ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ СИНТАКСИЧЕСКИХ ШАБЛОНОВ ИЗ КОРПУСА ТЕКСТОВ

Automatic Extraction of Grammatical Patterns with Clear Syntactical Structure

Студентка 2 курса
группы № БКЛ194

Чевелева Анастасия Николаевна

Научный руководитель

Клышинский
Эдуард Станиславович

доцент Школы
лингвистики Факультета
гуманитарных наук

Москва, 2021 г.

Оглавление

1. Введение

1.1. Виды синтаксического анализа

Одним из основных этапов автоматической обработки естественного языка является синтаксический анализ (parsing). Он необходим, например, при извлечении фактов из текста (Cowie & Lehnert 1996). Принято различать полный синтаксический анализ, при котором определяются связи всех слов в предложении и строится его синтаксическая модель (дерево) (Большакова и др. 2017: 101), и поверхностный (или частичный) синтаксический анализ, в ходе которого выделяются лишь некоторые синтаксически связанные группы соседних слов (chunks) (Abney 1991), при этом часто без установления конкретного типа связи. Тем не менее для решения многих задач достаточно поверхностного анализа, так как он более быстрый и менее ресурсозатратный по сравнению с полным (Feldman & Sanger 2007), поскольку анализирует меньшее количество единиц с меньшей степенью подробности.

Расцвет данного подхода пришёлся на начало XXI века (см., например, (Ножов 2003)), однако затем широкое распространение получили нейросетевые подходы за счёт их универсальности. Наиболее популярными синтаксическими парсерами на данным момент являются (UDPipe 2.0) и (spaCy v3.0), в основе обоих парсеров лежат нейросети. Несмотря на это, поверхностный синтаксический анализ ещё не исчерпал свой потенциал. В следующих разделах мы рассмотрим существующие методы поверхностного синтаксического анализа, а также предложим метод автоматического выделения синтаксических шаблонов с однозначным подчинением.

1.2. Методы поверхностного синтаксического анализа

Для проведения поверхностного синтаксического анализа могут использоваться различные методы. Так в (Molina, Pla 2002) используются скрытые Марковские модели, в (Кудинов 2013) — условные случайные поля, в (Большакова и др. 2007) — контекстно-свободные грамматики, а в (Смирнов, Шелманов 2013) упоминаются как конечные автоматы, так и различные методы машинного обучения.

Упомянутые методы можно разделить на две группы: эвристические методы (Добров 2017) и методы машинного обучения. Последние сами выделяют синтаксические конструкции из размеченного корпуса, обеспечивая большую полноту, чем эвристические методы, но при этом не всегда показывают сопоставимую точность и работают с меньшей скоростью. С другой стороны, в работе (Ножов 2003) для выделения именных групп используются конечные автоматы, обладающие высокой скоростью работы. Однако создание подобных автоматов требует длительного ручного труда экспертов-лингвистов по поиску кандидатов в извлекаемые конструкции. В особенности это относится к языкам с развитой морфологией и свободным порядком слов (Большакова и др. 2017: 103), поскольку данные факторы понижают полноту извлекаемых конструкций.

Тем не менее в данной работе мы утверждаем, что для русского языка, обладающего флективной морфологией, согласованием и относительно свободным порядком слов, существует достаточно много конструкций, синтаксические связи внутри которых можно с достаточной точностью определить, исходя только из частеречной принадлежности входящих в конструкцию словоформ.

1.3. Синтаксические шаблоны

В общем случае лингвистические шаблоны определяются как структурное описание языковой конструкции, отражающее некоторые её свойства, и которую затем ищут в тексте (Большакова и др. 2007). В данной работе под синтаксическим шаблоном мы будем понимать последовательность частеречных маркеров (например, *ADP, NOUN, PUNCT*), а под синтаксическими конструкциями — последовательность частеречных маркеров и связей, определяемых в пределах данного шаблона с долей верных ответов (accuracy) не ниже 0.97 (например, *ADP, 2, NOUN, PUNCT*, где числа обозначают номер слова (относительно длины шаблона), являющегося вершиной для слова, записанного слева от числа). Доля верных ответов не ниже 0.97 означает, что среди всех возможных реализаций данного шаблона это подчинение встречается в не менее чем 97% случаях. Данный порог выбран, во-первых, исходя из эмпирических представлений о вероятности ошибки, совершаемых при ручной разметке, и

во-вторых, в соответствии с результатами упоминавшихся выше синтаксических парсеров (для (spaCy v3.0) доля верно определяемых конструкций составляет 0.96).

Таким образом, цели настоящей работы заключаются в следующем:

- Выяснить для русского языка, какова доля конструкций с однозначной синтаксической структурой;
- Понять, возможно ли автоматически извлечь эти конструкции из корпуса с готовой синтаксической разметкой.

Задачами являются создание соответствующего метода и реализация его в виде алгоритма на языке программирования Python.

1.4. Материалы для исследования

Среди имеющихся синтаксических корпусов для русского языка для данного исследования был выбран глубоко аннотированный корпус SynTagRus, поскольку он является самым крупным из представленных в проекте морфосинтаксической разметки Universal Dependencies (61 889 предложений, 1 106 296 токенов) и также характеризуется полностью снятой морфологической и синтаксической омонимией (UD Russian). Тексты, размеченные в формате CoNLL-U, взяты из соответствующего репозитория (GitHub SynTagRus). Языковые примеры-иллюстрации, приведенные в данной работе, также взяты из корпуса SynTagRus.

В формате CoNLL-U (CoNLL-U Format) разбор каждого предложения начинается со строк с его идентификационным номером и с текстом предложения. Затем следует разбор всех элементов предложения (слов, знаков препинания, опущенных компонентов). Каждый элемент записывается с новой строки, каждая строка содержит 10 полей: номер элемента в предложении, словоформа/ пунктуационный знак, лемма или основа словоформы/ пунктуационный знак, частеречный маркер в формате UD и т.д. Однако для данного исследования понадобятся не все поля разметки (подробнее см. в разделе 3.1).

Ссылка на репозиторий с полным кодом и материалами работы размещена в Приложении[1].

[1] https://github.com/ancheveleva/grampatterns

2. Метод выделения синтаксических шаблонов

2.1. Извлечение n-грамм

На начальном этапе мы разбиваем синтаксически размеченный корпус на предложения. Дополнительно к существующей разметке добавляем маркеры начала и конца предложений, так как они могут быть хорошими маркерами-границами синтаксически связанных групп. Затем делим предложения на n-граммы, установленной ширины окна в n элементов, сохраняя для каждого элемента информацию о номере и части речи его родительской вершины. При этом мы не рассматриваем опущенные слова (empty nodes), а также исключаем из числа n-грамм такие, что содержат иностранные слова, так как подобные вхождения не несут практической пользы для нашего метода. В оставшихся n-граммах номера вершин назначаются относительно начала выбранного окна. Для слов, вершина которых оказалась снаружи n-граммы, постулируется отсутствие внутренней связи.

В результате полученные конструкции можно представить в виде строки POS1, #host1, POShost1, …, POSn, #hostn, POShostn, total_entries, где POSi, #hosti, POShosti — часть речи, номер родительской вершины и часть речи родительской вершины i-го слова n-граммы, а total_entries — число вхождений каждой конструкции в корпус.

На данном этапе при условии, что мы взяли достаточно большой корпус и для повышения полноты нашего метода извлекли цепочки разной длины (n, n+1, …, n+k-граммы), получается огромное число конструкций. Для уменьшения объёма данных можно было бы ограничиться только наиболее частотными шаблонами. Однако проблема заключается в том, что на один шаблон может приходиться разное количество конструкций, при этом данное отношение мало коррелирует с частотностью шаблона.

Так, для очевидно частотного шаблона *NOUN, ADJ, NOUN* синтаксические связи не определяются однозначно, то есть можно выделить как минимум две конструкции: последнее существительное относится к первому существительному (1a), или имеет родительскую вершину снаружи (1b).

(1) NOUN, ADJ, NOUN

a. <...> ***обострение холодной войны*** (конструкция *NOUN, ADJ, NOUN, 1*) [И к ним не зарастет народная тропа]

b. <...> *не представляли для* ***власти особой проблемы***. (конструкция *NOUN, ADJ, NOUN, _*) [Колчак, Александр Васильевич]

С другой стороны, неверно утверждать, что среди менее частотных шаблонов не найдется удачных: например, связь частицы и наречия в шаблоне *ADP, PART, ADV* (2) определяется однозначно, то есть выделяется всего одна конструкция.

(2) ADP, PART, ADV

В не менее *правильных советских фильмах* <...>. (конструкция *ADP, PART, 3, ADV*) [Обнищание обещаний]

Решение данной проблемы предлагается в следующих разделах.

2.2. Доли конструкций в шаблонах

Для получения шаблонов мы избавляемся от информации о родительских вершинах (об их номерах и частях речи) и группируем одинаковые строки по частям речи, то есть по POS1, ..., POSn, суммируя число вхождений. Затем всем конструкциям из прошлого этапа ставятся в соответствие их шаблоны с общим числом вхождений. После этого в каждой конструкции рассматриваем связь каждого слова по отдельности: вычисляем количество конструкций с данным подчинением рассматриваемого слова и долю этого подчинения среди шаблона.

Таким образом, мы получаем список новых конструкций, в которых информация о варианте подчинения, количестве вхождений и доле присутствует только у одного из слов (информация о родительских вершинах остальных слов n-граммы игнорируется).

Другим возможным вариантом анализа было бы рассмотрение долей сразу для всей конструкции, с максимальным количеством определённых связей, что сократило бы количество шаблонов и было бы удобнее в применении. Однако данный метод скорее всего будет проигрывать предложенному выше по полноте, так как пересечение множеств меньше объединения этих же множеств. Тем не

менее для точного ответа на данный вопрос требуется провести дополнительные исследования.

2.3. Отбор и суммирование конструкций

Финальный этап состоял в отборе конструкций с долей подчинения внутри шаблона не менее 97%. Иными словами, мы хотели отобрать относительно частотные конструкции, используя которые мы будем определять связь правильно с вероятностью не менее 0.97.

Поскольку на первом этапе для повышения полноты метода мы извлекли n, n+1, …, n+m-граммы, то требуется решить проблему дублей в более крупных конструкциях. Для этого применим рекурсивный подход: при анализе n-грамм удаляем из них такие, которые являются расширением контекста справа и слева для уже рассмотренных конструкций на уровнях n-1, n-2, …, n-k-грамм, где n-k — минимальная ширина уже рассмотренных шаблонов (считаем, что этот контекст избыточен). Таким образом, список проанализированных конструкций обновляется до ширины n-грамм.

Далее в рамках оставшихся n-грамм расширяем до ширины n такие n-1-граммы, что не прошли установленный порог по доле на уровне n-1, путём последовательного добавления элемента справа и слева. Среди расширенных контекстов снова отбираем те конструкции, относительная доля которых > 0.97, добавляем их к списку рассмотренных и исключаем из анализируемых n-грамм.

Наконец, из оставшихся n-грамм, не являющихся расширением контекста для каких-либо n-1-грамм, отбираем конструкции с долей не менее 0.97. Все “удачные” n-граммы последовательно записываются в отдельный файл вместе с числом их вхождений. Базисом (условием выхода) рекурсии для нас служат n-k-граммы. На их уровне мы сразу отбираем в число “хороших” конструкций те, что превышают порог по доле. Эти же конструкции будут первыми в списке уже проанализированных.

После того, как мы получили списки конструкций, удовлетворяющих требованию по доле подчинения внутри шаблона, можно вернуться к проблеме, затронутой в конце раздела 2.2: на практике нам бы хотелось пользоваться конструкциями, где сразу задано максимальное количество “хороших” связей. Для

этого мы группируем каждый список по частям речи, то есть по POS1, …, POSn. Важно отметить, что число вхождений подконструкций (каждая связь в отдельности) может различаться. Однако эта разница не может превышать 3% от общего числа конструкции в шаблоне, что является для нас приемлемым. Тем не менее в финальном перечне мы будем давать нижнюю оценку числа вхождений конструкций. В результате получаем максимально обобщенный для нашего метода список конструкций, в которых связи между словами могут быть определены лишь за счёт информации о порядке и частеречной принадлежности единиц. При последующем применении извлеченных конструкций мы сможем также задавать порог минимального числа вхождений.

3. Реализация алгоритма

3.1. Общие параметры входных и выходных данных

Как уже было сказано в разделе 1.4. мы используем файлы SynTagRus в формате CoNLL-U. Из них извлекаются 3-, 4-, 5- и 6-граммы. При этом мы проводим два варианта отбора конструкций: один на всех доступных файлах из (GitHub SynTagRus), а другой — только на ru_syntagrus-ud-train.conllu и ru_syntagrus-ud-dev.conllu. Во втором варианте мы оставляем файл ru_syntagrus-ud-test.conllu для более точной оценки качества полученных конструкций, но вместе с тем уменьшаем размеры коллекции для расчёта статистики по сравнению с первым вариантом. Сравнение результатов двух вариантов анализа см. в разделах 4.1 и 4.2.

3.2. Использованные библиотеки и созданные функции

Реализация первого этапа метода заключается в обработке файла, записанного в формате CoNLL-U. Для этого мы создали функцию get_sents() с параметром test, отвечающим за включение (test=True, дефолтное значение) или исключение (test=False) файла ru_syntagrus-ud-test.conllu из коллекции документов, на которой производится расчёт статистики. В рамках этой же функции происходит добавление маркеров начала и конца предложения и извлечение необходимой информации по каждому элементу: часть речи (поле 4 в

разметке элемента) и номер родительской вершины (поле 7). Номер самого элемента кодируется автоматически с помощью индексации элементов списка-предложения. Мы не используем модуль (PyPI conllu), поскольку среди всего множества полей разметки нам нужны только два. Другая важная функция из первого этапа — get_n_grams(). Она извлекает n-граммы, преобразуя при этом номера родительских вершин относительно границ n-граммы.

На втором и третьем этапах мы используем библиотеку pandas для удобной визуализации и обработки табличных данных (Python pandas). Все группировки, сопоставления и удаления частей таблиц (dataframes), описанные в разделах 2.2 и 2.3, производятся с помощью методов данной библиотеки.

Стоит отметить, что особенно долго выполняется код второго этапа для n-грамм от 5 и больше. Это напрямую связано с ростом количества анализируемых конструкций при переходе к более широким n-граммам. К сожалению, в настоящий момент оптимизация данного этапа ещё не проведена.

4. Результаты

4.1. Результаты, полученные на всей коллекции документов

4.1.1. Расчёт и анализ статистики

Количественные результаты извлечённых на каждом из этапов n-грамм представлены в Таблице 1.

Таблица 1. Количество n-грамм на всей коллекции документов

	Ширина n-граммы				Всего
	3	4	5	6	
Всего конструкций (частеречные n-граммы с синтаксическими связями)	10 831	75 126	237 452	461 198	784 607
Всего шаблонов (частеречные n-граммы)	2 874	22 114	96 735	256 716	378 439
Итоговые конструкции (после обобщения, точность выше 97%)	422	7 912	51 441	140 177	199 952
более 50 вхождений (из итоговых)	109 (25.83%)	359 (4.54%)	448 (0.87%)	194 (0.14%)	1 110 (0.56%)

более 5 вхождений (из итоговых)	170 (40.28%)	1 758 (22.22%)	6 577 (12.79%)	9 470 (6.76%)	17 975 (8.99%)

Несмотря на то, что с увеличением ширины окна n, общее количество n-грамм уменьшается, число уникальных конструкций и шаблонов, очевидно, возрастает. При этом отметим, что обобщение конструкций до шаблонов также тем значительнее, чем больше n. После применения этапа 2.3 получаем конструкции, удовлетворяющие условию точности не менее 97%. Заметим, что в абсолютных числах тренд снова положительный: чем больше n, тем больше конструкций извлекается. Однако если обратить внимание на то, как итоговые конструкции распределены по частоте вхождений, то мы увидим, что в процентном отношении большую часть 6-грамм (93.24%) составляют крайне редкие конструкции, встретившиеся менее 5 раз во всём анализируемом корпусе, тогда как максимальную долю относительно частотных (более 50 вхождений) конструкций наблюдаем среди 3-грамм (25.83%). Иными словами, с увеличением n прослеживается тенденция к снижению процента частотных n-грамм и к увеличению доли наиболее редко встречающихся конструкций.

Единственный тренд, не являющийся монотонно убывающим или возрастающим, это абсолютное число относительно частотных n-грамм: максимум достигается не на 6-, а на 5-граммах, тогда как количество первых оказывается меньше даже 4-грамм. Тем не менее для удовлетворительного объяснения данного эффекта требуется провести дополнительное исследование.

По всем n-граммам вместе статистика является весьма скромной в относительных числах. Заметим, однако, что формальное число конструкций может быть увеличено засчёт соединения таких контекстов n-грамм, при которых связи определяются одинаково. В данной работе такое укрупнение не проводилось для более удобного применения конструкций для оценки качества их работы, поскольку количество определяемых связей останется таким же.

4.1.2. Анализ наиболее частотных конструкций

Рассмотрим 5 наиболее частотных по числу вхождений 3-грамм (см. Таблица 2 в Приложении). На первом месте находится конструкция *ADP, ADJ, 3, NOUN* — предложная группа, в составе которой содержится именная группа с

модификатором-прилагательным. Очевидно, что при такой последовательности слов, прилагательное с вероятностью близкой к единице относится к последующему существительному (3a). Интересно же то, что наш метод не определил связь первого предлога и существительного как надёжную. Причина состоит в том, что на более чем 3% случаев приходятся конструкции с составными предлогами (3b) или устоявшимися словосочетаниями. Видимо, по правилам разметки SynTagRus в таких конструкциях связь предлога определяется по-другому: в примере (3b) вершиной *на* является *несмотря*.

(3) ADP, ADJ, NOUN

a. <...> ***по глубокому снегу*** (конструкция *ADP, 3, ADJ, 3, NOUN*) [Мертвые сраму не имут]

b. *Несмотря* ***на свойственную возрасту*** *впечатлительность* <...>. (конструкция *ADP, _, ADJ, 3, NOUN*) [Ежедневная симфония]

С другой стороны, связь предлога и существительного в предложной группе определяется с достаточной точностью, если после стоит знак препинания (<...> *например,* ***из глубины,*** <...> [Игрушки богов]). Конструкция *ADP, 2, NOUN, PUNCT* заняла третье место в топе.

На втором и четвёртом местах, как и на первом, оказались конструкции, определяющие связь прилагательного и существительного при наличии левого контекста — существительного (4) или глагола (5) соответственно. Конструкция (4) также является хорошей иллюстрацией того факта, что для русского языка, характеризующегося в основном правым ветвлением (наличие предлогов, основной порядок VO), тем не менее характерна препозиция прилагательного, то есть, находясь между двумя существительными, оно будет с намного большей вероятностью относится к последнему.

(4) NOUN, ADJ, 3, NOUN

<...> *при* ***помощи специального канала*** <...>. [Канал Discovery строит планы… на ваш туалет]

(5) VERB, ADJ, 3, NOUN

<...> ***говорит организатор конференции*** <...>. [Кому достанется Чингисхан]

Последняя конструкция в пятёрке лидеров также устанавливает связь существительного и его модификатора, в данном случае — детерминатора (6).

(6) DET, 2, NOUN, PUNCT

<...> *ссылаясь на* ***свою некомпетентность.*** [Кому достанется Чингисхан]

Перейдём к рассмотрению 6-грамм (Таблица 3 в Приложении). Первые три конструкции из топ-5 не представляют особого лингвистического интереса, так как указывают на вершину пунктуационного знака.

6-грамма с 4 места выделяет предложную группу, но с помощью довольно широкого контекста (7). Возможно, что такой контекст является избыточным. Однако нашим методом и на заданном уровне точности в 0.97 такая конструкция может быть получена только на этапе 6-грамм.

(7) NOUN, PUNCT, PRON, VERB, ADP, 6, NOUN

Проблема, которая встает перед дизайнерами <...> [Нанотехнологии вдохнули новую жизнь в бионику]

Последняя из самых частотных 6-грамм также определяет связь существительного с предлогом, но при этом не непосредственно примыкающих друг к другу, а разделённых двумя прилагательными (8). На данном примере также хорошо видна особенность наше метода: в рамках данной конструкции могут быть определены ещё как минимум две связи — обоих прилагательных с существительным. Однако эти связи уже были определены более узкими n-граммами, поэтому здесь игнорируются.

(8) NOUN, ADP, 5, ADJ, ADJ, NOUN, PUNCT

<...> *для* ***перехода к современной капиталистической экономике,*** <...> [Превратности развития капитализма в России]

Таким образом, наиболее частотными оказались n-граммы, связанные с определением связи прилагательного или определителя с существительным в именной и предложной группах. Подобные конструкции описаны, например, в (Кобзарева 2007), и их корректность с заданным уровнем точности не вызывает сомнений.

4.2. Результаты, полученные без test

4.2.1 Расчёт и анализ статистики

С уменьшением объёма анализируемого материала общее число конструкций и шаблонов, очевидно, уменьшилось (см. Таблица 5). Неожиданным стало то, что число итоговых конструкций, встретившихся более 50 раз, увеличилось для 3- и 4-грамм. Для 3-грамм также уменьшилась доля конструкций, встретившихся менее пяти раз, то есть 3-граммы стали более качественными. Однако общие показатели изменились незначительно, а число редких конструкций в целом даже увеличилось. В связи с этим, оснований для уменьшения количества анализируемых документов у нас нет.

Таблица 4. Количество n-грамм на коллекции без test

	Ширина n-граммы				Всего
	3	4	5	6	
Всего конструкций (частеречные n-граммы с синтаксическими связями)	10 557	71 592	221 790	423 614	727 553
Всего шаблонов (частеречные n-граммы)	2 844	21 544	92 321	240 263	356 972
Итоговые конструкции (после обобщения, точность выше 97%)	428	7 802	49 336	130 541	188 107
более 50 вхождений (из итоговых)	111 (25.93%)	365 (4.68%)	425 (0.86%)	155 (0.12%)	1 056 (0.56%)
более 5 вхождений (из итоговых)	176 (41.12%)	1 714 (21.97%)	6 122 (12.41%)	8 429 (6.46%)	16 441 (8.74%)

4.2.2. Анализ наиболее частотных конструкций

Большинство 3-грамм из топ-5 по частотности, выделенных без использования документа ru_syntagrus-ud-test.conllu (Таблица 5 в Приложении), совпадают с рассмотренными выше в разделе 4.1.2. Различия касаются конструкции *ADP, 2, NOUN, PUNCT*, которая в данном случае не смогла пройти порог по доле в шаблоне (0.968). За счёт этого на пятую строчку поднялась ещё одна конструкция, описывающая именную группу, которая в предыдущем варианте анализе была на шестом месте — *ADJ, ADJ, 3, NOUN* (9).

(9) ADJ, ADJ, 3, NOUN

<...> *до сооружения* ***парижского инженерного чуда*** <...>. [Нанотехнологии вдохнули новую жизнь в бионику]

Что касается списка 6-грамм (Таблица 6 в приложении), то от топ-5 в разделе 4.1.2 они отличаются только на 3 и 4 позициях — эти конструкции поменялись местами.

Суммируя, можно сделать вывод, что в целом набор наиболее “удачных” конструкций не сильно зависит от состава коллекции документов. В обоих случаях получаем среди самых частотных конструкции, связанные с именной и предложной группами. Однако при прочих равных рекомендуется использовать как можно больший объём данных, чтобы повысить статистическую мотивированность результатов.

4.3. Оценка точности и полноты и сравнение с результатами spaCy v3.0

В качестве оценки рассмотрим вариант анализа, основанный на всей коллекции документов. Суммарно удачные 3-граммы встретились в тексте более 146 000 раз, 4-граммы -- более 153 000. Это означает, что вместе они покрывают около 27.2% синтаксических связей корпуса (добавление маркеров начала и конца предложения увеличивает покрытие на 4.7%). Заметим, что подобная оценка покрытия является оценкой сверху, так как 3- и 4-граммы могут накладываться друг на друга. Если считать все 3-граммы, которые имеют два общих слова, за совпадения, необходимо сократить нижнюю границу примерно на 16 000 вхождений. Для 4-грамм снижение покрытия должно составить около 50 000 вхождений. Таким образом, нижняя граница полноты может быть оценена в 21.7%. Для точности нижней оценкой является наш порог на долю 97%.

Таким образом, с точки зрения полноты наш метод несомненно хуже с точки (spaCy v3.0), но по точности его вполне можно сравнить с современными нейросетевыми инструментами синтаксического анализа (см. Таблица 7).

Таблица 7. Сравнение с полным синтаксическим анализом

	Вся коллекция документов	spaCy (spaCy v3.0)
Precision	> 97%	96%

Recall	> 21.7%	100%

5. Заключение

Как было показано выше, более 21% синтаксических связей в русском тексте может быть восстановлено с использованием только информации о частях речи слов. При этом мы соблюдаем начальное условие на то, что конструкции должны быть корректными для 97% примеров, в которых они встречаются (с точностью до корректности разметки корпуса). Расширение до грамматических параметров должно ещё повысить долю правильных ответов, например, при соединении прилагательных и существительных, но как показывают эксперименты, не является обязательным.

Однако заметим, что точность работы метода зависит от точности предшествующего ему этапа снятия омонимии. Несмотря на то, что метод не чувствителен к неоднозначности леммы или грамматических параметров, он не сможет давать точные ответы без хорошего разрешения частеречной неоднозначности.

Полученные результаты несомненно представляют интерес для теоретической лингвистики, так как предлагают готовый материал для анализа, а также могут служить дополнительным аргументом к вопросу о том, насколько в русском языке свободный порядок слов. Помимо этого, в будущем у данного метода предполагается типологическое расширение, то есть его применение для извлечения синтаксически однозначных конструкций из других языков, обладающих достаточно большим синтаксически размеченным корпусом.

В области компьютерной лингвистики полученные конструкции могут использоваться при извлечении синтаксически связанных словосочетаний (например, (Klyshinsky et al. 2018)), а также для разработки методов поверхностного синтаксического анализа, или как вспомогательный инструмент при проведении полного синтаксического анализа, помогающий уменьшить “комбинаторный взрыв” при разборе предложения (Добров 2017). Метод выгодно отличается от существующих тем, что позволяет автоматически извлечь всех кандидатов, избегая ручного поиска.

Доклад по данному исследованию занял 2-е место на студенческой научно-практической конференции ФКН CoCoS'2021, а также принят к участию на международной научной конференции «Корпусная лингвистика 2021».

6. Список литературы

Большакова и др. 2007 — Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Е., Морозов С.С. *Лексико-синтаксические шаблоны в задачах автоматической обработки текста* // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2007». 2007. С. 70–75.

Большакова и др. 2017 — Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. *Автоматическая обработка текстов на естественном языке и анализ данных*: учеб. пособие. Москва: Изд-во НИУ ВШЭ, 2017. 269 с.

Добров 2017 — Добров А. В. *Глава 2. Компьютерный Синтаксис* // Прикладная и компьютерная лингвистика. Москва: Ленанд, 2017. С. 35–58.

Кобзарева 2007 — Кобзарева Т. Ю. *Некоторые свойства линейной структуры именных и предложных групп* // Вестник РГГУ. 2007. № 8. С. 113–130.

Кудинов 2013 — Кудинов М. С. *Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей* // Машинное обучение и анализ данных. 2013. Т. 1. № 6. С. 714–724.

Ножов 2003 — Ножов И. М. *Реализация автоматической синтаксической сегментации русского предложения*: дис. ... канд. техн. наук, Москва. 2003. 148 с.

Смирнов, Шелманов 2013 — Смирнов И. В., Шелманов А. О. *Семантико-синтаксический анализ естественных языков Часть I. Обзор методов синтаксического и семантического анализа текстов* // Искусственный интеллект и принятие решений. 2013. № 1. С. 41–54.

Abney 1992 — Abney S. P. *Parsing By Chunks* // Principle-Based Parsing: Computation and Psycholinguistics Studies in Linguistics and Philosophy. / ed. R. C. Berwick, S. P. Abney, C. Tenny. Dordrecht: Springer Netherlands, 1992. Pp. 257–278.

Cowie & Lehnert 1996 — Cowie J., Lehnert W. *Information extraction* // Communications of the ACM. 1996. T. 39. № 1. Pp. 80–91.

Feldman & Sanger 2006 — Feldman R., Sanger J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press, 2006.

Klyshinsky et al. 2018 — Klyshinsky E. S., Lukashevich N. Y., Kobozeva I. M. *Creating a Corpus of syntactic co-occurrences for Russian* // In Proc. of “Dialog-2018”. 2018. Pp. 317–331.

Molina & Pla 2006 — Molina A., Pla F. *Shallow Parsing using Specialized HMMs* // The Journal of Machine Learning Research. 2002. № 2. Pp. 595–613.

CoNLL-U — *CoNLL-U Format* [Электронный ресурс]. URL: https://universaldependencies.org/format.html (дата обращения: 20.05.2021).

GitHub SynTagRus — *GitHub: UD_Russian-SynTagRus* [Электронный ресурс]. URL: https://github.com/UniversalDependencies/UD_Russian-SynTagRus (дата обращения: 17.05.2021).

PyPI conllu — *PyPI conllu* [Электронный ресурс]. URL: https://pypi.org/project/conllu/ (дата обращения: 20.05.2021).

Python pandas — *Python Data Analysis Library: pandas* [Электронный ресурс]. URL: https://pypi.org/project/conllu/ (дата обращения: 20.05.2021).

spaCy v3.0 — *spaCy v3.0: Russian* [Электронный ресурс]. URL: https://spacy.io/models/ru (дата обращения: 20.05.2021).

UDPipe 2.0 — *UDPipe 2.0* [Электронный ресурс]. URL: https://ufal.mff.cuni.cz/udpipe/2 (дата обращения: 20.05.2021).

UD Russian — *Universal Dependencies: Comparison of Treebank Statistics for Russian* [Электронный ресурс]. URL: https://universaldependencies.org/treebanks/ru-comparison.html (дата обращения: 17.05.2021).

7. Приложение

Ссылка на исходный код программы: https://github.com/ancheveleva/grampatterns

Таблицы из текста работы:

Таблица 1. Количество n-грамм на всей коллекции документов

	Ширина n-граммы				Всего
	3	4	5	6	
Всего конструкций (частеречные n-граммы с синтаксическими связями)	10 831	75 126	237 452	461 198	784 607
Всего шаблонов (частеречные n-граммы)	2 874	22 114	96 735	256 716	378 439
Итоговые конструкции (после обобщения, точность выше 97%)	422	7 912	51 441	140 177	199 952
более 50 вхождений (из итоговых)	109 (25.83%)	359 (4.54%)	448 (0.87%)	194 (0.14%)	1 110 (0.56%)
более 5 вхождений (из итоговых)	170 (40.28%)	1 758 (22.22%)	6 577 (12.79%)	9 470 (6.76%)	17 975 (8.99%)

Таблица 2. Топ-5 самых частотных 3-грамм на всей коллекции документов

POS1	**#1**	**POS2**	**#2**	**POS3**	**#3**	**total_entries**
ADP	0	ADJ	3	NOUN	0	15168
NOUN	0	ADJ	3	NOUN	0	14308
ADP	2	NOUN	0	PUNCT	0	13452
VERB	0	ADJ	3	NOUN	0	9059

DET	2	NOUN	0	PUNCT	0	7049

Таблица 3. Топ-5 самых частотных 6-грамм на всей коллекции документов

POS1	**#1**	**POS2**	**#2**	**POS3**	**#3**	**POS4**	**#4**	**POS5**	**#5**	**POS6**	**#6**	**total_ entries**
NOUN	0	PUNCT	0	NOUN	0	PUNCT	5	NOUN	0	PUNCT	0	944
PUNCT	0	NOUN	0	PUNCT	4	NOUN	0	PUNCT	6	NOUN	0	495
PROPN	0	PUNCT	0	PROPN	0	PUNCT	5	PROPN	0	PUNCT	0	292
NOUN	0	PUNCT	0	PRON	0	VERB	0	ADP	6	NOUN	0	287
NOUN	0	ADP	5	ADJ	0	ADJ	0	NOUN	0	PUNCT	0	274

Таблица 4. Количество n-грамм на коллекции без test

	Ширина n-граммы				Всего
	3	4	5	6	
Всего конструкций (частеречные n-граммы с синтаксическими связями)	10 557	71 592	221 790	423 614	727 553
Всего шаблонов (частеречные n-граммы)	2 844	21 544	92 321	240 263	356 972
Итоговые конструкции (после обобщения, точность выше 97%)	428	7 802	49 336	130 541	188 107
более 50 вхождений (из итоговых)	111 (25.93%)	365 (4.68%)	425 (0.86%)	155 (0.12%)	1 056 (0.56%)
более 5 вхождений (из итоговых)	176 (41.12%)	1 714 (21.97%)	6 122 (12.41%)	8 429 (6.46%)	16 441 (8.74%)

Таблица 5. Топ-5 самых частотных 3-грамм на коллекции без test

POS1	**#1**	**POS2**	**#2**	**POS3**	**#3**	**total_entries**

ADP	0	ADJ	3	NOUN	0	13557
NOUN	0	ADJ	3	NOUN	0	12955
VERB	0	ADJ	3	NOUN	0	8129
DET	2	NOUN	0	PUNCT	0	6377
ADJ	0	ADJ	3	NOUN	0	6242

Таблица 6. Топ-5 самых частотных 6-грамм на коллекции без test

POS1	**#1**	**POS2**	**#2**	**POS3**	**#3**	**POS4**	**#4**	**POS5**	**#5**	**POS6**	**#6**	**total_ entries**
NOUN	0	PUNCT	0	NOUN	0	PUNCT	5	NOUN	0	PUNCT	0	878
PUNCT	0	NOUN	0	PUNCT	4	NOUN	0	PUNCT	6	NOUN	0	469
NOUN	0	PUNCT	0	PRON	0	VERB	0	ADP	6	NOUN	0	256
PROPN	0	PUNCT	0	PROPN	0	PUNCT	5	PROPN	0	PUNCT	0	252
NOUN	0	ADP	5	ADJ	0	ADJ	0	NOUN	0	PUNCT	0	249

Таблица 7. Сравнение с полным синтаксическим анализом

	Вся коллекция документов	spaCy (spaCy v3.0)
Precision	> 97%	96%
Recall	> 21.7%	100%