А.Н. Чевелева, Э.С. Клышинский
A.N. Cheveleva, E.S. Klyshinskiy

АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ КОНСТРУКЦИЙ ДЛЯ ПОВЕРХНОСТНОГО СИНТАКСИЧЕСКОГО АНАЛИЗА
EXTRACTION OF CONSTRUCTIONS FOR SHALLOW PARSING

Аннотация. В статье исследуется синтаксическая однозначность текстов русского языка на уровне n-грамм. На материале синтаксически размеченного корпуса СинТагРус мы показываем, что синтаксическая структура примерно 25% связей в тексте может быть восстановлена на уровне 3- и 4-грамм с долей верных порядка 97%. Наиболее частотными являются такие связи в именных и предложных группах.

Ключевые слова. поверхностный синтаксический анализ, автоматическое извлечение конструкций, русский язык

Abstract. In this article, we investigate the issue of syntactical unambiguity of Russian word n-grams. Using SynTagRus corpus, we demonstrate that about 25% of syntactic links in such a corpus can be obviously established using merely information of part-of-speech 3- and 4-grams with about 97% accuracy. The most frequent unambiguous connections here are connections in noun and prepositional phrases.

Keywords. shallow parsing, automatic extraction, the Russian language

1. Введение

Одним из вариантов проведения синтаксического анализа текста является предварительный поиск синтаксически связанных групп соседних слов (выделение синтаксических фрагментов текста, chunking) [Кудинов 2013]. В некоторых случаях, подобный

поиск может не принимать во внимание синтаксических связей между словами [Смирнов и др. 2013]. В случае, когда синтаксические связи не нужны, речь обычно идет о выделении терминов, описываемых определенными синтаксическими конструкциями (см., например, [Большакова и др. 2007]), именованных сущностей, извлечении фактов и решении похожих на них задач [Molina et al. 2002]. В случае определения синтаксических связей речь идет о поверхностном синтаксическом анализе. Среди прочего, проведение поверхностного синтаксического анализа позволяет ускорить работу полного синтаксического анализа, сократив количество анализируемых единиц и снизив синтаксическую неоднозначность текста. Подобный подход был популярен в начале 2000 годов (см. например, [Ножов 2003]), однако введение нейросетевых подходов отодвинуло его на второй план. Предложенные в рамках данного подхода методы всё ещё остаются актуальными, но для их практического применения должно быть соблюдено несколько условий: точность выделения фрагментов должна превышать точность работы существующих решений; применение результатов этапа выделения фрагментов должно повышать скорость работы системы.

Для проведения поверхностного синтаксического анализа могут использоваться различные методы. Так в [Molina et al. 2002] используются скрытые Марковские модели, в [Кудинов 2013] - условные случайные поля, в [Большакова и др. 2007] - контекстно-свободные грамматики, а в [Смирнов и др. 2013] упоминаются как конечные автоматы, так и различные методы машинного обучения. Упомянутые методы можно разделить на две группы: эмпирические методы и методы машинного обучения. В работе [Ножов 2003] для выделения именных групп используются

конечные автоматы, обладающие высокой скоростью работы. Однако создание подобных автоматов требует длительного ручного труда по поиску кандидатов в извлекаемые конструкции. Методы машинного обучения не всегда показывают сопоставимую точность и работают с меньшей скоростью, однако сами выделяют конструкции из размеченного корпуса, обеспечивая большую полноту.

Мы предлагаем метод автоматического выделения шаблонов для анализа фрагментов текста, обладающих синтаксически однозначной структурой, на основании статистической информации из синтаксически размеченных корпусов русского языка. Одной из задач являлось определение доли текстов, которые могут быть подвергнуты подобному анализу.

2. Метод выделения фрагментов с однозначными связями

На начальном этапе мы разбиваем синтаксически размеченный корпус на предложения и извлекаем из них частеречные 3- и 4-граммы с информацией о номере родительской вершины для каждого слова. При этом мы не рассматриваем опущенные слова (empty nodes), а также исключаем из числа n-грамм иностранные слова, так как они не несут практической пользы. 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, …, POSn, #hostn, total_entries*, где *POSi, #hosti* – часть речи и номер родительской вершины i-го слова n-граммы, a total_entries -- число вхождений каждой конструкции в корпус. Синтаксические связи рассматриваются как направленные не помеченные.

Очевидно, что даже в частотных конструкциях внутренние связи при поверхностном синтаксическом анализе могут определяться неоднозначно. Например, для частотной конструкции существительное, прилагательное, существительное синтаксические связи не определяются однозначно: последнее существительное может как относиться к первому существительному (*обострение холодной войны*), так и иметь родительскую вершину снаружи (*(не представляет для) власти особой проблемы*). С другой стороны, неверно утверждать, что среди менее частотных конструкций не найдется удачных: так, связь частицы и наречия в шаблоне *ADP, PART, ADV* (*В не менее (правильных советских фильмах)*) определяется однозначно.

Для решения этой проблемы мы продублируем n-грамму несколько раз, оставив в каждой только одну связь, и сгруппируем строки по списку частей речи. Для каждой такой n-граммы с одной связью рассчитывается её суммарное число вхождений. После этого, во всех общих конструкциях рассматриваем все слова в отдельности: вычисляются количество конструкций с данным подчинением и доля этого подчинения среди общего числа подобных n-грамм.

Финальный этап состоял в отборе относительно частотных конструкции (50 и более вхождений), используя которые мы будем определять связь правильно с вероятностью не менее 0,97. Уровень 50 вхождений выбран исходя из эмпирических соображений, точность 0.97 выбрана с тем, чтобы конкурировать с современными синтаксическими анализаторами, UAS которых достигает 0.96 (см., например, https://spacy.io/models/ru). В список 3-грамм, подходящих для поверхностного синтаксического анализа, отбираются те, что подходят по данным критериям. Далее

из списка 4-грамм удаляются те, что являются расширением контекста справа и слева для уже отобранных 3-грамм (считаем, что этот контекст избыточен). Затем каждая из 3-грамм, не прошедших порог, дополняется контекстом справа. Получившиеся 4-граммы с точностью ниже 0,97 удаляются. Из оставшихся к удачным 3-граммам добавляются те 4-граммы, которые встретились 50 и более раз. Аналогичная процедура расширения проводится для левого контекста. Наконец, из оставшихся 4-грамм, не являющихся расширением контекста для каких-либо 3-грамм, отбираем прошедшие установленный порог по доле и числу вхождений.

3. Анализ результатов

Эксперименты проводились с синтаксически размеченным корпусом SynTagRus в формате деревьев зависимостей Universal Dependencies (https://universaldependencies.org/) с добавлением маркеров начала и конца предложений. Размер корпуса превышает 1.1 миллиона словоупотреблений. После начального этапа мы получили 10 821 3-граммы и 75 126 4-граммы. После обобщений и отбора было выделено 109 3-граммы и 347 4-граммы (из них 7 и 45, соответственно, содержали маркеры начала и конца предложения). 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 в начале и в конце, за совпадения, необходимо сократить нижнюю границу примерно на 16 000 вхождений. Для 4-грамм снижение покрытия может составить

около 50 000 вхождений. Таким образом, нижняя граница покрытия может быть оценена в 21.7%.

Наиболее частотной оказали 3-граммы, связанные с определением связи прилагательного или определителя с существительным в именной и предложных группах: прилагательное/определитель и существительное, перед которыми находится предлог, ещё одно существительное, глагол, ещё одно прилагательное, определитель или число (например, *в последнее время, (при) помощи специального канала, говорит организатор конференции*). Подобные конструкции описаны, например, в [Кобзарева 2007], и их корректность с заданным уровнем точности не вызывает сомнений. Ещё одной частотной конструкцией стало подчинение прилагательного существительному, стоящему перед союзом (*молодой человек* и). Предлог подчиняется стоящему после него существительному, личному местоимению или имени собственному, если после них идут глагол, наречие, союз, знак препинания и некоторые другие части речи (*через месяц после, в комнату вошел, на улицу и*).

Среди прочих, были выделены такие редкие конструкции, как вспомогательный глагол, подчиняющийся прилагательному (*трудно было понять*), или союз, подчиняющийся глаголу через другой глагол (*и стал смотреть*).

Среди 4-грамм самой частотной оказалась зависимость предлога от существительного через прилагательное, когда перед ними стоят ещё одно прилагательное, предлог или существительное, глагол, некоторые другие части речи (*знакомой до последней трещинки, сказку в новый год*), или когда после них находятся союз, частица, знак препинания или другие части речи

(*по ценным бумагам и*). Из низкочастотных конструкций неожиданно была извлечена конструкция глагол, прилагательное, существительное, существительное, в которой последнее существительное зависит от первого (*измерять абсолютную сложность заданий*).

Наш метод не определил связь первого предлога и существительного как надёжную, так как более чем 3% случаев приходятся на конструкции с составными предлогами или устоявшимися словосочетаниями: *по глубокому снегу*, но (*Несмотря) на свойственную возрасту (впечатлительность)*.

Исходные коды программного обеспечения для получения статистики, а также результаты расчетов приведены в репозитории https://github.com/ancheveleva/grampatterns/ .

4. Обсуждение результатов

Как было показано выше, от 21.7% до 27.2% синтаксических связей в русском тексте может быть восстановлено с использованием только информации о части речи слов. Заметим, что покрытие может быть увеличено за счет снижения требований к частоте встречаемости кнструкций (их частоты распределены по закону Ципфа и имеют характерный «тяжелый хвост») или добавления 5- и 6-грамм. При этом мы соблюдаем начальное условие по которому конструкции должны быть корректными для 97% примеров, в которых они встречаются (с точностью до корректности разметки корпуса). Расширение до грамматических параметров должно повысить долю правильных ответов, например, при соединении прилагательных и существительных, но как показали наши эксперименты, не является обязательным. Заметим, что точность работы метода зависит от точности

предшествующего ему этапа снятия омонимии. Метод не сможет давать точные ответы без хорошего разрешения частеречной неоднозначности, но не так чувствителен к неоднозначности леммы или грамматических параметров.

Представленные результаты могут использоваться для разработки методов поверхностного синтаксического анализа или извлечения синтаксически связанных словосочетаний (например, [Klyshinsky 2018]). Метод выгодно отличается от существующих тем, что позволяет автоматически извлечь всех кандидатов, избегая ручного поиска.

Литература

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

2. *Кобзарева Т.Ю.* (2007), Некоторые свойства линейной структуры именных и предложных групп. Вестник РГГУ. Серия “Языкознание”. №8, с. 113-130.

3. *Кудинов М.С.* (2013) Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей. Машинное обучение и анализ данных, т. 1, № 6, с. 714-724.

4. *Смирнов И.В., Шелманов А.О.* (2013), Семантико-синтаксический анализ естественных языков Часть I. Обзор методов синтаксического и семантического анализа текстов. Искусственный интеллект и принятие решений, 1, с. 41-54.

5. *Ножов И.М.* (2003), Реализация автоматической синтаксической сегментации русского предложения. Диссертация на соискание степени кандидата наук, М.

6. *Klyshinsky E.S., Lukashevich N.Y., Kobozeva I.M.* (2018) Creating a Corpus of Syntactic Co-occurrences for Russian. In Proc. of "Dialog-2018", pp. 317-331.

7. *Molina A., Pla F.* (2002), Shallow Parsing using Specialized HMMs, Journal of Machine Learning Research, 2, pp. 595–613.

References

1. *Bolshakova E.I., Baeva N.V., Bordachenkova E.A., Vasil'eva N.E., Morozov S.S.* (2007), Leksiko-sintaksicheskie shablony v zadachah avtomaticheskoj obrabotki teksta [Lexico-Syntactic Patterns for Automatic Text Processing]. Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2007». Moscow, 2007.

2. *Kobzareva T.Yu.* (2007), Nekotorye svoystva lineynoy struktury imennykh i oredlozhnykh grupp [Some Properties of Linear Structure of Noun and Prepositional Phrases] RSUH Bulletin. Linguistics. №8, pp. 113-130.

3. *Kudinov M. S.* (2013) Shallow Parsing of Russian Text with Conditional Random Fields. Machine Learning and Data Analysis, vol. 1, № 6, pp. 714-724.

4. *Smirnov I.V., Shelmanov A.O.* (2013), Semantiko-sintaksicheskiy analiz estestvennykh yazykov [Semantic-syntactical analysis of Natural Languages], #1, p. 41-54.

5. *Nozhov I.M.* (2003), Realizatsiya avtomaticheskoy sintaksicheskoy segmentatsii pusskogo predlozheniya [Implementation of Automatical Syntactical Segmentation of a Russian Sentence]. PhD thesis, Moscow.

6. *Klyshinsky E.S., Lukashevich N.Y., Kobozeva I.M.* (2018) Creating a Corpus of Syntactic Co-occurrences for Russian. In Proc. of "Dialog-2018", pp. 317-331.

7. *Molina A., Pla F.* (2002), Shallow Parsing using Specialized HMMs, Journal of Machine Learning Research, 2, pp. 595–613.

Чевелева Анастасия Николаевна
Национальный исследовательский университет «Высшая школа экономики» (Россия).
Cheveleva Anastasia Nikolaevna
National research university "Higher School of Economics" (Russia).
E-mail: ancheveleva@edu.hse.ru

Клышинский Эдуард Станиславович
Национальный исследовательский университет «Высшая школа экономики» (Россия).
Klyshinskiy Eduard Stanislavovich
National research university "Higher School of Economics" (Russia).
E-mail: eklyshinsky@hse.ru