Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2020»

Москва, 17–20 июня 2020 г.

ДИФФЕРЕНЦИАЛЬНЫЕ СЕМАНТИЧЕСКИЕ СКЕТЧИ ДЛЯ РУССКОЯЗЫЧНЫХ ИНТЕРНЕТ-КОРПУСОВ

Деткова Ю. (julia.detkova@abbyy.com)
ABBYY Lab МФТИ

Новицкий В. (valeriy.novitskiy@abbyy.com),
Петрова М. (m.petrova@abbyy.com),
Селегей В. (vladimir.selegey@abbyy.com)
ABBYY

В статье описывается новый тип агрегированной корпусной выдачи — семантические скетчи, получивший пробную реализацию на одном из подкорпусов ГИКРЯ. Семантические скетчи являются естественным распространением идеи корпусных скетчей на анализ сочетаемости в терминах семантических отношений и семантических классов. Уточняющий атрибут «дифференциальный» означает возможность дополнительной параметризации скетчей метатекстовыми характеристиками. Разумеется, построение таких скетчей требует семантической разметки корпуса, в качестве которой в данной работе использовались частичные семантические разборы Compreno. В статье приводятся примеры построенных скетчей и оцениваются достоинства и проблемы корпусной статистики такого рода.

Ключевые слова: семантические скетчи; автоматическая семантическая разметка; семантические классы; глубинные позиции; лексическая сочетаемость, ГИКРЯ

DOI: 10.28995/2075-7182-2020-19-211-227

DIFFERENTIAL SEMANTIC SKETCHES FOR RUSSIAN INTERNET-CORPORA

Detkova J. (julia.detkova@abbyy.com)
ABBYY Lab MIPT

Novitskiy V. (valeriy.novitskiy@abbyy.com),
Petrova M. (m.petrova@abbyy.com),
Selegey V. (vladimir.selegey@abbyy.com)
ABBYY

The current paper suggests a new representation type of word collocations—the semantic sketches. It was first tested on one of the subcorpora of the General Internet-Corpus of Russian. The semantic sketches continue the idea of word sketches based on grammatical relations between words and expand it by adding the semantic information—word meanings and semantic relations between words. Moreover, the sketches can be additionally provided with metatextual characteristics.

Certainly, building such sketches demands the semantic markup of the corpora. Therefore, we have used partial semantic analysis of the Compreno parser for our purposes. The paper demonstrates the examples of the sketches, provides the quality evaluation of the markup they are based on, and shows the advantages and disadvantages of the given approach.

Keywords: semantic sketches; corpora semantic markup; semantic classes; semantic relations; lexical collocations, GICR, Compreno

1. Введение

Идея представления корпусной выдачи в виде статистики сочетаемости анализируемого слова с синтаксически связанными лексическими единицами была предложена и реализована Адамом Килгарифом в рамках проекта SketchEngine [9].

Информация о сочетаемости слова представляется в виде скетчей — обобщенных лексикографических портретов, где зависимые слова классифицируются по грамматическим отношениям: для дочерних зависимых указывается, являются ли они определением, субъектом, объектом или соответствуют другой синтаксической роли.

Подобные скетчи не только позволяют продемонстрировать в сжатом виде основную сочетаемость слова, но также являются удобным инструментом для сравнения сочетаемости разных слов, что бывает полезно при сопоставлении разных лексических единиц одного лексико-семантического поля.

Получение скетчей может основываться на уже имеющейся [частичной] синтаксической разметке корпусов или использовании так называемых скетч-грамматик, имеющих вид контекстно-свободных правил, которые строят отдельные типы синтаксических связей на основе автоматической морфоразметки корпуса [8].

Идея построения синтаксических скетчей оказалась очень продуктивной, а многоязычные интернет-корпуса, собранные в рамках проекта SketchEngine, стали востребованным инструментом корпусных лингвистических исследований.

Надежность и польза полученных скетчей для исследователя ограничена сегодня следующими факторами:

- состав корпуса (зависимость от состава корпуса очевидна, но редко анализируется [4]);
- качество установленных связей (его влияние на получаемую статистику также на удивление остается без внимания);
- невозможность строить скетчи с учетом лексической омонимии.

Последний пункт особенно важен: результаты агрегации по всем значениям лексической единицы сложно интерпретировать, и эту проблему нельзя решить на уровне синтаксиса.

Чтобы снять лексическую неоднозначность, необходимо дополнить имеющийся инструментарий семантическим анализом. Это позволило бы не только решить (насколько это возможно) проблему омонимии, но и учесть диатезное варьирование при реализации семантических отношений.

Целью данной работы является создание семантических скетчей. В них зависимые классифицируются уже не по грамматическим отношениям, а в терминах семантических ролей, таких как Агент, Экспериенсер, Объект, Локатив и т. п., и для каждого слова, помимо грамматических признаков, учитывается также его семантическое значение.

Кроме того, хотелось бы анализировать лексикографические портреты лексем с учетом жанрово-тематического и социолингвистического варьирования [4].

Для решения этой задачи нужны дифференциальные корпуса[1] и достаточно надежные семантические анализаторы, основанные на разметке, которая устроила бы лингвистов и лексикографов.

В качестве корпуса мы естественным образом выбрали ГИКРЯ, поскольку других дифференциальных корпусов для РЯ пока нет. «Ближайший» корпус со свойством дифференциальности [17] реализует пока только идею сегментной различимости данных для лингвистического анализа.

Сама идея скетчей на основании семантических отношений достаточно очевидна, но ее реализации препятствует отсутствие сегодня систематической корпусной семантической разметки. Создать такую разметку с помощью аналога скетч-грамматик, оперирующих синтаксическими отношениями, вряд ли возможно.

Естественным решением данной проблемы представляется прямое использование семантических парсеров. Остановимся подробнее на возможностях их применения для построения семантических скетчей с указанными характеристиками.

---

[1] При дифференциальном подходе [4] всякая корпусная статистика параметризуется метатекстовыми характеристиками: от года создания и социолингвистических характеристик автора до сегментов Интернета, что позволяет обнаруживать статистически значимые различия (смещения) в корпусной выдаче.

2. Использование семантических парсеров при создании семантических скетчей

В настоящее время исследования в области frame semantic parsing ведутся весьма активно, достаточно упомянуть хотя бы популярную для такого анализа DL-платформу SLING [13], основанную на разметке OntoNotes.

Возможность оценки SOTA в этой области (см. раздел 4) дает проведенная на ACL2019 shared task по семантической разметке [11]. Участникам было предложено пять фреймворков для семантической разметки: DELPH-IN MRS Bi-Lexical Dependencies [6], Prague Semantic Dependencies [7], Elementary Dependency Structures [10], Universal Conceptual Cognitive Annotation [1] и Abstract Meaning Representation [3]. Из перечисленных систем разметки только UCCA поддерживает атрибуты ребер, остальные же ограничиваются метками вида “argN”, сравнивать которые нецелесообразно.

Анализ результатов Shared Task on Cross-Framework Meaning Representation Parsing показал, что существующие open-source решения не подходят для семантического анализа русского языка по нескольким причинам.

1. Отсутствие большого разнообразного по составу качественно размеченного семантического корпуса. Это относится ко всем языкам, и к РЯ в частности. Например, на том же треке ACL’2019 для обучения использовались сравнительно небольшие датасеты, объем максимального из которых составляет 56 тысяч предложений. Некоторая семантическая разметка без снятия лексической омонимии реализована в подкорпусе НКРЯ [16], для построения скетчей она, разумеется, не подходит.

2. Отсутствие стандарта семантической разметки: на уже упоминавшемся треке ACL использовались 5 альтернативных систем разметок, и очевидно, акцент был сделан на машинном обучении, а не на сравнительном анализе подходов.

3. Предложенные фреймворки не используют информацию о семантических ролях зависимых, либо используют её очень ограниченно [11].

По указанным выше причинам мы решили использовать для разметки семантическую модель и парсер Compreno [2], предоставленный авторам для исследовательских целей по академической ограниченной лицензии[2].

Основные его характеристики представлены в [2], семантическая часть модели описана в [12], некоторые примеры также представлены в приложении. Отметим две составляющие модели Compreno, релевантные для целей настоящего исследования:

[2] Данная лицензия позволяет размечать и выводить в открытый доступ корпуса ограниченного объема. Кроме того, ABBYY как участник проекта ГИКРЯ предоставляет разработчикам корпуса возможность использовать разметку для целей лингвистических исследований в отдельных подкорпусах, включая отдельные SQL-запросы и агрегацию их результатов.

- во-первых, это организация лексики: все слова представлены в виде семантической иерархии тезаурусного типа, где каждому слову соответствует свой семантический класс (СК), определяющий значение слова и его место в иерархии. Разделение по семантическим классам позволяет рассматривать отдельно разные значения слова;
- во-вторых, — организация семантических отношений между словами: семантические отношения представлены в виде глубинных позиций (ГП). Если валентности в традиционном понимании покрывают, в основном, только актантные зависимые, то ГП описывают все возможные зависимые, как актантные, как и сирконстантные, что позволяет выводить в скетчах полную сочетаемость слова.

Мы использовали не полную семантико-синтаксическую разметку Compreno, а только подмножество разметки с указанием СК и ГП, актуальное для построения семантических скетчей.

Рассмотрим далее особенности выбранной разметки и корпус, на котором строились семантические скетчи.

3. Особенности корпуса и разметки

В настоящей работе для экспериментов был использован журнальный подкорпус ГИКРЯ.

Объем подкорпуса:

- 74 тысячи документов,
- 24,7 млн предложений,
- 326 млн слов.

Все тексты были размечены с помощью технологии ABBYY Compreno [15]. Результатом разметки является семантическая структура для каждого предложения в виде дерева семантических отношений между семантическими классами входящих в него слов.

Для целей данного исследования было решено ограничиться частичной семантической разметкой, где маркируются только поддеревья глаголов: все глаголы размечаются семантическими классами и глубинными позициями для их непосредственных зависимых; все зависимые, для которых определены ГП, маркируются также и по СК, что позволяет нам получать пары вида "глагол:СК [ГП: зависимая:СК]", например:

Окна выходят в сад.
"выходить:TO_FRONT" [Object: окно "окно:WINDOW_OF_BUILDING"]
"выходить:TO_FRONT" [в Locative_FinalPoint: сад "сад:сад:GARDEN"]

На семантических отношениях этих поддеревьев и строятся скетчи в первой версии.

Вершиной такого поддерева является семантический класс глагола (наиболее вероятный по результатам анализа), его дочерние элементы связаны с ним семантическими отношениями — ГП из некоторого набора Compreno

(примеры используемых для построения скетчей ГП и иллюстративный фрагмент иерархии семантических классов приведены в приложении).

Из атрибутов ядра оставлены лексема, семантический класс, морфологические граммемы категории GrammaticalType (часть речи) и синтаксические граммемы категории SyntacticCase (падеж). На рис. 1 ниже представлен разбор предложения с указанной информацией при лексемах и семантическими отношениями между ними.

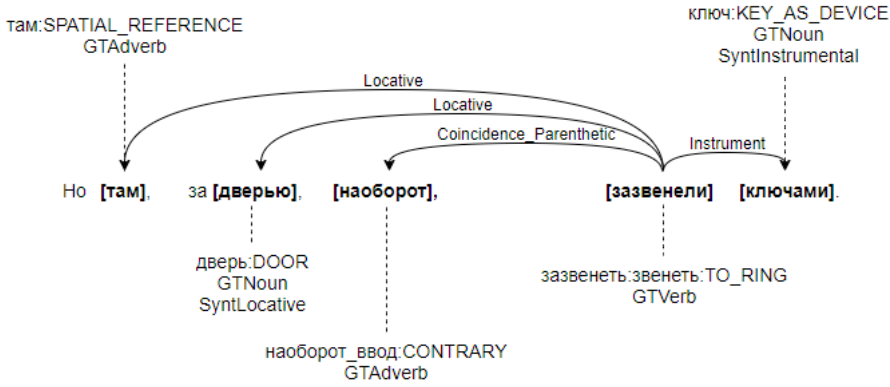


Рис. 1. Разбор предложения с указанными семантическими отношениями, семантическими классами и грамматическими значениями ядер составляющих

Таким образом, на данном этапе мы не размечали СК и ГП для прилагательных, именных групп при существительных и прочих неглагольных зависимых, для синтаксически перемещенных групп, а также зависимые при эллиптированных глаголах (в предложениях типа *Мы правы.* не маркируются ГП и СК при эллиптированном «быть») и сами эллиптированные зависимые (например, в предложении *Составители попросили каждого из поэтов сочинить…* подразумевается эллиптированное ядро «поэта» в составляющей «каждого (поэта) из поэтов»).

Большинство этих ограничений не является принципиальным, поскольку, во-первых, используемый парсер данную информацию предоставляет, а, во-вторых, степень влияния таких случаев на скетчевую выдачу предоставляется несущественной (что, впрочем, планируется проверить на следующих этапах исследования).

Местоимения размечаются классами вида PRONOUN в тех случаях, когда не удается установить по контексту связь с референтом, и классом референта — когда такая связь устанавливается. В соответствии с этим возможны такие варианты разметки:

#— А как зовут "звать:TO_CALL_AND_DESIGNATE" Object: *ее* "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"?

— референт «ее» не найден;

#— А как зовут "звать:TO_CALL_AND_DESIGNATE" Object: *ее* "болонка:PET_DOG"*?*

— связь с референтом «ее» установлена по контексту.

В итоге, размеченный корпус содержит:

- 55 тысяч различных семантических классов;
- 91 тысяча различных лексических классов (под лексическим классом понимается потомок семантического класса в конкретном языке);
- 305 различных глубинных позиций;
- 95 млн экземпляров семантических отношений — то есть, размеченных вхождений данных ГП в корпусе.

Разметка получена с помощью API, предоставленного ABBYY для исследований по ограниченной лицензии, и выгружена в локальную базу данных с примерами для сбора статистик. Также написан экспортер разметки в формат xml, необходимый для публикации корпуса-прототипа в open-source, ссылка на который представлена ниже.

4. Оценка качества семантической разметки

Для оценки качества разметки случайным образом было отобрано 200 предложений из журнального зала ГИКРЯ. Автоматическая разметка была проверена и исправлена вручную, и уже с полученной таким образом эталонной разметкой проводилось сравнение.

В качестве метрики была выбрана официальная метрика для ACL Shared Task 2019 —Meaning Representation Parsing [11]. Эта метрика (далее MRP F1) представляет собой невзвешенное среднее между F1-score для различных объектов разметки: узлы дерева разбора, метки узлов (семантические классы), атрибуты узлов (граммемы), привязка к тексту, связи, атрибуты связей (глубинные позиции).

Привязка к тексту сравнивается максимально мягко: не учитываются ошибки на пробельных символах, пунктуации, скобках и кавычках, не учитываются ошибки разбиения аннотации на несколько подряд идущих. MRP F1 разработана как универсальная метрика для сравнения деревьев разбора различных форматов представления семантической структуры и хорошо адаптируется к отсутствию любого из объектов разметки. Так, мы не учитывали качество определения граммем ядер составляющих, так как они не влияют напрямую на качество скетчевой выдачи. Кроме того, ввиду ограничений разметки, взятой для эксперимента, мы имеем дело не с полными деревьями разбора, а с их не обязательно связными подграфами. Выбранная метрика не противоречит этой модификации, но нужно учитывать, что сравнение с результатами MRP Shared Task 2019 не вполне корректно по нескольким причинам: мы оцениваем разметку на другом языке, другом корпусе и с упрощенной структурой.

В результате была получена следующая оценка для корпуса: MRP F1 = 96,78, что является достаточно высоким значением для автоматического

парсера. Для сравнения, максимальное значение MRP F1, которого удалось достичь на ACL 2019 Shared Task, составляет 0,84 (оценка для 100 случайных предложений из The Little Prince) [11].

Основное количество ошибок было связано с выбором СК для нужного значения глагола (например, «выйти» в буквальном значении движения в «*статья вышла в свет*» вместо «выйти» в значении «появиться»), с выбором СК для зависимой (например, "нос:NOSE_AS_FRONT_PART" в «*из носу кровь*» вместо ожидаемого «нос:NOSE»), а также с выбором ГП — как правило, в случае их возможной омонимии (например, Locative для «[*в книге*] *написано*» вместо предполагаемого MetaphoricLocative (разбор с неметафорическим локативом был бы уместен в предложении типа «[*в книге*] *лежала закладка*»)).

Более подробно с корпусом, использованным для оценки качества разметки, можно ознакомиться по ссылке: SemanticSketchCorpora.

5. Семантические скетчи

Описанная семантическая разметка позволяет нам строить семантические скетчи с обозначенными выше характеристиками, а именно — с учетом разделения слова на разные значения, с группировкой зависимых по семантическим отношениям и возможностью единообразно представлять одинаковые семантические роли независимо от их разного синтаксического выражения при разных предикатах.

Рассмотрим теперь технические средства построения скетчей данного формата.

5.1. Обозначения

Здесь и далее используются обозначения в заголовках таблиц с примерами:

- **Lexeme** — лексема ядра составляющей
- **SemanticClass** — семантический класс составляющей
- **SlotName** — семантическая роль дочерней составляющей
- **ChildText** — текст заполнителя
- **Count** — поддержка коллокации в корпусе
- **F** — частота встречаемости коллокации, умноженная на 10^6 для удобства отображения
- **MI** — метрика ассоциации Mutual Information
- **Dice** — метрика ассоциации logDice

5.2. Метрики

Ранжирование только по частоте встречаемости сочетания в корпусе не представляет большого интереса, так как не несёт никакой информации о том, насколько то или иное словосочетание характерно для исследуемого слова или семантического класса. Поэтому мы предусмотрели возможность ранжировать скетчевую выдачу по нескольким метрикам:

Частота встречаемости: $f(x,y)$

Mutual Information: $MI(x,y) = log_2\left(\frac{f(x,y)}{f(x)f(y)}\right)$

logDice: $Dice(x,y) = 14 + log_2\left(\frac{2f(x,y)}{f(x)+f(y)}\right)$

Так как метрики MI и logDice чувствительны к случайным сочетаниям, было решено использовать их вместе с ограничением на частоту встречаемости в корпусе. Примеры, для которых она ниже некоторого устанавливаемого пользователем порога, не попадают в скетчевую выдачу. Преимущество метрики logDice хорошо видно на следующих примерах:

(1) Топ-5 коллокаций глагола "сообщать:TO_INFORM" по частоте встречаемости:

SlotName	ChildText	Count	f	MI	Dice
Agent	он	1042	287.53	3.97	6.75
Theme	об этом	694	191.5	8.3	9.65
Addressee	ему	580	160.04	7.36	9.14
Agent	она	394	108.72	3.93	6.65
Addressee	ей	247	68.16	7.18	8.78
Theme	о том	215	59.33	8.46	9.44

(2) Топ-5 коллокаций глагола "сообщать:TO_INFORM" по logDice:

SlotName	ChildText	Count	f	MI	Dice
Ch_EvaluationOfHumanTemperAndActivity	доверительно	144	39.73	14.09	12.08
Ch_Information	по секрету	49	13.52	14.56	11.49
Metaphoric_Route	по рации	22	6.07	14.34	10.84
TextStructure	настоящим	11	3.04	15.34	10.79
Ch_Information	под большим секретом	13	3.59	15.67	10.72
Ch_Information	конфиденциально	13	3.59	15.45	10.68

5.3. Механизм построения скетчей

Все данные, необходимые для построения скетчей, выгружаются из корпуса в базу данных SQLServer. Там же хранятся списки ГП и СК. Для ускорения вычисления метрик MI и logDice в базе сохранены предпосчитанные частоты встречаемости для лексических и семантических классов ядер составляющих. Непосредственно вычисления оформлены в виде хранимых процедур T-SQL. Для получения данных из базы и визуализации скетчей использованы библиотеки Python: pyodbc, pandas, seaborn.

5.4. Семантические классы

Основным преимуществом нового инструмента является взаимодействие с семантической разметкой предложений. Это позволяет снять омонимию в большинстве контекстов: на этапе разметки для каждого слова при помощи парсера определен наиболее вероятный класс, к которому слово может принадлежать в данном контексте. Как следствие, мы можем строить скетчи для различных значений слова. Можно продемонстрировать разницу на примере глагола «выходить»:

(3) Примеры для различных семантических классов глагола «выходить»:

Класс	Пример употребления
TO_FRONT	*Окна выходят в сад.*
TO_WALK	*Мальчик вышел из комнаты.*
TO_TAKE_PLACE	*Из этой затеи ничего не вышло.*
TO_TREAT_AND_CURE	*Выходить больного щенка.*

Приведем семантические скетчи для некоторых из данных значений (для удобства сравнения представим скетчи в виде таблиц без частотных характеристик, но с разделением по ГП, аналогично тому, как в Sketchengine представлено разделение по синтаксическому выражению). Здесь и далее примеры приведены в порядке убывания метрики logDice:

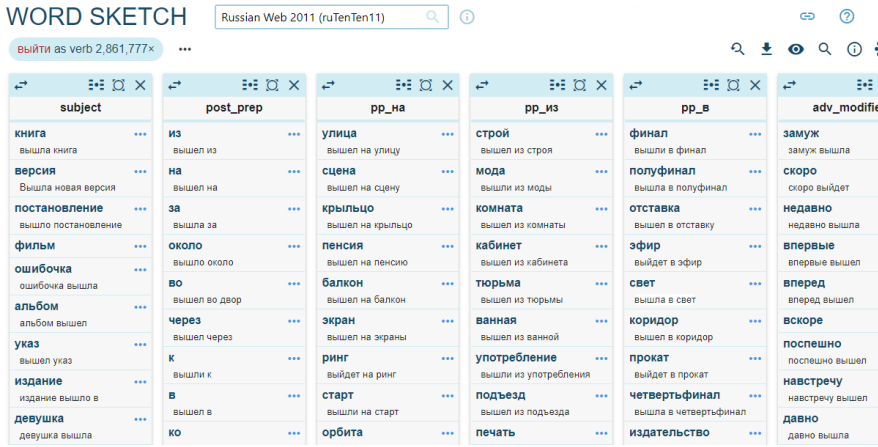
(4) Семантический скетч для глагола "выходить:TO_TAKE_PLACE":

Ch_Relation_Coincidence	Modality	Object_Situation	Ch_Evaluation	Time	Locative
наоборот вышло наоборот	само_собой вышло само собой	заминка вышла заминка	складно выходило складно	четверть выходила за четверть	сумма вышло в сумме
такой вышло так	так вышло так	скандалить вышел скандал	нехороший вышло нехорошо	как-то вышло как-то	рассказ выходит в рассказах гуцко
иной вышло по-иному	на_деле вышло на деле	размолвка вышла размолвка	скверный выходило скверно	восьмой вышла восьмого марта	практик выходит на практике
похожий выходило очень похоже	неправдоподобный вышло крайне неправдоподобно	казус вышел казус	красивый вышло красиво	иной_раз выходит иной раз	конкурентка выходит у конкурентки
другой вышло по-другому	криво вышло криво	неприятность вышли неприятности	ничего вышло ничего	подчас выходило так подчас	Земский выходит у него
этакий вышло этак	кривовато выходило кривовато	ссориться вышла ссора	паскудный вышло паскудно	ноябрь выходило на прошлогодний ноябрь	Вяльцев выходит у вяльцева

(5) Семантический скетч для глагола "выходить: TO_FRONT":

Locative_FinalPoint	Object	Locative_PartAsOrientation	Locative_Orientation_FinalPoint	Time	Locative
двор выходили во двор	окно выходили окна	окно выходила окнами	запад выходили на запад	поныне выходят и поныне	Соня выходят у нас
сад выходили в сад	балкон выходил балкон	дверь выходили дверями	север выходят на север	частенький выходил частенько	ярус выходили на втором ярусе
дворик выходило во внутренний дворик	фасад выходил его главный фасад	фасад выходили фасадом на тверскую	восток выходят на восток	параллельный выходила параллельно	n_этажка выходит в каждой пятиэтажке
улица выходили на улицу	окошко выходило окошко	конец выходит одним концом	юг выходит на юг	как_раз выходило как раз	Питер выходят в питере
проспект выходили на проспект	веранда выходила веранда	стена выходил северной стеной	сторона выходит в сторону	иной_раз выходили иной раз	крыша выходили поверх низких крыш
север выходили на север	подъезд выходили ни подъезды		напротив выходили напротив моих	осень выходил под осень	кафе выходили у него

Для сравнения приведем также скетч для «выйти», получаемый на sketchengine.eu:



Как видно, выдача содержит разные значения данного глагола: «выйти на крыльцо, на балкон» соответствует СК TO_WALK выше, — «выйти в свет, в прокат, вышел альбом» — значению «появиться», «ошибочка вышла» — «произойти» (TO_TAKE_PLACE) и так далее.

Конечно, в ряде случаев омонимия разрешается неправильно и в нашей модели. Для примера приведем скетч "выйти:TO_WALK":

Locative_FinalPoint	Locative_InitialPoint	Time	Agent	Agent_Metaphoric	OrderInTimeAndSpace
улица вышел на улицу	дом вышел из дома	утро вышел утром	человек выходили люди	книга вышла книга	наконец вышел наконец
двор вышел во двор	комната вышел из комнаты	только_что вышел только что	женщина вышла женщина	издание вышло второе издание	потом вышел потом
коридор вышел в коридор	из_дому вышел из дому	минута вышел через минуту	мужчина вышел мужчина	срок вышел срок	наконец-то вышел наконец-то
сцена выходит на сцену	кабинет вышел из кабинета	вечер вышел вечером	девушка вышла девушка	сборник вышел сборник	затем вышел затем
крыльцо вышел на крыльцо	машина вышел из машины	ранний вышел рано	старик вышел старик	что_interrog вышло чего	снова вышел снова
свет вышел в свет	подъезд вышел из подъезда	час вышел через полчаса	жена вышла жена	роман вышел роман	опять вышел опять

Как видно, примеры, попавшие в колонку с ГП Agent_Metaphoric, не на данное значение. Данная ошибка — результат неправильного семантического анализа, на который, в данном случае, влияет как модель, так и статистика. Это позволяет надеяться, что при дальнейшем использовании и обучении на текстах большего объема число подобных ошибок будет уменьшаться.

5.5. Глубинные позиции

Семантическая разметка позволяет оперировать не поверхностными, синтаксическими позициями, а глубинными, семантическими. Это, с одной стороны, позволяет различать разную семантику в отношениях, имеющих одинаковое синтаксическое выражение, как в примерах (6) ниже:

(6) Примеры, в которых зависимые слова имеют одну поверхностную позицию Object_Dative, но разные глубинные:

Глубинная позиция	Пример
Addressee	*Рассказать сказку* [детям].
Experiencer	*Позволить* [гостю] *войти*.
Possessor	*Дать* [другу] *книгу*.

С другой стороны, дает возможность сопоставить с ГП релевантные семантические отношения даже в случае их разного синтаксического выражения. Сравним, к примеру, семантические скетчи двух лексических классов из СК TO_INFORM — «сообщать» (7) и «извещать» (8):

(7) Семантический скетч для глагола "сообщать:TO_INFORM":

Agent	Addressee	Theme	Object_Situation	Time	Object
газета сообщали газеты	читатель сообщает читателю	это сообщил об этом	адрес сообщил адрес	мимоходом сообщает мимоходом	новость сообщил новость
радио сообщило радио	родитель сообщил родителям	смерть сообщила о его смерти	подробность сообщить какие-либо подробности	немедленный сообщить немедленно	сведения сообщал сведения шпионского характера
автор сообщает автор	жена сообщил жене	то сообщает о том	номер сообщили выигравший номер	между_делом сообщил между делом	весть сообщил радостную весть
агентство сообщает агентство	мать сообщил матери	результат сообщить о результатах	собираться сообщила и собирается бросать университет	незамедлительный сообщать незамедлительно	известие сообщить пренеприятное известие
СМИ сообщают сми	друг сообщил другу	событие сообщает об этом событии	следовать сообщить куда следует	вскользь сообщается вскользь	факт сообщать факты
источник сообщает источник	отец сообщил отцу	приехать сообщить о своем приезде	вещь сообщил важные вещи	срочный сообщить срочно	информация сообщить информацию

(8) Семантический скетч для глагола "извещать:TO_INFORM":

Addressee	Theme	Agent	Time	Object_Situation	Agent_Metaphoric
полиция известить полицию	прибыть известила о прибытии	Григоров извещал григоров	своевременный извещены своевременно	пожечь известили и пожгли	объявление извещало объявление
граф извещен граф николай петрович румянцев	решать извещу о моем решении	доброжелатель известят редакционные доброжелатели	непосредственно извещать непосредственно	предать_суду известили и предан суду	табличка извещала табличка на заборе
родственник известить родственников	кончина извещены о кончине владыки	генерал-губернатор известил генерал-губернатор граф ростопчин	заранее известить заранее	принимать известил что статья принята	бейдж известил бейдж на директорском пиджаке
родные известить родных	приехать известить о вашем приезде	Измайлов извещал измайлов со вздохом облегчения	предварительный извещать предварительно	причаливать известил что причаливает	пометочка извещает пометочка другой рукой
общественность извещают и общественность	рожать извещать о рождении наследника престола	газета известили газеты	нескоро известят нескоро	миновать известил что опасность бомбежки миновала	Известия извещают известия
Штейнер извещал штейнера	визит известить о визите бога	яхта известила яхта наша	заблаговременный извещены заблаговременно	подлежать извещает и тоже подлежит наказанию	мейл извещал второй мейл

Как видно, ГП Addressee имеет разное синтаксическое выражение при данных глаголах и соответствует прямому объекту при «извещать» и дативу — при «сообщать». Однако это не мешает продемонстрировать общность семантической модели данных глаголов в отношении валентности адресата.

5.6. Семантические скетчи для семантических классов

Выше мы рассматривали примеры скетчей для отдельных слов. Тем не менее, интересным может быть также создание скетчей для целых семантических классов, что позволило бы сравнивать разные лексемы одного семантического поля (или, в терминах нашего формализма, разные лексические классы одного СК).

В качестве примера приведем фрагмент скетча для класса TO_COMMIT — лексической функции, где многие глаголы имеют весьма ограниченную сочетаемость:

Object_Situation	Object	Agent	Time	Agent_Metaphoric	Modality
роль	это	человек	год	это	правильный
сыграли свою роль	сделать это	сделал он	делали в прошлом году	произвело это	сделал правильно
впечатление	вопрос	автор	жить	что_Interrog	так
произвести впечатление	задавать вопросы	делает автор	сделал в жизни	сделал я	сделал так
участвовать	решение	власти	сейчас	книга	действительно
приняли участие	принято решение	делает власть	делать сейчас	производит книга	сделал действительно
шагать	вывод	поэт	тогда	оно	конечно
сделал шаг	делать выводы	делает поэт	делать тогда	сделал он	сделал конечно
дело	что_Interrog	отец	время	жить	возможно
сделано дело	делать что	делал он	сделал в свое время	дала жизнь	сыграло возможно
должное	оно	писатель	теперь	рука	может_быть
отдать должное	взять их	делает писатель	делать теперь	делают руки	сделал может быть

Наиболее показательна сочетаемость с объектными позициями, где проявляется лексикализованность сочетаемости рассматриваемых глаголов (*принять участие* vs *задать вопрос* vs *сыграть роль* и под.).

5.7. Другие возможности

Помимо перечисленного, в предлагаемом инструменте реализованы следующие возможности:

- ограничение части речи и падежа ядра зависимой составляющей;
- ограничение части речи и падежа ядра родительской составляющей;
- ограничение на количество упоминаний сочетания в корпусе — при желании можно включать в скетчи только зависимые, встретившиеся больше n раз.

Кроме того, при создании семантических скетчей в качестве зависимых модель позволяет выводить как фрагменты размечаемых предложений (либо равные составляющим, либо ядра составляющих с необходимыми грамматическими характеристиками — в релевантных предложно-падежных формах, например), так и ядра семантических классов, лексических классов или лексем.

6. Анализ результатов

Новизной настоящего исследования является эксперимент по построению семантических скетчей на значительном корпусном материале.

Семантические скетчи позволяют получать сравнительные портреты слов из одного семантического поля независимо от разницы в их синтаксических моделях, а также получать данные о сочетаемости целых семантических классов и учитывать конкретные значения рассматриваемых слов.

Реализация на основе разметки Compreno показала вполне приемлемые результаты, хотя некоторой проблемой является большое количество типов глубинных позиций — более 300. Эту проблему еще предстоит решить при включении механизма семантических скетчей в функционал ГИКРЯ (мы надеемся на фидбэк от пользователей корпуса).

7. Заключение и планы на будущее

В настоящий момент проводится обучение SLING-подобных технологий на корпусе RuSemSketches. Стоит задача попробовать применить обученную нейросеть для семантической разметки всего корпуса ГИКРЯ (при этом скетчевая разметка все же проще, чем полная семантическая разметка, что дает надежду получить приемлемые для пользователей результаты). Мы рассчитываем, что на основе RuSemSketches удастся провести shared task по семантической разметке на следующем Диалоге.

В ближайшее время планируется включить семантические скетчи на основе разметки Compreno в функциональность новой версии ГИКРЯ. Существенной особенностью скетчей в ГИКРЯ является их дифференциальность: возможность сравнения результатов с учетом всех доступных в ГИКРЯ метатекстовых признаков. Это позволит в ближайшей перспективе проводить исследования по дифференциальной лексической семантике.

8. Благодарности

Данная работа выполнена совместно исследовательской лабораторией ABBYY Lab МФТИ (проект ГИКРЯ) и отделом Advanced Research Development компании ABBYY. Мы благодарим всех коллег, и в особенности Константина Дружкина и Евгения Инденбома за полезную критику и помощь в проведении исследования.

Литература

1. *Abend O., Rappoport A.* (2013), UCCA. A semantics-based grammatical annotation scheme, Proceedings of the 10th International Conference on Computational Semantics, Potsdam, Germany, pp. 1–12.
2. *Anisimovich K., Druzhkin K., Minlos F., Petrova M., Selegey V., and Zuev K.* (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", vol. 11, pp. 91–103.

3. *Banarescu L., Bonial C., Cai S., Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn P., Palmer M., Schneider N.* (2013), Abstract Meaning Representation for sembanking, Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, pp. 178–186.
4. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.,* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation, Web as Corpus Workshop (WAC-8).
5. *Belikov V., Sharoff S., Kopylov N. et al.* (2015), Corpus with automatically removed morphological ambiguity: methodology of linguistic research [Korpus s avtomaticheski snyatoi morfologicheskoi neodnoznachnostju: K metodike lingvisticheskih issledovanij], Computational Linguistics and Intellectual Technologies [Kompi'uternaia lingvistika i intellektualnye tekhnologii], Vol. 14, № 1. pp. 84–95.
6. *Hajic J., Hajicova E., Panevova J., Sgall P., Bojar O., Cinkova S., Fucıkova E., Mikulova M., Pajas P., Popelka J., Semecky J., Sindlerova J., Stepanek J., Toman J., Uresova Z., Zabokrtsky Z.* (2012), Announcing Prague Czech-English Dependency Treebank 2.0., Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp. 3153–3160.
7. *Ivanova A., Oepen S., Øvrelid L., Flickinger D.* (2012), Who did what to whom? A contrastive study of syntacto-semantic dependencies, Proceedings of the 6th Linguistic Annotation Workshop, Jeju, Republic of Korea, pp. 2–11.
8. *Kilgarriff A., Rychlý P., Smrž P., Tugwell D.* (2004), The sketch engine, Information Technology.
9. *Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V, Michelfeit J., Rychlý P., Suchomel V.* (2014), The Sketch Engine: ten years on, Lexicography, 1: 7–36.
10. *Oepen S., Lønning J. T.* (2006), Discriminant-based MRS banking, Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, pp. 1250–1255.
11. *Oepen S. et al.* (2019), MRP 2019: Cross-Framework Meaning Representation Parsing, CoNLL 2019.
12. *Petrova M. A.* (2014), The Compreno Semantic Model: The Universality Problem, International Journal of Lexicography, Vol. 27, Issue 2, pp. 105–129.
13. *Ringgaard M., Gupta R., Pereira F. C. N.* (2017), SLING: A framework for frame semantic parsing.
14. *Rychlý P.* (2008), A Lexicographer-Friendly Association Score, Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN, Brno, pp. 6–9.

Интернет-ресурсы

15. https://www.abbyy.com/ru-ru/science/technologies/compreno/
16. http://www.ruscorpora.ru/new/corpora-sem.html
17. https://tatianashavrina.github.io/taiga_site/
18. http://www.webcorpora.ru/

Приложение

Семантические классы

Семантический класс — это универсальная (разделяемая всеми языками системы) единица лексического описания. Включает базовое лексическое значение конкретного языка со всеми его продуктивными грамматическими и семантическими дериватами.

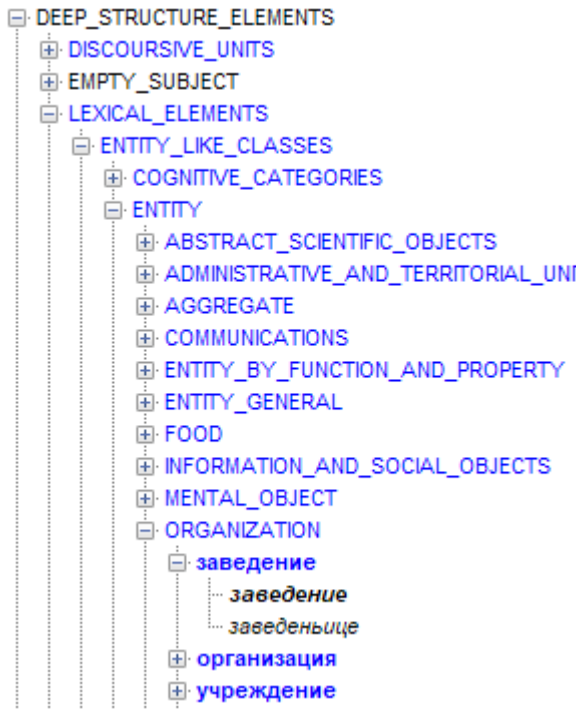


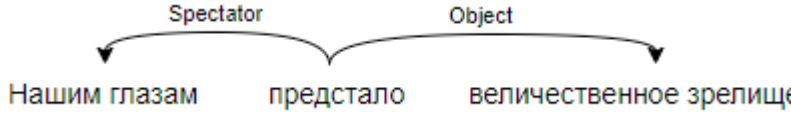
Рис. 2. Общий вид семантической иерархии

Каждому классу соответствует набор ограничений, накладываемых на семантические отношения, которые представители класса могут заполнять.

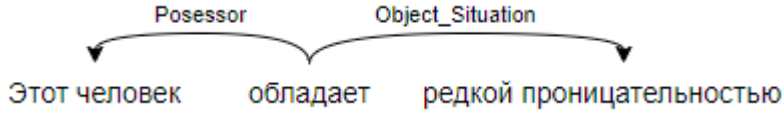
Семантические отношения

Ниже приведены краткие описания некоторых глубинных позиций и примеры размеченных фрагментов текста.

- **Object** — объект действия в широком смысле.
- **Spectator** — Свидетель, зритель — лицо, при котором происходит какое-то действие, но которое не вовлечено в это действие.



- **Object_Situation** — ситуационный объект.
- **Possessor** — посессор, обладатель.



- **Agent** — агенс (в широком смысле).
- **Locative_InitialPoint** — исходный пункт, начальная точка.