Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2023»

Москва, 14–17 июня 2023 г.

Binary classification model as a tool to detect sentences with microsyntactic units[1]

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 морфо-синтаксическом 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 наX-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*** прострации. [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 СинТагРуса 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 СинТагРуса 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 токенизиацию 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 мироединица ***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 рандомно 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 переназначить 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 *звенит и звенит, 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, видавший 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 искомых выражений в конкретном предложении текста, тем самым позволяя детектировать микросинтаксические единицы и существенно увеличить полноту размекти 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 микросинтаксческой 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 leksemoj 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.