УДК 004.912

ОБЗОР СИСТЕМ АНАЛИЗА ЕСТЕСТВЕННОГО ТЕКСТА НА РУССКОМ ЯЗЫКЕ

М. А. Артемов, А. Н. Владимиров, К. Е. Селезнев

Воронежский государственный университет

Поступила в редакцию 12.07.2013 г.

Аннотация. В данном обзоре представлены системы анализа текста, а также подходы к их построению, представленные в архиве конференции «Диалог-21» за 2006-2012 год, а также некоторые другие системы (ЭТАП-3, Abbyy Compreno, Томита-парсер, LINGVO-MASTER, Treeton, DictaScope Syntax, и пр.) Статья описывает основные особенности описанных систем, которые выделяют их среди остальных, а также акцентирует внимание на некоторых минусах представленных систем. В конце статьи дается общее заключение, в котором перечисляются основные недостатки, общие для большинства систем, а также формулируются требования к системе анализа текстов.

Ключевые слова: парсер, синтаксический анализ, анализ текста, синтаксическое дерево, естественный текст.

Annotation. This review introduce most successful systems for Russian language analysis and describes how they work. These systems were represented in Russian linguistics conference "Dialog-21" from 2006 to 2012 years(ЭТАП-3, Abbyy Compreno, Томита-парсер, LINGVO-MASTER, Treeton, DictaScope Syntax, etc.) The topic describes most important differences of each systems and shows disadvantages of it. Finally, authors shows some common disadvantages of these systems. In the end, some requirements for text analysis systems are given.

Keywords: parser, syntax analysis, text analysis, syntax tree, natural language.

Количество систем для анализа естественного текста неукоснительно растет, изобретаются новые алгоритмы, новые подходы, пересматриваются методики анализа и используемые структуры.

За последние годы можно выделить несколько систем, реально существующих и показывающих хорошие результаты при анализе русского языка.

Система LINGVO-MASTER .Данная система разработана специалистами ИПИ РАН и включает в себя лексикографический, морфологический, терминологический и синтактико-семантический анализаторы. На этапе морфологического анализа все слова приводятся к начальной форме. Терминологический анализатор выделяет термины из заранее определенного множества с учетом вариантов сокращений. Синтактико-семантический анализатор использует контекстные правила, изначально задаваемые в системе.

Система имеет в себе прямой лингвистический процессор ,позволяющий преобразовывать естественный текст в структуры знаний, а также обратный процессор, позволяющий на основе некоторых знаний строить предложения на естественном языке. Работа данной системы на примере анализа резюме рассмотрена в [1]. Судя по данной статье, система требует четкого определения возможных слов (терминов) и возможной структуры предложений (для конкретной предметной области), без определения которых система не имеет возможности работать.

Система TREETON. Данная система разрабатывается на факультете ВМиК МГУ и представляет собой исследовательскую систему для анализа естественных текстов [2,3]. Интерес представляет синтаксический анализатор, работающий в рамках данной системы.

Морфологический анализатор выдает в качестве выходных данных множество аннотаций для каждого из слов представленного текста. Данные множество поступает на вход синтаксического анализатора, который идет последовательно по словам, иногда откатываясь назад.

В рамках синтаксического анализатора авторы расширяют понятие «аннотация» до поня-

тия «тринотация», которая отличается от аннотации тем, что она дополнительно снабжена внутренней структурой — лесом с именованными связями, в узлах которых находятся другие тринотации. Таким образом формируется древовидная структура предложений.

Проблема экспоненциального роста времени обработки в данной системе решается на основе специальной системы штрафов, которые назначаются за следующие вещи:

1. Штрафы за количество синтаксических связей (штрафы на повторение).
2. Штрафы за непроективные предложения (штрафы на зацепление).
3. Штрафы на пропуски.
4. Штрафы на применение правил.

На основе данных штрафов строится общая оценка построенной синтаксической структуры, и на основе этого отбирается наиболее подходящий вариант с наименьшим количеством штрафов.

Очевидным минусом данной системы является то, что она требует четкой формализации большого множества правил языка. Применяемая система штрафов представляет большой интерес, однако возникает очевидная проблема определения размера штрафов для различных ситуаций и их балансировки. В силу большого количества различных связей и синтаксических правил, которые требуют определения в системе, балансировка может стать достаточно трудоемкой, если вообще разрешимой за конечное время задачей.

Среда из РГТУ. В рамках создаваемой в Институте лингвистики РГТУ объектной лингвистической среды[4], помимо стандартных модулей анализа текста, используется дополнительный модуль поверхностно-синтаксического, или сегментационного анализа.

Этап сегментационного анализа предшествует этапу синтаксического анализа и заключается в выделении частей предложений, называемых сегментами. Причем сегменты бывают 2 типов — α-сегменты, отвечающие за мелкие части предложений, вроде причастных и деепричастных оборотов, и β-сегменты, отвечающие за части сложноподчиненных и сложносочиненных предложений. Вложенность сегментов, в принципе, не ограничена. К каждому из сегментов применяется внутрисегментный синтаксический анализ, в результате которого строится синтаксический граф сегмента. И в конце происходит отдельный этап межсегментного анализа, который соединяет построенные синтаксические графы сегментов в один единый синтаксический граф.

Данный подход представляет интерес как способ сокращения сложности непосредственно синтаксического анализа, потому как анализ происходит не на уровне предложения, а на уровне сегмента. Однако на этапе межсегментного анализа возможны сложности с соединением построенных синтаксических графов. Ещё одним недостатком является то, что при неправильном выделении границ сегментов можно получить совершенно непредсказуемые результаты анализа (например при наличии некоторой авторской пунктуации).

DictaScope Syntax. Данная система разрабатывается специалистами компании «Диктум» и представляет собой достаточно мощный синтаксический анализатор [5]. Синтаксический анализ происходит на основе взвешенных графов, что позволяет свести задачу нахождения наиболее правдоподобной синтаксической интерпретации к решению задачи поиска минимального остовного дерева в построенном в процессе работы синтаксического анализатора.

Однако такой подход требует расстановки весов в дереве, что является затруднительной задачей для русского языка. Для других языков имеется достаточно большой языковой корпус, например для английского, или, что ближе, для чешского. Для русского же такого корпуса нет. Поэтому авторы данной системы вводят систему синтаксических правил. Каждое синтаксической правило содержит пару слов с указанным набором морфологических характеристик, возможные варианты окружающих их слов и собственно коэффициенты, которые применяются для вычисления весов. На 2012 год таких правил в системе было около 200.

Поиск проективного дерева осуществляется при помощи алгоритма Эйснера, за счёт чего сложность данного поиска — кубическая, то есть происходит за время O(n^3) от размера дерева, что является достаточно хорошим результатом среди множества парсеров.

Также в системе имеется возможность исправления опечаток, осуществляемая при нахождении неизвестного слова в тексте.

SemSin. Данный анализатор[10] использует в своей работе специальную морфологическую базу данных, построенную на основе сло-

варя Тузова, который, в свою очередь, основан на словаре Зализняка. В базе данных содержится информация о морфологических свойствах лексемы, номер класса и актанты вызываемых ею лексем. Морфологический анализатор, используя эту базу данных, на выходе выдает информацию в виде начальной формы слова, морфологических характеристик, а также класс с набором актантов. Помимо основной базы данных применяется база фразеологизмов и база предлогов.

Полученный в результате морфологического анализа набор токенов подвергается применению продукционных правил, в результате чего получается синтаксическое дерево. Количество правил — порядка 210.

ЭТАП-3. Данная система разработана в Институте проблем передачи информации РАН и представляет собой систему для анализа естественного текста на русском и английском языке[6]. Его работа осуществляется на основе нескольких сотен синтаксических правил, и специального комбинаторного словаря, на данный момент насчитывающего около 100000 слов.

Каждое правило записывается на специальном языке и состоит из 2 частей : последовательности групп предикатов и действия. Предикаты записываются в виде дизъюнктивных нормальных форм, элементарными предикатами которых являются различные морфологические и синтаксические характеристики слов. Применяются предикаты следующим образом — последовательность слов должна соответствовать всем предикатом из нечётных групп, и не соответствовать предикатам из чётных. Действие представляет собой создание потенциальной связи определенного типа между двумя словами.

Комбинаторный словарь представляет собой набор элементов, каждый из которых, помимо самого слова и его морфологических характеристик, содержит:

- дескриптор слова, или семантический признак слова,
- варианты согласования с другими словами с указанием параметров этих слов,
- синонимы,
- антонимы,
- однокоренные слова других частей речи,
- коллокации (слова, встречающиеся рядом).

После применения правил ко всем элементам предложения получается синтаксическое дерево, в котором достаточно много лишних связей. Процесс удаления лишних связей основан на применении 3 типов правил.

- Так называемые «интерсинтаксические» правила, позволяющие расставить некоторые абсолютные веса дугам синтаксического дерева.
- Правила выбора члена предложения, который является корневым элементом в синтаксическом дереве.
- Прочие правила, применяемые к синтаксическому дереву. Отличие этих правил от интерсинтаксических в том, что они применяются при помощи алгоритма с возвратом (бэктрекинг), то есть применяется последовательность правил, пока не произойдет блокировка, то есть пропадет связь некоторого элемента со всеми остальными.

Данный анализатор применяется в системе машинного перевода текста, в системе синтеза речи, а также для создания первого размеченного синтаксического корпуса русского языка (SynTagRus).

ABBYY Compreno. Система ABBYY Compreno разрабатывается компанией ABBYY в течении последних 15 лет[7], и представляет собой интеллектуальную систему анализа текста, построенную на основе универсального дерева понятий, или USH (Universal Semantic Hierarchy). Это дерево имеет корневыми элементами некоторые области человеческой жизни, которые впоследствии ветвятся на более конкретные области, а листьями в этом дереве являются непосредственно слова. Это очень полезная структура для системы машинного перевода, так как позволяет искать перевод слов, двигаясь по этому смысловому дереву, которое практически одинаково для любых языков.

Синтаксический анализ в данной системе выдает выходную информацию в виде синтаксического дерева, расширенного дополнительными связями[8]. Связи дерева показывают синтаксическую зависимость одного слова от другого, в то время как дополнительные связи могут содержать указания на удаленное управление слов, анафоры и прочие зависимости не между соседними элементами.

Также система использует шаблоны для описания возможного порядка слов в виде набора слотов, которые могут принимать различ-

ные значения, причём шаблоны хранятся отдельно от информации об ограничениях элементов для заполнения слотов.

Томита-парсер. Данный продукт разрабатывается компанией Яндекс, и предназначен для выделения из структурированного текста цепочек слов или фактов[11]. Парсер имеет в своем составе токенизатор для выделения слов, сегментатор для выделения предложений и морфологический анализатор под названием *mystem*[12]. В качестве лингвистической информации данный продукт использует словарь ключевых слов, набор правил, записанных на языке контекстно-свободных грамматик и множество описаний типов фактов, порождаемых грамматиками в процессе анализа.

Большим плюсом данной системы является то, что она никаким образом не завязана на какой-либо язык или набор правил — всё, что определяет её работу, задается пользователем. Однако система решает достаточно конкретную задачу, поэтому не может рассматриваться как система для полноценного анализа текста.

Прочие системы анализа текста. Помимо вышеописанных, в области синтаксического анализа существует ещё несколько систем:

• Синтактико–семантический анализатор русского языка группы SemanticAnalyzer Group. Данный проект принимал участие в соревнованиях парсеров в рамках конференции «Диалог-2012» [9]. Однако судя по информации, доступной в сети, это достаточно новый проект, который ещё не описан достаточно в рамках каких-либо статей или описания работы.

• Проект AotSoft также указан среди участников соревнований 2012 года, однако информации, изложенной на сайте проекта, явно недостаточно, чтобы делать какие-то выводы и описания.

• Система SynAutom имеет лишь несколько упоминаний о ней в рамках соревнования парсеров, однако никаких описаний в сети недоступно.

Заключение. После проведенного анализа продуктов, представленных в области анализа естественного текста, можно выделить характерные для всех перечисленных в обзоре систем черты:

1. В большинстве систем применяется стандартный порядок анализа текста — лексический, морфологический, синтаксический, семантический(если есть) анализ. Ни одна из представленных систем не предполагает возврат или смешение этапов анализа текста.

2. Все системы предполагают наличие корректного текста, лишь единицы предполагают хотя бы минимальное исправление ошибок.

3. Системы можно подразделить на 2 группы по применимости к предметной области — первая группа систем предполагает наличие конкретной предметной области с определенными правилами, словарями и структурой данных, вторая — наличие большого массива правил, позволяющих анализировать практически любой текст.

4. Основные улучшения стандартных подходов к анализу текста направлены на ускорение процесса перебора вариантов для разрешения омонимии на этапе синтаксического анализа.

5. Большинство систем показывают хорошие результаты лишь на грамматически корректных, построенных по всем правилам русского языка текстах. Отклонения в сторону неполных, безличных, сложно построенных предложений вызывают у большинства доступных для тестирования систем трудности с интерпретацией.

Также становится понятно, что большинство из них обладает существенными недостатками:

• Для описания правил, применяемых на различных этапах анализа, особенно на этапе синтаксического анализа, нужно очень хорошо представлять структуру и свод правил языка, для которого пишутся данные правила.

• Каждый анализатор имеет сою машину вывода со своим набором правил. Правила, созданные для одного анализатора, не всегда можно адаптировать для другого.

• Количество правил, применяемых в системах, достаточно большое, не менее нескольких сотен. С учетом такого огромного количества правил, нужны средства их диагностики и отладки.

• Некоторые системы требуют тщательной калибровки и подбора различных параметров и коэффициентов.

• Наконец, самый, пожалуй, существенный недостаток всех представленных систем – мало внимания уделяется некорректным и преднамеренно искаженным текстам. Некоторые системы имеют систему определения простейших опечаток и их автокоррекции. Однако даже такую коррекцию нельзя считать достаточной для полноценного анализатора естественного текста, так как во многих случаях текст далек

от идеально правильного, и требует тщательной обработки ошибок.

Таким образом, анализ текстов очень похож на экспертную систему диагностического типа: по входному тексту определяются его возможные трактовки.

Естественный язык — это живой язык, меняющийся со временем, приобретающий некоторые оттенки в зависимости от окружения, откуда взят тот или иной текст. Язык двух совершенно разных документов может быть одновременно корректным с точки зрения языковых правил, и иметь совершенно разную структуру и стиль, что очень сильно влияет на анализ данного текста. Поэтому надо понимать, что невозможно построение «универсального» анализатора на все случаи жизни.

Сформулируем требования, которым должен удовлетворять анализатор текста, чтобы его применение было простым и гибким:

1. Система правил всегда должна задаваться извне, чтобы изменение каких-либо норм языка, изменение среды, в которой работает анализатор, не влекло за собой переработку внутренней структуры системы, а могло регулироваться пользователем системы извне.

2. Должны быть средства отладки и диагностики системы правил.

3. Машина вывода должна быть простой, но внутри себя иметь мощную оптимизацию, позволяющую бороться с комбинаторным взрывом.

4. Нужно ориентироваться на обработку неправильных и преднамеренно искаженных текстов. Особенную важность данное требование приобретает для обработки текстов в сети Интернет.

СПИСОК ЛИТЕРАТУРЫ

1. *Кузнецов И.П.* Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных / И.П. Кузнецов, А.Г. Мацкевич. – URL : http://www.dialog-21.ru/digests/dialog2006/materials/html/KuznetsovI.htm (дата обращения: 28.06.2013)

2. *Мальковский М.Г.* Модель синтаксиса в системе морфосинтаксического анализа «Treeton» / М.Г. Мальковский, А.С. Старостин – URL: http://www.dialog-21.ru/digests/dialog2006/materials/html/Starostin.htm (дата обращения: 28.06.2013)

3. *Старостин А.С.* Синтаксический анализатор «Treevial». Принцип динамического ранжирования гипотез / А.С. Старостин , Н.В. Арефьев , М.Г. Мальковский. – URL : http://www.dialog-21.ru/digests/dialog2010/materials/html/71.htm (дата обращения: 28.06.2013)

4. *Баталина А.М.* Экспериментальная реализация сегментационного анализа русского предложения. / А.М. Баталина [и др.]. – URL : http://www.dialog-21.ru/digests/dialog2007/materials/html/04.htm (дата обращения: 28.06.2013)

5. *Скатов Д.С.* Синтаксический анализатор естественного языка Dictascope Syntax. / Д.С. Скатов [и др.]. – URL : http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Скатов.pdf

6. *Иомдин Л.Л.* Синтаксический анализатор системы ЭТАП: современное состояние. / Л.Л. Иомдин [и др.]. – URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Iomdin.pdf (дата обращения: 29.06.2013)

7. URL: http://habrahabr.ru/company/abbyy/blog/115226/ (дата обращения: 03.07.2013)

8. *Анисимович К.В.* Синтаксический и семантический парсер, основанный на лингвистических технологиях ABBYY Compreno. / К.В. Анисимович [и др.]. – URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/anisimovich.pdf (дата обращения: 03.07.2013)

9. *Толдова С.Ю.* Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка. / С.Ю. Толдова [и др.]. – URL : http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Toldova.pdf

Артемов Михаил Анатольевич – заведующий кафедрой математического обеспечения и администрирования информационных систем Воронежского государственного университета, доктор физико-математических наук, профессор. E-mail: atremov_m_a@mail.ru

Владимиров Александр Николаевич – аспирант, Воронежский государственный университет, факультет прикладной математики информатики и механики, кафедра программного

Artemov Mikhail A. – Head of Department Software & Information System Administering, Voronezh State University, doctor of Physics-math. Sciences, Professor. E-mail: atremov_m_a@mail.ru

Vladimirov Alexander – postgraduate, Voronezh State University, Applied Mathematics, Informatics and Mechanics department. E-mail: alcobass@gmail.com

обеспечения и администрирования информационных систем. E-mail: alcobass@gmail.com

Селезнев Константин Егорович – доцент, Воронежский государственный университет, факультет прикладной математики информатики и механики, кафедра программного обеспечения и администрирования информационных систем. E-mail: konstantin.seleznyov@gmail.com

Seleznyov Konstantin – docent, Voronezh State University, Applied Mathematics, Informatics and Mechanics department. E-mail: konstantin.seleznyov@gmail.com