УДК 004 **Информационные технологии**

Таржуманянц Ашот Самвелович, *техник ОИТСиЗИ*

Студент, КУБГТУ

Янаева Марина Викторовна, *зав. кафедрой ИСП ФГБОУ ВО*

к.т.н., доцент, КУБГТУ

СИСТЕМЫ ТЕКСТОВОГО ПОИСКА И АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация: В данной статье рассматриваются методы анализа которые представляют собой 4 этапа. Графематический, морфологический, синтаксический, семантический. С помощью данных методов можно извлекать информацию из текстов, представленных на естественных языках. В статье так же приводится подход анализа с помощью потоков интерпретаций, который является отличным решением в рамках данной проблемы.

Ключевые слова: анализ текста, извлечение текста, графематический анализатор, идентификация, морфологический анализатор, дерево синтаксического разбора, естественный язык, семантический анализ, онтология.

Annotation: This article discusses the methods of analysis that represent 4 stages. Graphematic, morphological, syntactic, semantic. Using these methods, it is possible to extract information from texts presented in natural languages. The article also provides an analysis approach using interpretation streams, which is an excellent solution to this problem.

Keywords: text analysis, text extraction, graphematic analyzer, identification, morphological analyzer, parsing tree, natural language, semantic analysis, ontology.

Приведем ниже общее описание процесса анализа текста, а также список основных проблем. На вход любой системе для анализа текста, как правило,

поступает текст в исходном формате, то есть в популярных форматах документов DOC, OAT, RTF, PDF, HTML [7]. Такие документы кроме самого текста содержат также форматирование, сноски, «лишние» участки текста (реклама). Первой сложностью при анализе текста является извлечение текста из документов [1; 2], поданных на вход. Далее, полученный текст следует пропустить через графематический анализатор, задача которого - определить границы слов, предложений, абзацев, а также отметить числа, знаки препинания и прочие символы и последовательности символов, не являющиеся словами, но играющие важную роль в предложении [4]. Сложность данного этапа состоит в том, что некоторые последовательности символов часто бывает невозможно идентифицировать на этом этапе анализа. Если принять во внимание необходимость обработки [5] опечаток, проблема идентификации слов становится непростой задачей.

Затем, когда предложения размечены, и слова локализованы, начинает работу морфологический анализатор. Применяя имеющийся словарь и набор словоизменительных правил, этот анализатор дополняет каждое слово набором грамматических характеристик. Среди них известные всем падежи, рода и числа, а также некоторые менее известные грамматические значения. На этапе морфологического анализа также возникает ряд сложностей с обработкой опечаток, особенно если они допущены в окончании слова, играющем, как известно, ключевое значение в определении набора грамматических характеристик. Ряд сложностей возникает при анализе неизвестных слов, а также заимствованных из других языков [9].

Следующий этап анализа - синтаксический. Наиболее популярным подходом к этому анализу является построение дерева синтаксического разбора предложения. В процессе построения дерева исходное предложение перестраивается в древовидную структуру, представляющую собой аналог дерева синтаксического разбора некоторой формальной грамматики. К сожалению, зачастую доступная в сети информация представлена в виде разговорной речи и, как следствие, слабо поддается полному синтаксическому

разбору, так как даже опытный лингвист не сможет построить дерево синтаксического разбора для подобных предложений. В качестве решения этой проблемы существует альтернативный подход к синтаксическому анализу, а именно - частичный синтаксический анализ. При частичном синтаксическом анализе полное дерево разбора не строится, а вместо этого анализатор [6] сосредотачивается на поиске заранее определенных синтаксических конструкций в тексте.

Такой подход вполне приемлем, так как естественный язык не имеет грамматики в привычном математикам смысле. Естественный язык не имеет терминальных и нетерминальных символов, правил вывода и не может быть отнесен ни к одному типу в иерархии Хомского [1]. Естественные языки, как правило, имеют достаточно вольные «правила» построения грамматических конструкций, слабо поддающиеся формализации.

В достаточной степени формализуемы лишь некоторые синтаксические отношения [10]:

1) Отношения согласования и зависимости при построении словосочетаний.

2) Валентности глаголов и отглагольных причастий.

3) Ссылки на контекст (различные типы местоимений).

4) Общие правила построения предложений (простое предложение, сложносочиненное, сложноподчиненное предложение и другие сложные предложения).

Заключительным этапом анализа текста является семантический анализ [11] результатов, полученных на этапе синтаксического анализа. Для данного этапа не существует устоявшихся моделей и подходов. В большинстве систем роль семантического анализатора играет эвристически реализованный модуль, осуществляющий поставленную перед ним задачу. Ясно, что даже при наличии дерева полного синтаксического разбора текста, все равно сложно реализовать алгоритм, извлекающий полезную информацию на основе полученного дерева.

В рамках данной работы проводился ряд исследований, направленных на

извлечение информации из текстов естественного языка для автоматизированного построения онтологий [3]. Одним из результатов является новый подход к анализу текста: «потоки интерпретаций» [13]. Потоком интерпретаций является множество пар (участок текста) + (некоторая информация). При этом не ставится никаких ограничений на тип информации и ее формат. Таким образом, основное отличие предлагаемого подхода от традиционного состоит в том, что, если традиционно принято делить процесс анализа на этапы, жестко разделяя при этом порядок исполнения этапов анализа, а также результат каждого этапа, то в подходе, основанном на потоках интерпретаций, предлагается фиксировать лишь формат результатов анализа.

Приведем пример потока интерпретаций для некоторого текста. Рассмотрим следующий текст:

«Аналитическая геометрия (раздел геометрии) исследует геометрические фигуры средствами алгебры» [14].

Для данного текста составляются следующие интерпретации [12]:

1. Интерпретации, соответствующие всем словам и знакам препинания.
2. Интерпретации, соответствующие результатам морфологического анализа слов.
3. Интерпретация, рассматривающая уточнение в скобках как пробельный символ.
4. Интерпретации, соответствующие терминам «аналитическая геометрия» и «геометрические фигуры».
5. Интерпретация, несущая информацию о разрешении валентности глагола «исследует»: субъект = аналитическая геометрия, объект = геометрические фигуры.
6. Любые другие виды информации об участках текста.

Анализаторы, действующие как преобразователи данных в традиционной модели [15] анализа текста, заменяются обработчиками потока интерпретаций, основная задача которых - дополнять поток интерпретаций новой информацией, на основе существующей. При этом не накладывается никаких ограничений на

количество альтернативных интерпретаций одного и того же участка текста.

Замена преобразователей информации обработчиками потока позволяет произвольным образом комбинировать этапы анализа, в том числе возвращаться на уже завершенный этап.

Возможность добавления альтернативных интерпретаций одного участка текста позволяет легко порождать альтернативные ветки анализа. Например, при добавлении двух разных интерпретаций одного участка текста следующий обработчик будет вынужден разветвить процесс анализа на две ветки: по одной на каждую альтернативную интерпретацию.

Существует возможность отмечать любой участок предложения как незначительный, добавив в качестве его интерпретации пробельный символ. Тогда все обработчики на одной из веток анализа будут считать, что вместо данного участка в тексте стоит пробельный символ.

Подобным образом легко реализуются различные автоматические «исправители» опечаток. Все возможные варианты просто добавляются в качестве альтернативных интерпретаций.

В модели потоков интерпретаций легко реализуется возврат на предыдущие этапы анализа и даже оценка достоверности полученных результатов. Обработчики потока интерпретаций независимы друг от друга, и поэтому анализ с их помощью представляет собой легко масштабируемый, прозрачный процесс [8].

Разумеется, подход, основанный на потоках интерпретаций, не лишен недостатков. Основной недостаток данного подхода - значительно большая сложность реализации обработчиков по сравнению с анализаторами. Увеличение сложности связано со сложностями обработки потока интерпретации, который может содержать альтернативные интерпретации одного участка текста, не интерпретированные участки и др. Кроме того, структура обработчиков, напрямую независимых друг от друга, безусловно, сложнее планарной структуры анализаторов. Тем не менее, получаемые удобства стоят затраченных усилий на разработку сложных алгоритмов

обработчиков потоков интерпретаций.

Выше описывается разработанный подход к анализу текстов на естественном языке при помощи потоков интерпретаций. Данный подход является естественным усовершенствованием традиционного поэтапного подхода к анализу. Имея достаточно приемлемый список недостатков, потоки интерпретации представляют собой удобный инструмент, позволяющий эффективно вмешиваться в любой этап процесса анализа.

Вместо набора преобразователей предлагается использовать обработчики потока интерпретаций, каждый из которых работает независимо и имеет своей целью дополнение потока новой информацией. Отсутствие ограничений на количество альтернативных интерпретаций одного и того же участка текста позволяет легко порождать альтернативные ветки анализа, сложно реализуемые в традиционном подходе.

Библиографический список:

1. Власов Д. Ю., Пальчунов Д. Е., Степанов П. А. Извлечение отношений между понятиями из текстов на естественном языке // Вестник НГУ. 2010. Т. 6. № 3.С. 23-33.

2. Зализняк А. А. Грамматический словарь русского языка. Словоизменение. Около 100000 слов. М.: Русский язык, 1997.

3. Пальчунов Д. Е. Решение задачи поиска информации на основе онтологий // Бизнес-информатика. 2008. № 1. С. 3-13.

4. Современный русский язык: учеб. для филол. спец. высших учебных заведений / В. А. Белошапкова, Е. А. Брызгунова, Е. А. Земская и др. 3-е изд., испр. и доп. М.: Азбуковник, 1997.

5. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. - М.: МИЭМ, 2011. -272с.

6. Попов Э.В. Общение с ЭВМ на естественном языке, изд. 2. - М.: Эдиториал УРСС, 2004. - 360 с.

7. Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск. - М.: Вильямс, 2011. - 528 с.

8. Daniël de Kok, Harm Brouwer Natural Language Processing for the Working Programmer. - 2011. URL: https://www.researchgate.net/publication/ 259572969_Draft_Natural_Language_Processing_for_the_Working_Pro grammer.

9. Statistical Natural Language Processing / In M. Lothaire, editor, Applied Combinatorics on Words. Cambridge University Press, 2005.

10. Brin, S. The Anatomy of a Large-Scale Hypertextual Web Search Engine / Sergey Brin, Lawrence Page// - Режим доступа: http://infolab.stanford.edu/pub/papers/google.pdf.

11. Некрестьянов, И.С. Латентно-семантический анализ: Введение в латентно-семантический анализ. - Режим доступа: http://meta. math.spbu.ru/~igor/papers/lsaprg/node2.html.

12. Studer, R. Knowledge Engineering: Principles and мMethods/ Studer R., Benjamins V.R., Fensel D. // In Data & Knowledge Engineering, 25, 1998. - P.161 - 197.

13. Когаловский, М.Р. Перспективные технологии информационных систем / М.Р. Когаловский. -М.: Компания АйТи, 2003. - 288 с.

14. Солтон, Дж. Динамические библиотечноинформационные системы. - М.: Мир, 1979.

15. Лифшиц, Ю. Модели информационного поиска. - Режим доступа: http://yury.name/internet/03ianote.pdf.