Федюкова Екатерина Алексеевна

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
katefedyukova@gmail.com
Научный руководитель — А. В. Добров, канд. филол. наук

КОМПЬЮТЕРНАЯ ГРАММАТИКА НЕПОСРЕДСТВЕННЫХ СОСТАВЛЯЮЩИХ И ЗАВИСИМОСТЕЙ ДЛЯ АНАЛИЗА ПРЕДЛОЖЕНИЙ С НЕОДНОЗНАЧНЫМИ ОБСТОЯТЕЛЬСТВЕННЫМИ ДЕТЕРМИНАНТАМИ

Ключевые слова: синтаксическая неоднозначность, грамматика непосредственных составляющих, грамматика зависимостей, обстоятельственный детерминант.

Статья посвящена исследованию автоматического синтаксического разбора предложений с неоднозначными конструкциями на русском языке, а также созданию грамматики на платформе NLTK. Описано имеющееся программное обеспечение, алгоритм работы, и проведен анализ проблем и дальнейших перспектив.

Fedyukova Ekaterina

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

IMMEDIATE CONSTITUENT AND DEPENDENCY COMPUTATIONAL GRAMMAR FOR THE ANALYSIS OF SENTENCES WITH AMBIGUOUS ADVERBIAL DETERMINANTS

Keywords: syntactic ambiguity, immediate constituent grammar, dependency grammar, adverbial determinant.

The article addresses the study of an automatic syntactic analysis of sentences with am-biguous constructions in Russian, as well as the creation of a grammar on the NLTK platform. The available software and the algorithm of work are described, and an analysis of problems and further prospects is proposed as well.

Введение

В статье пойдет речь о синтаксическом разборе неоднозначных предложений с обстоятельственными детерминантами.

Метод объединения грамматик зависимостей и непосредственных составляющих решает проблему синтаксической неоднознач-

ности относительно тех предложений, неоднозначность в которых не различают упомянутые методы, примененные независимо.

Были исследованы подходы к объединению грамматик зависимостей и непосредственных составляющих, а также произведена попытка объединения, для чего был создан корпус предложений с подходящими неоднозначными конструкциями. Для объединения была выбран формат грамматики, используемый в проекте AIIRE (Artificial Intelligence Information Retrieval Engine).

Детерминанты и обстоятельственные детерминанты

Для работы были выбраны предложения с неоднозначностью типа «обстоятельственный детерминант — обстоятельство. Выбор обусловлен тем, что для работы требовался такой тип неоднозначности, который не был бы отражен при разборе методами непосредственных составляющих и зависимостей по отдельности.

Согласно Н. Ю. Шведовой, «детерминант — это такой член предложения, который распространяет предложение в целом и формально не связан ни с каким определённым членом предложения» [Шведова, 1970, с. 624]. Он может присоединяться как к распространенному предложению, так и к нераспространённому.

Некоторые примеры предложений с обстоятельственными детерминантами: *Со стороны сада раздаются голоса. Начиная с субботы на стройку доставляются горячие обеды. Из-за отсутствия помещения не было возможности работать.* [Шведова, 1970, с. 625].

Синтаксическая неоднозначность

Синтаксическая неоднозначность (синтаксическая омонимия) — это ситуация, в которой можно построить более одного варианта синтаксической структуры на основании одной и той же последовательности минимальных единиц.

Согласно А. В. Гладкому, можно выделить следующие виды синтаксической неоднозначности:

1) разметочная — омонимия субъектной и объектной связей: *он платил за конферансье,*
2) стрелочная — множества синтаксических групп совпадают, но отношения подчинения различны: *он боялся с самого начала рассердить учителя,*
3) конституентная омонимия — множества синтаксических групп не совпадают: *я вижу только два дерева* [Гладкий, 1985, с. 111]

Различные типы неоднозначности могут комбинироваться между собой. Для обстоятельственных детерминантов характерная конституентная омонимия, соответственно, именно она актуальна для данной работы.

Структура непосредственных составляющих

Модель (система, структура) непосредственных составляющих была предложена лингвистом Л. Блумфилдом. Эта идея получила развитие в работах З. Харриса, Р. Уэллса, а позднее — у Н. Хомского.

В модели составляющих предложение (S) рассматривается как линейно упорядоченная цепочка минимальных единиц, в качестве которых могут выступать морфы, словоформы или даже единицы большие, чем одна словоформа [Буторов, с. 144].

В отличие от грамматики зависимостей, НС-грамматика в том виде, в котором она существует начиная с работ Хомского, не только отражает строение предложение, но и объясняет способ его порождения (деривацию). Грамматика задает список правил (правил переписывания), позволяющих преобразовывать вышеупомянутые абстракции (Предложение, Именная группа и т. д.) в более мелкие элементы (Именную группу можно преобразовать в два существительных).

Грамматики зависимостей

Основоположником метода грамматик зависимостей (ГЗ) считается французский лингвист Л. Теньер (1893–1954). Особенностью ГЗ является тот факт, что связь между словоформами в ней подчинительная (однонаправленная). Синтаксические связи устанавливают между словами отношение зависимости — одно из двух слов является главным, а другое — зависимым. Например, в словосочетании *большое красное яблоко* есть две связи: *яблоко* (какое?) *большое* и *яблоко* (какое?) *красное*, в обоих случаях *яблоко* — главное слово.

Согласно Я. Г. Тестельцу, «С помощью зависимостей легко анализируются непроективные структуры, а для того, чтобы отразить их в структурах составляющих, требуется усложнить формальный аппарат, например, введя два разных уровня представления предложения и особый грамматический компонент, устанавливающий соответствие между уровнями» [Тестелец, 2001, с. 146].

Объединение грамматик непосредственных составляющих и зависимостей

На материале русского языка самым подробным изложением такого подхода является теория синтаксических групп (ТСГ) Гладкого (см. [Гладкий, 1985, с. 34]). В рамках ТСГ, предложение — совокупность составляющих, связанных отношениями зависимости. Данная теория допускает установление зависимости не между отдельными словоформами, а между их совокупностями.

Для того, чтобы доказать наличие связей между составляющими, совокупность которых образует предложение, можно привести пример. У предложения *Из-за отсутствия помещения не было возможности работать* можно выделить два значения:

1) не было помещения, соответственно, работать было нельзя,
2) работать было невозможно только из-за отсутствия помещения, все остальные факторы этому не мешали.

В первом случае связь устанавливается между детерминантом *из-за отсутствия помещения и* всем предложением, а во втором — между обстоятельством *из-за отсутствия помещения и* глаголом *было.* Случай, когда связь устанавливается между детерминантом и предложением, будет отражен при разборе данного предложения синтаксическим парсером, использующим грамматику непосредственных составляющих. Второй же случай — при помощи парсера, использующего грамматику зависимостей, и связь тогда будет между «было» и «из-за». Очевидно, что при этом парсеры, использующие только какой-то один вид разбора, не отражают неоднозначность примера. И тогда одним из способов отразить неоднозначность при разборе является объединение грамматик.

Важно отметить, что, несмотря на наличие работ об объединении грамматик на английском языке (см, например, [Hays, 1960]), данная стратегия более актуальна для языков с менее жестким порядком слов, таких как арабский (см. [Skatov et al., 2013]), китайский (см. [Xiaona Ren et al., 2013]) , русский, о грамматиках для которого говорилось выше, и других.

Сбор материала

При помощи синтаксического подкорпуса НКРЯ был создан корпус предложений с неоднозначными обстоятельственными детерминантами объёмом 250 единиц. Все предложения можно клас-

сифицировать по критериям, которые использовались для поиска в корпусе. Некоторые примеры:

1) комбинации «предлог + существительное» в сочетании с «другой предлог + существительное; в данном примере — «на» + существительное + «в» + существительное. *На Съезде народных депутатов СССР в выступлении многих депутатов звучала озабоченность нерешенностью национальных проблем, накопившихся за долгие годы административно-командного правления, тревога за судьбы нашей федерации.*

2) Существительное + глагол, обстоятельственная связь. *Кораблями он ввозил в Россию краску индиго, промышлял также селитрой для пороха.*

При создании корпуса в предложениях были исправлены орфографические и пунктуационные ошибки.

Создание грамматики

Была создана НС-грамматика для данного корпуса предложений. Для проверки грамматики был выбран пакет NLTK для языка Python. На этом этапе было важно добиться максимальной полноты вариантов, которые предоставляет грамматика. При создании грамматики было решено допустить, что ветвление может быть только унарным и бинарным.

Синтаксические структуры в системе AIIRE — это размеченные структуры составляющих, причем разметка содержит в себе информацию о зависимостях (синтаксическом подчинении) и линейном порядке единиц.

В грамматике, реализованной в AIIRE, в отличие от NLTK, допускаются непроективные порядки слов, т. е. можно присоединить обстоятельство в начале предложения к глагольной группе, а не к вершине предложения.

Цель исследования подразумевает попытку перевода грамматики из формата NLTK в формат AIIRE, что налагает некоторые условия на НС-грамматику NLTK. Так, необходимо было решить, какая из составляющих в бинарном ветвлении всегда главная, а какая — всегда зависимая. Учитывая то, что, например, для именной группы единственного числа в родительном падеже нельзя решить, зависимая она или главная, было решено ввести условное обозначение для зависимой группы. Другим условием для трансформации грамматики была уникальность терминальных вершин — обозначе-

ния, используемые для терминальных вершин, не могут встречаться где-либо выше в дереве.

Заключение

В результате работы удалось:

1) создать корпус неоднозначных предложений на материалах синтаксического подкорпуса, а также основного корпуса НКРЯ,
2) создать НС-грамматику в формате NLTK для полученного корпуса,
3) предпринять успешную попытку перевести НС-грамматику в формате NLTK в формат AIIRE,
4) изучить источники на заданную тему.

В дальнейшем планируется успешно завершить перевод грамматики в формат AIIRE, рассмотреть возможности интегрирования других лингвистических модулей в программу и практического использования полученной программы.

ЛИТЕРАТУРА

Буторов, 1996 — *Буторов В. Д.* Моделирование синтаксиса естественного языка. СПб., 1996. С. 142–160.

Гладкий, 1985 — *Гладкий А. В.* Синтаксические структуры естественного языка в автоматизированных системах общения. М., 1985.

Тестелец, 2001 — *Тестелец Я. Г.* Введение в общий синтаксис. М.: РГГУ, 2001.

Шведова, 1970 — *Шведова Н. Ю.* Грамматика современного русского литературного языка. М.: Наука, 1970. С. 624–625.

Hays, 1960 — *Hays D.* Grouping and dependency theories [*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960].

Skatov et al., 2013 — *Skatov D.* Parsing Russian: a Hybrid Approach / Dan Skatov, Sergey Liverko, Vladimir Okatiev, Dmitry Strebkov // Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 34–42.

Xiaona et al., 2013 — *Xiaona Ren.* Combine Constituent and Dependency Parsing via Reranking / Xiaona Ren, Xiao Chen, Chunyu Kit // Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2013, pp. 2155–2161.