Автоматический синтаксический анализ русских текстов

Анастасия Леонтьева, Ильдар Кагиров

Санкт-Петербургский институт информатики и автоматизации РАН
{an_leo, kagirov}@iias.spb.su

Аннотация

В данной статье описывается структура разработанного модуля автоматического синтаксического анализа (МСА). Данный модуль был разработан на основе модуля морфологического анализа русского языка SMART [1]. МСА служит для анализа структуры простых предложений и выделения в них синтаксических групп. Планируется внедрение МСА в систему распознавания русской слитной речи для повышения качества распознавания.

1 Введение

МСА – это программа или часть программы, выполняющая синтаксический анализ. На сегодня создание автоматического МСА является одной из самых актуальных задач в компьютерной лингвистике, решение которой позволило бы достичь высокого уровня формализации языковых структур в разнообразных прикладных целях: от создания систем автоматического распознавания речи до поисковых систем в Интернет.

Под целью синтаксического анализа в настоящей статье понимается вычленение базовых синтаксических структур и установление синтаксических связей между ними. На выход МСА поступает цепочка слов, разбитая на группы, причем каждая группа имеет связь с другими группами. Кроме того, МСА идентифицирует такие синтаксические категории, как подлежащее и сказуемое.

Однако создание МСА для русского языка упирается в большое количество сложностей, связанных с недостаточно разработанной теоретической базой в общем и прикладном языкознании; структуры человеческого языка отличаются разнообразием и часто высоким уровнем сложности, предусмотреть который чрезвычайно тяжело. В связи с этим в настоящей статье предлагается структура МСА, работающего с простыми синтаксическими структурами; создание МСА, справляющегося с текстом на русском языке любой сложности, представляется на настоящем этапе невозможным.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

Предложенный нами модуль синтаксического анализа является частью анализатора текста русского языка SMART (на нынешний момент разрабатывается), ориентированного на грамматический разбор текстов на литературном русском языке. Наш МСА выгодно отличается от многих разработок в этой области своей ориентированностью на синтаксический уровень языка, а не исключительно на формальные (внешние) признаки словоформ. Иными словами, на выход МСА подается *синтаксическая структура*, а не цепочка словоформ с простейшей разметкой по морфологическому согласованию.

Например, в МСА, демонстрируемом на сайте http://www.aot.ru, реализован именно «согласовательный» подход, при котором МСА ищет морфологические характеристики, свидетельствующие о зависимости двух и более членов предложения, тогда как структуры типа «очень хорошо» или «в пальто» остаются неразобранными. К сожалению, такой подход на сегодняшний день является хоть и самым примитивным, но и довольно надежным – зачастую представить более сложные синтаксические зависимости в терминах грамматики составляющих оказывается проблематичным.

2 Синтаксический анализ предложения

Поскольку число предложений бесконечно, при синтаксическом разборе имеет смысл ориентироваться на более мелкие единицы – фразовые категории (ФК). ФК – это группа, в которой имеется одна вершина, а также может быть одно или несколько зависимых от этой вершины [2], [3]. Таким образом, алгоритм автоматического анализа сводится к вычленению ФК в составе предложения и поиску связей между ними.

Для разработки модуля автоматического синтаксического анализа был использован корпус текстов, состоящий из клауз с нераспространенной синтаксической структурой из [4]. Клаузы составлены в соответствии с нормами литературного русского языка. Этот корпус, безусловно, нуждается в расширении и усложнении, но на нынешнем этапе разработки модуля синтаксического анализа он отвечает основному поставленному требованию: идентификация отдельных ФК в структуре клаузы и определение связей между ними.

На основании анализа используемого корпуса были выделены пять основных синтаксических групп: именная группа (ИГ), глагольная группа (ГГ), группа прилагательного (ПГ), предложная группа (ПрГ), инфинитивная группа (ИнфГ). Для удобства за каждой группой был закреплен порядковый номер. Каждая синтаксическая группа имеет вершину, то есть слово, от которого зависят все остальные слова в группе. Вершиной ИГ является имя существительное или личное местоимение. Вершиной ГГ – личные формы глагола. Вершиной ПГ выступает краткое прилагательное. Вершиной ПрГ является предлог. Вершиной ИнфГ – инфинитив. Связи и соотношения слов внутри групп представлены в таблице 1.

Символ «*» означает, что элементы группы могут стоять также и в обратном порядке. При этом тип связи между ними сохраняется неизменный.

Следует отметить, что глагольные группы представлены двумя уровнями. Глагольная группа второго уровня ГГ" включает в себя, помимо глагольной группы первого уровня ГГ', другие элементы.

Таблица 1

Типы ФК, используемых в модуле синтаксического анализа

ИГ {Сущ}/{М}	ГГ' {ГГ"}	ПГ {КрПрил}	ПрГ {Пр}	ГГ" {Глаг}	ИнфГ {Инф}
–	ИГ* управление	/ИГ управление		–	
/{ИГ} управление	ИГ/ управление /ИГ управление			({AUX} → {КрПрил})	
{Прил}/ согласование	ПрГ* примыкание			{Нар}* примыкание	
/CON-{Сущ}	ИГ/ управление /ПрГ примыкание	{Нар}/ примыкание	/ИГ управление	{Нар}/ примыкание /{Нар} примыкание	ИГ* управление
	/ИГ управление //ПрГ) примыкание				
	/ИнфГ примыкание				

ИГ – вершиной является Имя существительное Сущ или местоимение М; ГГ" – вершиной является финитный глагол Глаг; ГГ' – вершиной является группа ГГ"; ПГ – вершиной является краткое прилагательное КрПрил или прилагательное П; ПрГ – вершиной является предлог Пр; ИнфГ – вершиной является инфинитив Инф; AUX – вспомогательный глагол.

Настоящая таблица представляет систематизацию ФК, встречающихся в 500 тестовых предложениях. В каждой ФК действуют подчинительные связи одного из трех типов; на уровне морфологии это находит отражение в том, что при согласовании зависимое слово (Прил) принимает те же показатели рода, числа и падежа, что и вершина (Сущ или М); при примыкании наблюдается простое синтаксическое соположение вершины и неизменяемого слова-зависимого без дополнительного маркирования на морфологическом уровне, а при управлении зависимое слово (Сущ или М) стоит в определенном косвенном падеже, причем выбор падежа определяется по словарю, в характеристиках слова-вершины. Для определения падежа, в котором стоит зависимое слово при подчинительной связи, используется словарь [5]. Предполагается со временем создать свой словарь, специально приспособленный для нужд автоматического синтаксического анализа.

3 Описание алгоритма автоматического анализа синтаксической структуры предложения

Работа анализатора основывается на базе данных по ФК. В тексте ищутся только такие ФК, которые внесены в базу. Предложением (в случае с нераспространенными предложениями это клауза) называется отрезок текста между двумя показателями конца предложения – точкой / восклицательным знаком / вопросительным знаком + пробел и точкой / восклицательным знаком / вопросительным знаком.

Анализ начинается с того, что модуль морфологического анализа определяет морфологические характеристики и частеречную принадлежность анализируемого слова. Дальше начинается формирование гипотез о текущей ФК. Любая ФК может быть представлена в виде: $ХГ = (x; Г)$, где $ХГ$ –

название ФГ, x – вершина, YG – зависимое. YG может принимать значения $YG = 0$, где 0 – это пустое множество:

$$XG = волк; \quad x = волк; \quad YG = 0.$$

Или $YG = y; ZG$ где ZG – ФК, идентичная по структуре YG.

Теоретически, разложение YG на составляющие может быть бесконечным: $_1$[кот, $_2$[который пугает и ловит синицу, $_3$[которая часто ворует пшеницу, $_4$[которая $_5$[в тёмном чулане]$_5$ хранится $_6$[в доме, $_7$[который построил (Джек)]$_1$]$_2$]$_3$]$_4$]$_5$]$_7$

На следующем шаге по базе ФК определяются типы ФК, в которые может входить анализируемая словоформа. Это решается за счет простого перебора НС, которые возможны в текущем контексте при данных морфологических характеристиках слова.

После формирования гипотезы относительно ФК анализатор переходит к поиску и анализу следующей словоформы; далее аналогичным методом ищутся вершины и зависимые (см. раздел 5). По сути, все предложение разлагается на две группы – группу подлежащего (ИГ) и группу сказуемого (ГГ), построенных по модели $XG = (x; YG)$. Причем, если словоформа *A* – существительное в именительном падеже, у него нет вершины, оно объявляется подлежащим, если словоформа *B* – глагол в личной форме, у него нет вершины, он объявляется сказуемым и согласуется в числе и лице с подлежащим.

Анализ идет до тех пор, пока все словоформы в предложении не будут связаны друг с другом.

4 Программная реализация МСА

Структурная схема анализатора представлена на рис. 1.

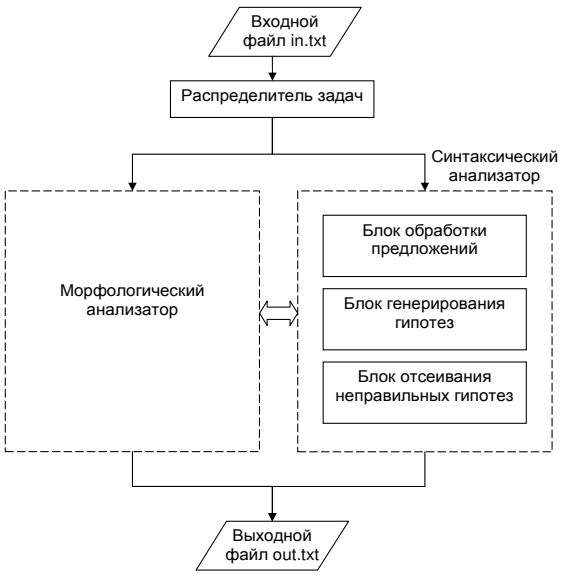

Рис. 1. Структурная схема модуля синтаксического анализа

Как было сказано выше, входные данные представляют собой список простых предложений. Распределитель задач передает эти данные в блок синтаксического анализа.

В блоке обработки предложения обрабатываются пословно. Исходная словоформа предается в блок морфологического анализа [6], в котором для нее подбираются все возможные варианты основ и соответствующие грамматические показатели. Каждое предложение считывается пословно, после чего словоформа поступает на вход морфоанализатора. В результате определяются все возможные основы и соответствующие грамматические показатели. Если данной словоформе соответствует только одна основа, она поступает в процедуру построения гипотез. В зависимости от части речи и грамматических показателей выделяется соответствующая синтаксическая группа. После предварительной обработки словоформа поступает в блок генерирования гипотез. Этот блок является основным. На его вход поступает словоформа. Если это первое слово в предложении, то в соответствии с частью речи определяется синтаксическая группа.

В зависимости от частей речи группа может определяться однозначно, либо могут быть варианты. Например, если на вход поступило имя существительное, то первая группа в предложении будет именной. В этом случае запоминается порядковый номер группы и обрабатывается следующее слово. Если же первым словом в предложении является наречие, то оно может относиться как к глагольной группе, так и к группе прилагательного. В этом случае для однозначного определения группы требуется следующее слово.

Начиная со второго слова в предложении, важную роль играет не только часть речи данной словоформы, но и информация о группе или группах, которые выделены на данный момент. Поступившая на вход словоформа может, как принадлежать текущей синтаксической группе, так и выделяться в другую синтаксическую группу. В этом случае формируется дополнительная гипотеза, и рассматриваются оба варианта. В конечном счете, в предложении выделяются группа подлежащего и группа сказуемого. Важно отметить, что некоторые группы, например, глагольная, могут содержать в себе другие группы. История выделения слов в группы сохраняется в виде индекса.

Таким образом, из-за морфологической и синтаксической неоднозначности для одного предложения может быть сформировано несколько гипотез, ср. [7], [8]. Далее эти гипотезы поступают в блок отсеивания неправильных гипотез. Данный блок имеет два уровня проверки. На первом уровне проверяется согласование синтаксических групп в рамках одного предложения. Это согласование определяется, исходя из грамматических характеристик вершин групп.

Выходной файл представляет собой список предложений, каждое из которых разбито на синтаксические группы. Если предложение содержит слово, отсутствующее в словаре, то оно выводится без разбора.

5 Заключение

Для тестирования разработанного модуля были выбраны случайным образом 50 простых предложений из ГОСТ Р 50860-95. Они были представлены в виде списка и обработаны. Результаты выглядят следующим образом: 32% предложений разобрано не было. При анализе результатов были выявлены следующие ошибки:

1. Наличие в анализируемом тексте синтаксических конструкций, отличных от исходных синтаксических групп – 26%.

2. Наличие в тексте слов, отсутствующих в словаре – 6%.

3. Семантическая (морфологическая) неоднозначность слов, порождающая несколько вариантов разбора предложения. Данный вид ошибки приводил к выводу неправильных гипотез, которые программа не смогла отсеять – 22%.

Для решения этих ошибок необходимо расширить базу синтаксических групп, за счет анализа большого количества текстов. Также планируется разработка модуля для автоматического пополнения словаря. И одной из важнейших задач является разработка частотного словаря, что позволит уменьшить количество гипотез для одного предложения и увеличить скорость и качество работы синтаксического анализатора.

В результате описанных в настоящей статье разработок был создан модуль синтаксического анализа текста русского языка. Синтаксический модуль позволяет производить разбор предложения. Следует заметить, что данная работа является только первым шагом при разработке полноценного синтаксического анализатора. В будущем планируется сделать более информативный вывод, представляя предложения в виде дерева. Таким образом, будут наглядно показаны синтаксические зависимости внутри предложений и указаны типы связи. Также планируется провести тестирование анализатора на произвольном тексте, взятом из художественной или научной литературы. На данном этапе разработки разбор производится только для простых предложений, так что планируется разработать и реализовать алгоритм разбора сложных предложений. Модуль синтаксического анализа может работать как самостоятельная программа, а может выступать в качестве синтаксической составляющей декодера для русской слитной речи. Поскольку словарь распознавателя содержит в себе отдельно основы и окончания, то в связи с вариативностью русского языка могут быть ошибки при подборе окончаний. Разрабатываемый синтаксический анализатор позволит устранить подобные ошибки, таким образом, повысив качество распознавания русской слитной речи.

Литература

[1] An. Leontieva, “The Module of Morphophonetic Word Processing for Composing a Vocabulary for Russian Continuous Speech Recognizer”. Scientific-theoretical journal «Artificial intelligence», Donetsk, Ukraine, Vol. 3, 2007, pp. 319–327.

[2] Фитиалов С.Я. Об эквивалентности грамматик НС и грамматик зависимости // Проблемы структурной лингвистики, 1967. М.: Наука, 1967. 145 с. С. 16–43.

[3] Тестелец Я.Г. Введение в общий синтаксис. М.: Российский государственный гуманитарный университет, 2001. 800 с.

[4] Передача речи по трактам радиотелефонной связи: Требования к разборчивости речи и методы артикуляционных изменений. ГОСТ 16600 – 72. М.: Издательство стандартов, 1973. 90 с.

[5] Большой толковый словарь русского языка / Под ред. Д.Н. Ушакова. М.: Альта-принт, 2005. 1239 с

[6] Kagirov I.A., Leontyeva An.B. Grammar-Based Speech- and Word-splitting // Proceedings of 3rd Language & Technology Conference. October 5–7, Poznac, Poland. Poznac: Fundacja Uniwersytetu im. A. Mickiewicza, 2007. 578 s. P. 413–417.

[7] Попов Э.В. Общение с ЭВМ на естественном языке. М.: Едиториал УРСС, 2004. 260 с.

[8] Mel'иuk I. A. Dependency syntax: Theory and practice. Albany, NY: SUNY Press, 1988. 428 p.

Automatic syntactic analysis of russian texts

Anastasia Leontyeva, Ildar Kagirov

The paper presents a syntactic approach for analysis of Russian texts. The approach was used to develop the module of the syntactic analysis, which could be used as a single program or as the auxiliary module for speech recognition. At the first stage of development the module analyses only the sentences with a siple structure. During the analysis the syntactic groups are singled out and dependencies between the constituents are determined.