УДК 004.8

ПРИНЦИПЫ ПОСТРОЕНИЯ СИСТЕМ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ ДЛЯ СОЗДАНИЯ СЕМАНТИЧЕСКИХ БИБЛИОТЕК

Касекеева А. Б.
Докторант 3 курса ЕНУ имени Л.Н.Гумилева

ЕНУ имени Л.Н.Гумилева, г.Астана, Казахстан
E-mail: aibike_7474@mail.ru

Касекеева А. Б.
Окончила АГУ имени Абая, физико-математический факультет, докторант 3 курса ЕНУ имени Л.Н.Гумилева, г.Астана, Казахстан.

Аннотация. Приводится краткий обзор компонентов систем обработки текста. В обзор включены классические основы современных методов обработки. Рассматриваются виды анализов текста. Приводится пример использования анализатора Link Grammar Parser.

Ключевые слова: компьютерная лингвистика, обработка текста, естественный язык, семантическая библиотека.

В настоящее время происходит активизация в области исследования лингвистических проблем формальными методами и в области применения для этих целей компьютеров. В области формализации естественных языков и создании систем автоматической обработки текстов задействовано большое количество людей и мощностей, работающих в самых разных направлениях [1].

Это, прежде всего, связано с ростом производительности вычислительных систем, что позволяет в реальное время выполнять алгоритмы обработки текстов, которые раньше выполнить в реальное время было невозможно. Можно сказать, что увеличение вычислительных мощностей сделало возможным применение трудоемких лингвистических алгоритмов на больших объемах данных.

Одним из привлекательных направлений являются исследования, в которых предпринимаются попытки формализации семантических понятий, в том числе, понятия смысла текста. Результаты такого рода исследований могут быть применены в самых разных областях, начиная от фундаментальной лингвистики и до прикладных областей. В автоматизированных системах акцепции информации из текстов на естественном языке, интеллектуальных системах поиска информации в сети, при построении систем автоматического аннотирования, электронных переводчиков и словарей, систем безопасности, работающих с текстами на естественном языке и при создании семантических библиотек.

Компоненты, составляющие структуру систем анализа текстов, – лингвистические процессоры, которые друг за другом обрабатывают входной текст. Вход одного процессора является выходом другого. Выделяют [2] следующие компоненты систем обработки текстов.

1. Графематический анализ. Происходит выделение слов, цифровых комплексов, формул и т. д. Кроме деления текста на слова, на данном этапе анализатор разбивает текст на абзацы и предложения.

2. Морфологический анализ. Морфологический компонент осуществляет морфоанализ и лемматизацию словоформ. Морфоанализ – приписывание словоформам морфологической информации, лемматизация – приведение текстовых форм слова к словарным. При лемматизации для каждого слова входного текста морфологический процессор выдает множество морфологических интерпретаций следующего вида:

- лемма;
- морфологическая часть речи;
- множество наборов граммем.

Лемма – это нормальная форма слова. Например, для существительных – это единственное число (если оно есть у существительного), именительный падеж.

Граммема – это элементарный морфологический описатель, относящий словоформу к какому-то морфологическому классу, например, словоформе стол с леммой СТОЛ будут приписаны следующий набор граммем: «мр, ед, им, но», «мр, ед, вн, но». Таким образом, морфологический анализ выдает два варианта анализа словоформы стол с леммой СТОЛ внутри одной морфологической интерпретации: с винительным (вн) и именительным (им) падежами.

Также большую роль здесь играет омонимичность словоформ. Например, у словоформы стали могут быть следующие интерпретации: сталь – существительное и стать – глагол.

Таким образом, видно, что морфологического анализа явно не достаточно для выбора одной конкретной морфологической интерпретации слова. К тому же, выбор одной интерпретации может повлиять на выбор интерпретации для соседних слов. Поэтому программы работают с целым набором возможных морфологических интерпретаций, постепенно выделяя наиболее вероятные на следующих этапах анализа.

3. Фрагментационный анализ – деление предложения на неразрывные синтаксические единства (фрагменты), большие или равные словосочетанию (синтаксической группе), и установление частичной иерархии на множестве этих единств. Фрагменты – это главные и придаточные предложения в составе сложного, причастные, деепричастные и другие обособленные обороты. Иерархия отражает тот факт, что в предложении некоторые фрагменты синтаксически зависимы от других. Так, фрагмент «причастный оборот» будет подчиняться фрагменту, содержащему определяемое слово, придаточное предложение – главному.

4. Синтаксический анализ. Цель синтаксического анализа – построение групп в предложении. Фрагментационный анализ проходит после морфологического анализа и до синтаксического, поэтому на вход синтаксическому анализу предложение поступает по фрагментам. Это во много раз сокращает время работы синтаксического анализатора. Границы фрагментов не должны пересекать синтаксических связей, так что при правильной работе фрагментационного анализа не будут построены такие паразитические синтаксические связи, которые может допускать морфология. Таким образом, синтаксический анализатор решает одну из его основных задач – удаление значительной части морфологического шума и омонимичности словоформ.

5. Семантический анализ состоит в построении семантического графа текста. В отличие от морфологического и синтаксического, на семантическом этапе появляется

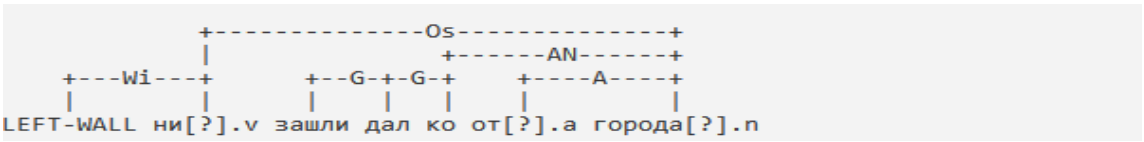
формальное представление смысла текста. В сферу семантического анализа входит построение семантической интерпретации слов и конструкций и установление «содержательных» семантических отношений между элементами текста, которые уже не ограничены размером одного слова. Результирующее представление, в котором решены эти задачи, является наиболее полным и законченным из всех возможных представлений, которые можно построить, пользуясь только лингвистическими средствами. Этим объясняется вся значимость такого анализа.

Семантика – раздел лингвистики, изучающий смысловое значение единиц языка: отдельных слов, словосочетаний, предложений, фрагментов текста. На данный момент существует ряд машинно-ориентированных методов отображения смысла высказываний.

Например, И.А. Мельчук ввел понятие лексической функции, развил понятия синтаксических и семантических валентностей и рассмотрел их в контексте толково-комбинаторного словаря [3]. В.Ш. Рубашкин и Д.Г. Лахути ввели иерархию синтаксических связей для более эффективной работы семантического анализатора. Подход И.А. Мельчука поддержан в программной системе Dialing [4].

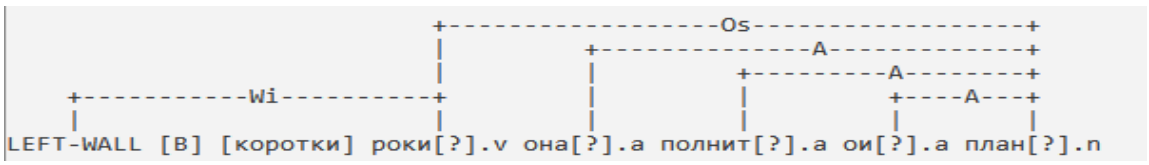
Еще один подход – это использование синтаксического анализатора Link Grammar Parser [5], разработанного в университете Корнеги-Меллона, базирующегося на некоторой специальной теории синтаксиса. Отметим, что данная теория, вообще говоря, отличается от классической теории синтаксиса. Получив предложение, система приписывает к нему синтаксическую структуру, которая состоит из множества помеченных связей (коннекторов), соединяющих пары слов. Приведем несколько примеров:

1. Они зашли далеко от города.



Здесь:
Os – соединяет подлежащий со сказуемым
Wi – находит подлещего
AN – соединяет существительное с определяющим его существительным
G – соединяет глагол с наречием
A - соединяет существительное и относящееся к нему прилагательное

2. В короткие сроки она выполнит все свои планы



Здесь:
Wi – находит подлещего
Os – соединяет подлежащий со сказуемым
A – соединяет существительное и относящееся к нему прилагательное

В статье описан процесс создания базы знаний, содержащей информацию в текстах на естественном языке. Формально описан процесс обнаружения перефразированных предложений, содержащих некоторые понятия. В настоящее время ведется работа по наполнению базы знаний. В результате проведенного исследования был сделан вывод о

целесообразности применения таком инструменте, как Link Grammar Parser, для решения поставленной задачи. Link Grammar Parser – достаточно необычная система, главной причиной, по которой анализатор называют семантической системой, можно считать уникальный по полноте набор связей.

Список литературы

[1.] Т.В. Батура, Ф.А. Мурзин. Машинно-ориентированные логические методы отображения семантика текста на естественном языке. – Новосибирск.: «Прайс-курьер», 2008. – 248 с.

[2.] А.В. Сокирко Семантические словари в автоматической обработке текста // Канд. дисс., МГПИИЯ, – Москва, 2000. – 108 с.

[3.] Мельчук И.А. Опыт теории лингвистических моделей типа «Смысл Текст». – М., 1974. – 315 с.

[4.] Сокирко А.В. Реализация первичного семантического анализа в системе ДИАЛИНГ // Тр. Международного семинара «Диалог'2000» по компьютерной лингвистике и ее приложениям. – Протвино, 2000. – 7 с.

[5.] Link Grammar Documentation, 2015, http://www.abisource.com/projects/link-grammar.

RESEARCH METHODS OF INFORMATION RETRIEVAL

A. Kassekeyeva
Doctoral student of the 3nd course of the
Eurasian National University named after
L.N.Gumilyov

Eurasian National University named after L.N.Gumilyov, Astana, Kazakhstan
E-mail: aibike_7474@mail.ru

Absrtact. A brief overview of the components of word processing systems. The review includes the classic foundations of modern processing methods. The types of text analysis are considered. An example of using the Link Grammer Parser analyzer is given.

Keywords: computer linguistics, word processing, natural language, semantic library.