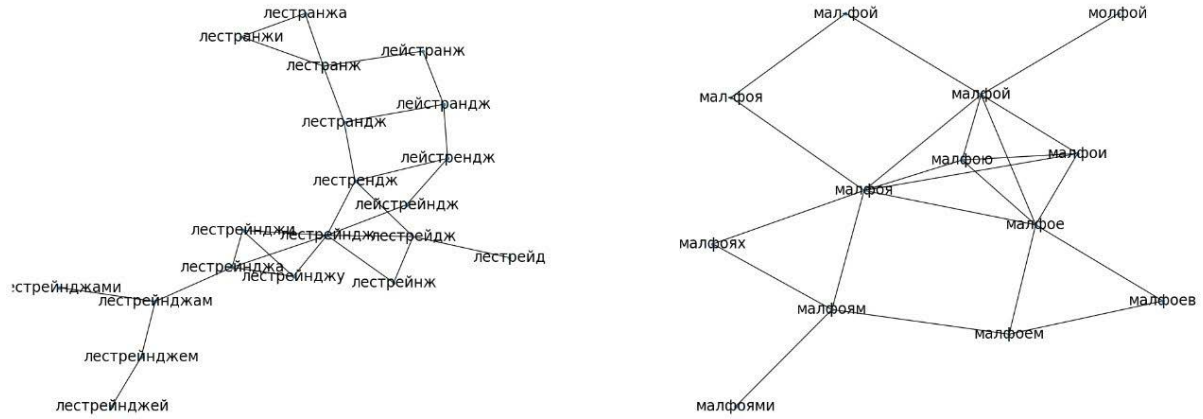


Figure 1: Graph of tokens having the Levenstein distance equal to 1

After retraining, the solution for the problem of incorrect pos-tagging (that actually means that we had adverbs or adjectives marked as entities) happened to be found. It appears that almost all incorrectly tagged words formed a single connected component (see Fig. 2), so we merely removed from the training data all the pairs including these tokens. Then we retrained the model again.

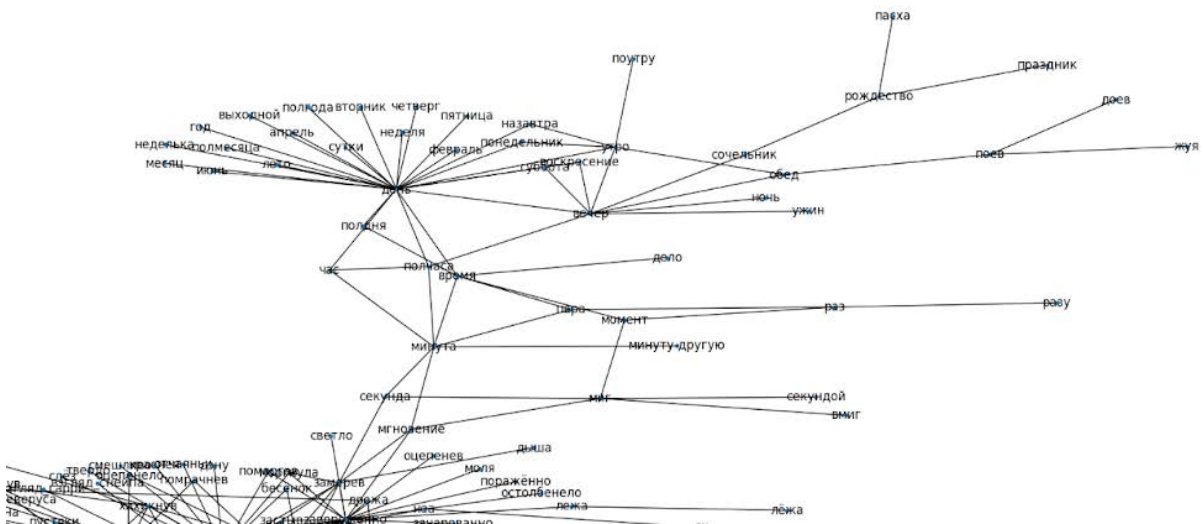


Figure 2: Graph of entities connections for the component with pos-tagging mistakes

Having a final model, we repeated all the process of graph building. The resulting graph had an interesting structure: there was a big connected component which contained almost all main characters and needed to be divided into parts (see Fig. 3) and a lot of small ones. As can be seen on the picture, the main connected component itself has several easily distinguished parts such as the one in the upper-left part which is devoted to Ministry of Magic characters or the bottom part whose components are all about geographical names both real and magical. Sadly, we still do have a small amount of mistagged words but there are already much less of them than there were in the previous iterations.

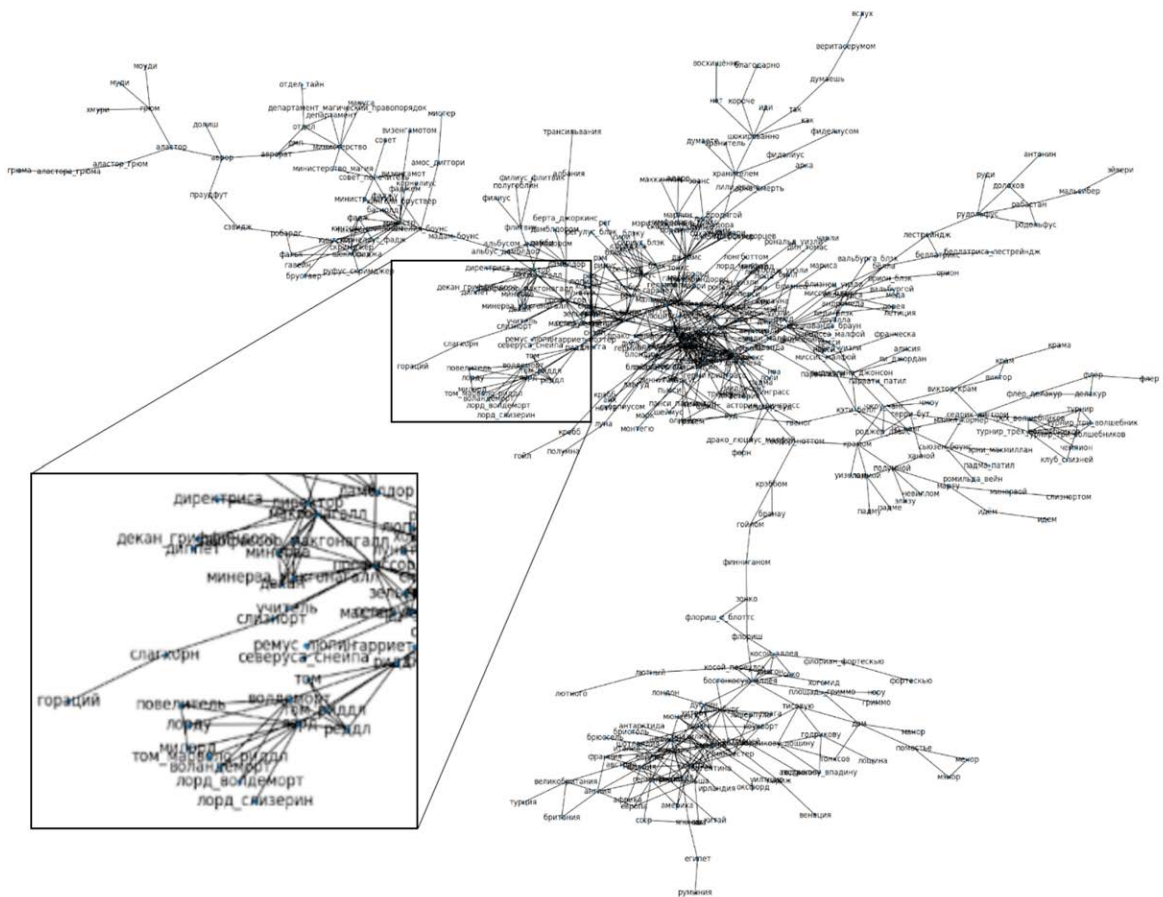


Figure 3: Graph of entities connections for the biggest connected component

As for the resolution parameter, we can start with 10 as the lower values of the algorithm does not really distinguish anything (it should be mentioned that this parameter has to be selected independently for every graph). However, this value appears to be too low and provides us with large groups of connected yet not the same characters (see Fig. 4).

- министр магии, робардс, фаджа, сэвидж, фаджу, визенгамотом, шеклболт, фадж, корнелиус, скримджер, министр, кингсли, бруствер, визенгамот
- америка, уэльс, испания, канада, шотландия, франция, германия, швейцария, греция, италия, ирландия, китаи, бразилия
- петрификус тоталус, инкарцери, секо, экспеллиармус, петрификусом, протего, ступефаем, петрификус, ступефай, экспеллиармусом, непростительный
- косой аллея, зонко, лютный, коукворт, лондон, манчестер, флориш, косой переулоч, косую аллея, хогсמיד, эдинбург
- учитель, директор, профессор, флитвик, филиус, минерва, мастер зелий, минерва макгонагалл, диплет, декан, макгонагалл
- риддл, том риддл, лорду, лорд, реддл, том, волдеморт, лорд Волдеморт, повелитель, воландеморт, милорд
- джим, лили эванс, эванс, лили, сохатый, шмэри, марлин, адаре, джеймс поттер, джеймс, скорп дин томас, уизли, рон, рон уизли, рональд, рыжик, уизел, дин, рональд уизли, перси
- блэка, тонкс, блек, бродяга, нимфадора, сириус, блэку, блэк, сириус блэк, сири
- чарли, джордж, фред уизли, анджелина, фред, кэти белл, джордж уизли, билл, кэти, близнец

Figure 4: Biggest 10 communities detected by the Louvain algorithm with resolution parameter equal to 10

If we increase the value to 20, we will still not get enough precise groups (see Fig. 5), but they will become more clean.

министр_магии, фаджа, визенгамотом, фадж, корнелиус, скримджер, министр, фаджу, визенгамот
 андромеда, друэлла, дорея, нарцисса, цисси, нарцисса_малфой, вальбурга, меда, нарси
 совы, филин, букля, буклю, сова, сову, хедвиг, сов
 блэка, блек, бродяга, сириус, блэку, блэк, сириус блэк, сири
 джордж, фред уизли, анджелина, фред, кэти белл, джорджуизли, кэти, близнец
 когтеван, гриффиндором, хаффлпафф, слизерин, пуффендуя, слизерином, рейвенкло, гриффиндор
 долохов, мальсибер, руди, рудольфус, антонин, рабастан, родольфус, эйвери
 кэрри, геланор, гарольд, гарриет, марьяна, эмили, гарри
 дурсли, мардж, петунья, эйлин, туни, вернон, дурслей
 косой аллея, зонко, косой переулок, косую аллея, лютный, хогсмид, флориш

Figure 5: Biggest 10 communities detected by the Louvain algorithm with resolution parameter equal to 20

Finally, the resolution parameter equal to 30 provides us with the best compromise between precision of the groups and the proportion of connections detected (see Fig. 6). As it can be seen on the picture, there are still groups which need further decomposition (such as the group with Harry which contains him and some marginal characters that are even not the part of Harry Potter original characters), but the fact that we can connect naming that are not connected in any way but semantically (such as Remus Lupin and Moony) make us consider this algorithm to have a pretty high precision and be at least a method of getting the baseline for more precise yet computationally complex models.

совы, хедвиг, филин, букля, буклю, сова, сову, сов
 блэка, блек, бродяга, сириус, блэку, блэк, сириус блэк, сири
 кэрри, геланор, гарольд, гарриет, марьяна, эмили, гарри
 дурсли, мардж, петунья, эйлин, туни, вернон, дурслей
 хогсмид, косой аллея, зонко, косую аллея, лютный, флориш, косой переулок
 лунатик, люпина, люпин, рем, ремус люпин, ремус, римус
 долиш, аврорат, министерство, отдел, отдел тайн, министерство магия, аврор
 кричер, кикимер, добби, тилли, эльф, винки, домовик
 министр_магии, фадж, фаджа, скримджер, корнелиус, фаджу, министр
 лорду, лорд, воландеморт, милорд, волдеморт, лорд волдеморт, повелитель

Figure 6: Biggest 10 communities detected by the Louvain algorithm with resolution parameter equal to 30

5 Conclusion

In this article, we present a method of anaphoric proper names detection in fictional texts using Word2Vec model and algorithms of community detection on graphs. This method allows grouping different namings of a single entity and can be useful as a part of preprocessing texts for further analysis such as building social networks or training neural models.

We applied our method to a large collection of fanfiction texts devoted to the Universe of Harry Potter (22252 of texts, 336 mln of tokens) and obtained 109 groups consisting of 356 named entities, including characters names, nicknames, faculties, toponyms, and artifacts.

Note that results need manual post-processing since there are some misplaced entities in communities. Our algorithm needs a pretty big collection of texts on the same topic since the Word2Vec model needs a big amount of contexts. Thus, it can be useful for preliminary processing of big textual collections as a builder of lists of anaphoric links.

One of the limitations of our method is it requires a large collection to train a Word2Vec model. It is hardly applied to a single masterpiece, even a huge one.

In further research we are planning to formulate the requirements for a corpus more precise since it can be useful for understanding the ways of overcoming existing limitations. Also, the algorithm of contextualized detections should be invented.

References

- [1] Barabasi A. *Network Science*. — Cambridge university press, Cambridge. 2016. 453 p.
- [2] Watts D., Strogatz H. Collective dynamics of «Small-world» networks // *Nature*, 1998. — Vol. 393. — P. 440–442.
- [3] Trilcke P. et al. Theatre Plays as ‘Small Worlds’? Network Data on the History and Typology of German Drama, 1730–1930 // *Digital Humanities 2016: Conference Abstracts* — Kraków, Poland, 2016. — P. 385–387.
- [4] Skorinkin D. (2017) Extracting Character Networks to Explore Literary Plot Dynamics // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2017” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2017”]*. Moscow, pp. 257–270.
- [5] Blondel V. D., Guillaume J., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks. — *Journal of Statistical Mechanics: Theory and Experiment*. — Vol. 10, p. 10008
- [6] Stiller J., Nettle D., Dunbar R.I.M. The small world of Shakespeare's plays. — *Human Nature*. — Vol. 14, No. 4, pp. 397–408.
- [7] Toldova S. Ju., Roytberg A., Ladygina A. A. et al. (2014) RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2017” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2017”]*. Moscow, pp. 681–694.
- [8] Honnibal M., Montani I. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. 2017
- [9] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. 2020.
- [10] Burtsev M., Seliverstov A., Airapetyan R. et al. (2018) DeepPavlov: Open-Source Library for Dialogue Systems // *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pp. 1–6
- [11] Mikolov T., et al. Efficient Estimation of Word Representations in Vector Space // *In Proceedings of Workshop at ICLR*. 2013
- [12] Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals // *Soviet Physics Doklady*. 10 (8), 1966, pp. 707–710.
- [13] Fan C., Li Y. Network Extraction and Analysis of Character Relationships in Chinese Literary Works // *Proc. of Computational Intelligence and Neuroscience*. Vol. 2, 2022, pp. 1-10.

Abstracts

SYNTAX AND PROSODY OF SPLIT SCRAMBLING: AN EXPERIMENTAL APPROACH

Belova D. D., Lomonosov Moscow State University, Moscow, Russia

The work is devoted to the experimental study of structures with split scrambling. With the help of acceptability assessment methods with the use of the Likert scale, self-paced reading, and prosodic experiment, the possibility of separating the left element (determinator or possessor) from the noun head within DP or PP is compared with the separation of the nominal head from its complement (dependent infinitive or prepositional phrase). The results show that separating the head for Russian speakers is not only possible but also rated higher than separating the left element from the head. This pattern is explained by the requirements of the information structure: the left element that has been fronted cannot form the only topic of the clause. The low scores are consistent with existing experimental studies; however, the results of the reading time appear to be inconsistent with existing views about the cognitive load required to process split sentences.

ENHANCING THE SEMANTIC AND CONCEPTUAL DESCRIPTION OF ANCIENT GREEK VERBS IN WORDNET WITH VERBNET AND FRAMENET: A TREEBANK-BASED STUDY

Biagetti E.¹, Villa L. B.¹, Zanchi C.^{1,2}, Luraghi S.¹, ¹University of Pavia, ²University of Bergamo, Italy

This paper presents a pilot study intended to enhance the semantic and conceptual description of Ancient Greek verbs in WordNet with information from two other resources, VerbNet and FrameNet, and to enrich a treebank of Ancient Greek texts with semantic information extracted from the three resources. We provided semantic annotation for verbs based on their morphosyntactic behavior, and performed a number of queries in order to extract occurrences from the Ancient Greek treebank that intended to match the different meanings of each verb. The manual check of the data extracted shows that, in spite of a limited number of mismatches, our queries yielded reliable results. The queries can be further refined in the future and complemented with a rule-based algorithm to map frame elements to dependency structure.

PAINTING THE SENATE #GREEN: A CORPUS STUDY OF TWITTER SENTIMENT TOWARDS THE ITALIAN ENVIRONMENTALIST BLITZ

Bianco A., Combei C. R., Zanchi C., University of Bergamo/Pavia, Italy

This study analyzes the reactions of the Italian Twitter community to an environmental demonstration that occurred in Rome on January 2nd, 2023. We compiled a corpus of 368,531 tokens consisting of 11,780 tweets, collected during a 7-day period. We propose a mixed-method approach that combines automated and manual corpus analyses of sentiment, emotions, and implicit language. Our findings offer insights into how tweets reflected the users' attitudes toward a variety of subjects and entities. Although the sentiment of the overall debate was distributed rather evenly, the incident itself seems to have sparked negative sentiment and emotions among Twitter users. The results of our manual analyses revealed some issues with respect to the automatic classification of sentiment, due to the fact that some tweets contained irony, sarcasm, and slurs. Non-literal interpretations were ignored by the tools at hand that could not account for complex rhetorical-argumentative strategies.

SCALED DOWN LEAN BERT-LIKE LANGUAGE MODELS FOR ANAPHORA RESOLUTION AND BEYOND

Bolshakov V.^{1,2}, Kolobov R.¹, Borisov E.^{1,3}, Mikhaylovskiy N.^{1,3}, Mukhtarova G.¹, ¹NTR Labs, Moscow/Tomsk, Russia; ²Bauman Moscow State Technical University Moscow, Russia; ³Higher IT School of Tomsk State University Tomsk, Russia

We study performance of BERT-like distributive semantic language models on anaphora resolution and related tasks with the purpose of selecting a model for on-device inference. We have found that lean (narrow and deep) language models provide the best balance of speed and quality for word-level tasks, and opensource RuLUKE-tiny and RuLUKE-slim models we have trained. Both are significantly (over 27%) faster than models with comparable accuracy. We hypothesise that the model depth may play a critical role for performance as, according to recent findings each layer behaves as a gradient descent step in autoregressive setting.

DIACHRONICON: A NEW RESOURCE FOR THE STUDY OF RUSSIAN CONSTRUCTIONS IN A MICRODIACHRONIC PERSPECTIVE

Budennaya E.^{1,2}, Bazhukov M.¹, Barkova L.¹, Kharlamova D.¹, Dugrichilov A.¹, Reznikova T.¹, Yakovleva A.¹, Litvintseva K.¹, Andreeva A.¹, ¹HSE University; ²Institute of Linguistics, Moscow, Russia

The article is devoted to the linguistic characteristics of the database "Diachronicon" and describes the features of the diachronic markup of Russian language constructions, as well as tags specially designed for searching through a diachronic database. A special comment field used in the database is separately justified. In addition, the computer interface of the "Diachronicon" is presented and described.

The developed resource provides extensive opportunities for systematic study of not only specific constructions, but also general mechanisms of idiomatization and grammaticalization. The database allows the researcher to simultaneously compare several separate plots, search through a list of constructions and their characteristics in diachrony, track the history of syntactic and semantic changes and limitations of compatibility of different constructions.

WHO IS ANSWERING TO WHOM? MODELING REPLY-TO RELATIONSHIPS IN RUSSIAN ASYNCHRONOUS CHATS

Buyanov I.¹, Yaskova D.², Sochenkov I.¹, ¹FRC CSC RAS, Moscow, Russia; ²MTS AI Moscow, Russia

The study highlights the asynchronous nature of modern group chats and related problems such as retrieving relevant information on the asked question and understanding reply-to relationships. In this work, we formalize the reply recovery task as a building block toward solving described problems. Using simple heuristics, we try to apply the result reply recovery model to a thread reconstruction problem. As a result, we show that modern pre-trained models such as BERT show great results on the task of reply recovery compared to more simple models, though it cannot be applied to thread reconstruction with just simple heuristics. In addition, experiments have shown that model performance depends on the chat domain. We open-sourced a model that can automatically predict which message the particular reply responds to and provide a representative Russian dataset that we built from Telegram chats of different domains. We also provide a test set for a thread reconstruction task.

BINARY CLASSIFICATION MODEL AS A TOOL TO DETECT SENTENCES WITH MICROSYNCTACTIC UNITS

Chaga A. V., Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

We consider a model of binary classifier predicting occurrence of microsyntactic units in sentences. The model is based on AWD-LSTM architecture with an encoder pre-trained on the Russian version of Wikipedia and further trained on a dataset built from the SynTagRus corpus supplied with a microsyntactic markup. We present the structure of the model and discuss its output. The study showed that binary classification allows targeting of microsyntactic markup and helps to significantly improve its recall.

CORPUS OF ACCENTUATED BYZANTINE WRITTEN MONUMENTS AND METHODS OF ITS MARKUP

Evdokimova A. A., Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

This article presents a corpus of Byzantine accentuated texts (BGAT) created since 2008. It currently includes 1010 Byzantine inscriptions, 950 papyri from various collections from the 1st to the 9th centuries, 132 seals from the collection of Dumbarton Oaks, and a selection of 100 Athos manuscripts from the 8th to the 15th centuries. Based on the collected data, we developed a method for markup such texts, which later makes it possible to create a database of accentuated texts from them and, based on the entire corpus, to train neural networks for classifying texts according to accentuation systems and recognizing them in images. As a result of marking texts, in addition to the previously known Alexandrian, Byzantine, and Dorian systems of accentuation, new accentuation systems were identified, including logical or semantic, with a shift of an accent mark to the right, with a displacement of an accent mark to the left, and mixed. For each group of monuments, their variants of using accentuation systems, especially the Alexandrian one, were identified, which show different aspects of the accentuation of the Byzantine Greek language. When creating a glossary of accentuated word forms based on the corpus, we determined that there were lexemes that retained their accentuation regardless of the influence of the dialect, meter, or traditions characteristic of the masters. However, a comparison of identical texts, even found in the same region of the Byzantine Empire, showed that the accentuation was not replicated when quoting.

EXPLORING EVALUATION TECHNIQUES IN CONTROLLED TEXT GENERATION: A COMPARATIVE STUDY OF SEMANTICS AND SENTIMENT IN RUGPT3LARGE-GENERATED AND HUMAN-WRITTEN MOVIE REVIEWS

Margolina A. V., Kolmogorova A. V., HSE University, Saint-Petersburg, Russia

The paper describes the proposed strategy for evaluation controlled text generation with the sentiment as attribute. Our approach mainly consists of automatic sentiment analysis (ruBERT) and topic modelling (BERTopic), which are applied to a parallel corpus with artificially produced and human-written texts. The model for evaluation is fine-tuned on the parsed reviews from big Russian movie-related website ruGPT3Large with the sentiment as prompt. The results of the analysis demonstrate that the proposed methods can offer a more comprehensive understanding of the advantages and limitations in the context of semantics and sentiment. Additionally, the paper employs metrics such as BERTscore and self-BLEU to further evaluate the generated text. The proposed methodology provides a novel approach for evaluating the quality of generated text and may have implications for future studies in the field.

THE TYPOLOGICAL CONSTRUCTION DATABASE

Muravyev N., University of Hamburg, Hamburg, Germany, **Gordeev N., Makarchuk I., Kukushkina M., Buzanov A.**, NRU HSE, Moscow, Russia

The report introduces a new resource: the Typological Construction database. This resource contains an inventory of constructions of selected semantic fields in a number of languages of different areal and genetic affiliation. The constructions are labeled according to a number of semantic and morphosyntactic parameters and provided with a detailed description and illustrative examples.

ABSTRACT USER GOALS IN OPEN-DOMAIN DIALOG SYSTEMS

Petukhova K., Smilga V., Zharikova D., Moscow Institute of Physics and Technology, Moscow, Russia

In task-oriented dialog systems, conversational agents have the means to plan the dialog to accomplish user tasks (e.g., order pizza). In chit-chat systems, there are no such straightforward tasks. Yet, in chit-chat dialogs people still pursue goals, but these goals are more abstract and thus less formalizable. In this work, we describe the development process of two goal-aware prototypes of a chatbot. The first prototype features entirely human-crafted scenarios for seven topic-specific (low-level) goals and a Goal Tracker service that detects these goals and monitors the process of their achievement. The other one combines pre-written utterances with response generation using DialoGPT model to cover the scenarios of four general (high-level) goals. The results show that introducing the concept of goals improves performance of a chit-chat dialog system. Qualitative analysis of conversations with the High-Level goals prototype demonstrates cases where a goal-aware chatbot outperforms the original one.

REPRESENTATION OF LEXICAL POLYSEMY IN THE DATABASE (SEMANTIC SHIFT 'SUN/DAY')

Russo M., Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

This paper, based on the data from more than 600 languages collected during the work on the database of semantic shifts in the languages of the world, addresses the semantic transition 'sun/'day'. We analyze the geographic and genealogical distribution of this semantic shift, the predominant direction of semantic development, and the patterns of morphological derivation associated with the shift.

GENERATIVE QUESTION ANSWERING SYSTEMS OVER KNOWLEDGE GRAPHS AND TEXT

Turganbay R.¹, **Surkov V.**¹, **Evseev D.**¹, **Drobyshevskiy M.**², ¹Neural Networks and Deep Learning Lab, MIPT, Dolgoprudny, Russia; ²Ivannikov Institute for System Programming of the Russian Academy of Sciences

In this paper we describe a generative question answering system which relies on text or knowledge graphs to find supporting evidence. The goal of generative QA is to provide a natural full sentence answer relying on the relevant evidence. Unlike existing models, the system proposed here can generate full answers using knowledge base triplets as evidence and is not restricted to simple questions consisting of one triplet. The generation module is a pretrained encoder-decoder transformer. Additionally, we constructed a new dataset DSberQuAD to train and evaluate the generative QA system in Russian. The new dataset was constructed in a rule-based manner and is an extension of SberQuAD with full sentence answers for each question. The proposed model is a new SOTA for Russian KBQA on RuBQ2.0 dataset. All the code and data from this project are available on GitHub 1 under Apache license.

DEVELOPMENT OF A MORPHOLOGICAL ANALYSER FOR SIBERIAN INGRIAN FINNISH

Ubaleht I., Omsk State Technical University, Omsk, Russia

This paper presents our work on the development of a morphological analyzer for Siberian Ingrian Finnish. Siberian Ingrian Finnish is a low-resource language. In this paper, we present an algorithm for analyzing nouns of Siberian Ingrian Finnish and show an example of analysis.

THE UNIVERSAL DATABASE FOR LEXICAL TYPOLOGY

Voloshina E., University of Gothenburg, Gothenburg, Sweden, **Leonova P.**, HSE University, Moscow, Russia

The paper presents the principles of creating a database for research in lexical typology and describes the possibilities of its use as a linguistic resource. The database is built around semantic fields and frames, i. e. units of analysis in the frame-based theory of lexical typology.

The database provides a universal format for storing the data; therefore, any project in lexical typology can be easily added. The database does not only store the data from previous research projects but allows anyone who wants to contribute to submit data via its web interface. The database includes examples provided by native speakers and manually annotated with translations, semantic fields, and frames, following the annotation principles adopted within the frame approach to lexical typology

PRE-EDITING STRATEGY BASED ON AUTOMATIC EVALUATION OF TRANSLATION COMPLEXITY TO IMPROVE THE QUALITY OF SPECIALIZED TEXTS MACHINE TRANSLATION INTO ENGLISH

Zhivotova A. A., **Berdonosov V. D.**, Komsomolsk-na-Amure State University, Komsomolsk-na-Amure, Russia

The study addresses the issue of applying optimizing pre-editing of Russian-language texts in order to improve the quality of machine translation into English. A probabilistic assessment of translation task complexity is proposed to be used for selecting a pre-editing strategy. A generalized model of the translation process is presented. A mathematical model and algorithm for automated assessment of translation task complexity are proposed. Test of the model on specialized texts of oil and gas industry is described, which showed that the estimate correlates with an estimate of translation quality and can be used in selecting a strategy for optimizing pre-editing of source texts in machine translation tasks.

REMUS, LUPIN AND MOONY WALK IN A BAR... GROUPING OF PROPER NAMES RELATED TO THE SAME DENOTATION IN LARGE LITERARY TEXTS COLLECTIONS

Zykova V. I., National Research University "Higher School of Economics", Moscow, Russia, **Klyshinsky E. S.**, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia

In this article, we present a method of anaphoric proper names detection in fictional texts using Word2Vec model and algorithms of community detection on graphs. This method allows grouping different namings of a single entity and can be useful as a part of preprocessing texts for further analysis such as building social networks or training neural models. The method uses large text collection, related to the same domain. The foundation of the method is training of a Word2Vec model using information on direct characters interactions. This model allows building a social graph of characters. Then, the Louvain algorithm is used to divide the graph into communities containing different names of characters related to the same denotation.

Авторский указатель

Андреева А.	1041	Дугричилов А.	1041	Марголина А. В.	1082
Бажуков М.	1041	Евдокимова А. А.	1071	Михайловский Н.	1032
Баркова Л.	1041	Евсеев Д.	1112	Муравьев Н.	1091
Белова Д. Д.	1001	Жарикова Д.	1097	Мухтарова Г.	1032
Бердоносков В. Д.	1141	Животова А. А.	1141	Петухова К.	1097
Биаджетти Э.	1009	Занки К.	1009, 1021	Резникова Т.	1041
Большаков В.	1032	Зыкова В. И.	1150	Руссо М.	1108
Борисов Е.	1032	Клышинский Э. С.	1150	Смилга В.	1097
Буденная Е.	1041	Колмогорова А. В.	1082	Соченков И.	1052
Бузанов А.	1091	Колобов Р.	1032	Сурков В.	1112
Буянов И.	1052	Комбей К. Р.	1021	Турганбай Р.	1112
Бьянко А.	1021	Кукушкина М.	1091	Убалехт И.	1127
Вилья Л. Б.	1009	Леонова П.	1133	Харламова Д.	1041
Волошина Е.	1133	Литвинцева К.	1041	Чага А. В.	1061
Гордеев Н.	1091	Лураги С.	1009	Яковлева А.	1041
Дробышевский М.	1112	Макарчук И.	1091	Яськова Д.	1052

Author Index

Andreeva A.	1041	Evdokimova A. A.	1071	Petukhova K.	1097
Barkova L.	1041	Evseev D.	1112	Reznikova T.	1041
Vazhukov M.	1041	Gordeev N.	1091	Russo M.	1108
Belova D. D.	1001	Kharlamova D.	1041	Smilga V.	1097
Berdonosov V. D.	1141	Klyshinsky E. S.	1150	Sochenkov I.	1052
Biagetti E.	1009	Kolmogorova A. V.	1082	Surkov V.	1112
Bianco A.	1021	Kolobov R.	1032	Turganbay R.	1112
Bolshakov V.	1032	Kukushkina M.	1091	Ubaleht I.	1127
Borisov E.	1032	Leonova P.	1133	Villa L. B.	1009
Budennaya E.	1041	Litvintseva K.	1041	Voloshina E.	1133
Buyanov I.	1052	Luraghi S.	1009	Yakovleva A.	1041
Buzanov A.	1091	Makarchuk I.	1091	Yaskova D.	1052
Chaga A. V.	1061	Margolina A. V.	1082	Zanchi C.	1009, 1021
Combei C. R.	1021	Mikhaylovskiy N.	1032	Zharikova D.	1097
Drobyshevskiy M.	1112	Mukhtarova G.	1032	Zhivotova A. A.	1141
Dugrichilov A.	1041	Muravyev N.	1091	Zykova V. I.	1150

Научное издание

**Компьютерная лингвистика
и интеллектуальные технологии**

По материалам ежегодной
международной конференции «Диалог»

Выпуск 22, 2023
Дополнительный том

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**