УДК 004.5+004.82

РАЗРАБОТКА МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ДЛЯ ПОСТРОЕНИЯ ПОНЯТИЙНОГО АППАРАТА ЭЛЕКТРОННОЙ БИБЛИОТЕКИ КАФЕДРЫ АСУ

Бажанова А.И., Мартынеко Т.В., Андриевская Н.К.
Донецкий национальный технический университет,
кафедра автоматизированных систем управления
A_Bazh@rambler.ru

Бажанова А.И., Мартыненко Т.В., Андриевская Н.К. Разработка морфологического анализатора для построения понятийного аппарата предметной области. *Проведен обзор методов морфологического анализа текста. Сформулирована математическая постановка задачи построения морфологического анализатора. Разработан алгоритм и программное средство для построения понятийного аппарата предметной области.*

Общая постановка проблемы

Задачи обработки текстов возникли практически сразу вслед за появлением вычислительной техники. Но несмотря на полувековую историю исследований в области искусственного интеллекта, огромный скачок в развитии ИТ и смежных дисциплин, удовлетворительного решения большинства практических задач обработки текста пока нет. Актуальность проблемы морфологического анализа и синтеза словоформ определяется тем, что блок морфологического анализа является необходимой частью большинства работающих с естественно-языковыми текстами программ самого различного уровня и назначения.

Не исключением являются системы семантического поиска информации, основанные на онтологиях. Онтологически модели обеспечивают сведение ресурсов, относящихся к одной области знаний в единое информационное пространство, обеспечивают возможность открытого и удобного доступа к ним, а также автоматизируют оперативный сбор и индексацию новой информации, поступающей в текстовом неструктурированном виде, а задача разработки информационных порталов знаний является одной из самых актуальных на сегодняшний день [2].

Для построения онтологии предметной области необходимо выполнить несколько этапов, первым из которых является построение терминологического словаря предмтной области. Построение полноценного терминологического словаря предметной области в ручном режиме практически невозможно, поэтому предлагается для решения данной задачи использовать морфологический анализ текста. Морфологическим анализом называется установление по словоформе исходного слова — лексемы, а также морфологических характеристик данной словоформы, таких как род, падеж, число и т.д. Разрабатываемый морфологический анализатор должен будет выполнять морфологический анализ, и выявлять существительные-понятия данного текста. Результатом такого анализа должно быть построение понятийного аппарата предметной области кафедры АСУ.

Математическая постановка задачи морфологического анализа

Для выделения понятий в научном тексте в начале выполняется лексический и морфологический анализы текста. В результате анализов осуществляется преобразование текста в поток лексем с характеристиками, отражающими морфологические признаки выявленных в тексте понятий.

Лексемы делятся на классы. В данном случае, грамматические классы слов: существительные, прилагательные, глаголы и т. д.

Лексический анализатор должен выдавать следующую информацию:

- поток основ слов или множество векторов лексем $L=\{ l_i \mid i=1..k \}$, где k – общее количество лексем в потоке

– множество $L^s = \{ p_i^{ls} \mid i = 1..k' \}$, где k′ – количество разновидностей лексем в потоке, k′ ≤ k, вектор p_i^{ls} содержит статические характеристики лексемы l_i.
– множество векторов $L^v = \{ p_i^{lv} \mid i = 1..k \}$, которые описывают динамические характеристики лексемы l_i, зависящие от контекста.

Вектор p_i^{ls} содержит значения параметров лексемы l_i, которые характеризуют лексему в общем:

$$p_i^{ls} =< n_i, l_i, f_i, m_i, p_i >, \quad (1)$$

где n_i – уникальный номер вектора p_i^{ls}; l_i – основа лексемы; f_i – частота встречаемости лексемы в тексте; m_i – класс лексемы (здесь часть речи); p_i – указатель на группу векторов, описывающих динамические параметры лексемы.

Вектор p_i^{lv} содержит значения таких параметров, которые отражают морфологические и синтаксические характеристики лексемы, такие, как падеж и число лексемы для существительных и прилагательных, адрес лексемы в тексте:

$$p_i^{lv} =< p_i, n_i, c_i, a_i >, \quad (2)$$

где p_i – уникальный номер вектора лексемы p_i^{lv}; n_i – уникальный номер вектора лексемы p_i^{ls}; c_i – морфологическая информация, такая как род, число и падеж; a_i – адрес лексемы l_i;

a_i=<n_{il}, n_{is}, n_{ip}, n_{id}, n_{ic}>, где n_{il} – порядковый номер лексемы в предложении; n_{is} – порядковый номер предложения в документе; n_{ip} – номер параграфа; n_{id} – номер раздела; n_{ic} – номер главы.

Таким образом, результаты анализов текста будут представлены множествами L, L^S и L^v [1].

Разработка морфологического анализатора

Для выполнения морфологического анализа будут использоваться следующие словари S:
– готовых словоформ, выраженных существительными (слова исключения, не подвергающиеся склонению) S1;
– основ существительных S2;
– окончаний существительных S3;
– основ прилагательных и причастий S4;
– окончаний прилагательных S5;
– основ глаголов S6;
– окончаний глаголов S7;
– наречий S8;
– словарь служебных частей речи и коротких часто используемых слов S9;

Словарь готовых (неизменяемых) словоформ – это упорядоченный по алфавиту перечень лексем-существительных, неизменяемых в зависимости от грамматической формы. К таким словам относятся заимствованные из других языков понятия.

Словарь служебных частей речи и коротких часто используемых слов – словарь предлогов, союзов, частиц, также некоторых вводных слов и прочих, которые не могут повлиять на смысл текста и его семантику, а значит, являются не существенными для морфологического разбора и могут быть отброшены.

Словарь наречий – упорядоченный по алфавиту перечень наречий со слитным написанием.

Словарь окончаний существительных – это перечень всех возможных окончаний имен существительных.

Все выше перечисленные словари должны быть заполнены перед началом морфологического анализа. Наиболее существенным является словарь окончаний. Он используется для определения необходимой морфологической информации, а именно числа и падежа для имен существительных и прилагательных, склонения для глаголов.

Словарь основ существительных, прилагательных, глаголов может формироваться во время

работы морфологического анализа.

В связи с тем, что термин может состоять из нескольких лексем, для построения категориально-понятийного аппарата научно-технического текста необходимо выявлять подобные словосочетания и в дальнейшем они будут принимать участие в разборе текста, как единая лексическая единица. Для выявления составных терминов необходимо использовать модель именных словосочетаний, выражаемую схемой: согласуемое слово + существительное. В общем случае именные словосочетания могут включать в свой состав следующие классы слов: существительные, прилагательные, предлоги, сочинительные союзы и наречия. Количество слов в именных словосочетаниях колеблется от двух до пятнадцати и в среднем составляет три слова. В работе [1] представлен перечень всех возможных именных словосочетаний. Необходимым является выделение словосочетаний всех приведенных в таблице структур.

Алгоритм автоматической обработки текста

Для морфологического разбора текста будет использован комбинированный метод. Т. е. будет использован процедурный метод в общем случае и декларативный в случае, когда слово нельзя разбить на основу и аффикс, в этом случае будет использоваться словарь исключений [3].

В общем виде схема морфологической обработки текста показана на рисунке 1.

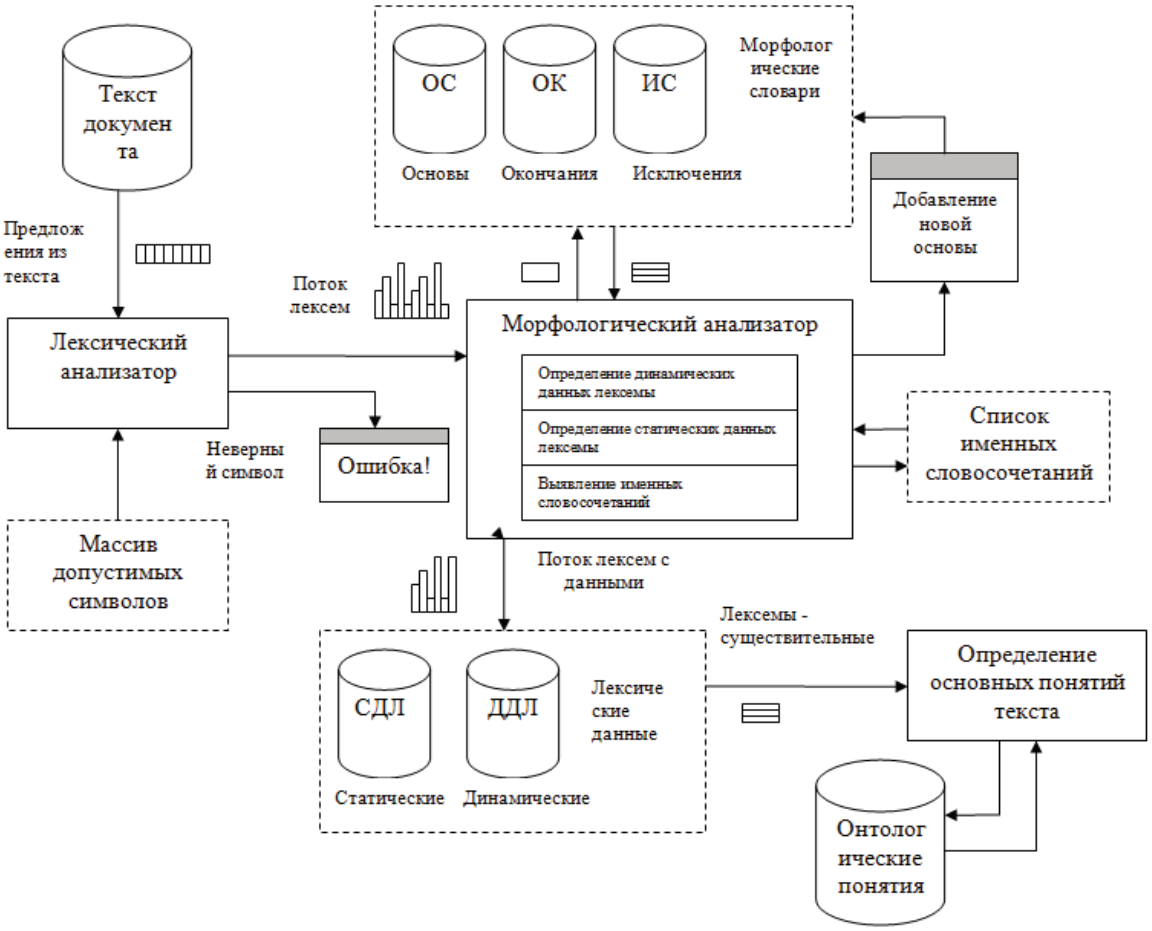

Рисунок 1. Схема морфологического анализатора

Сначала проводится лексический анализ, т. е. проверяются допустимые символы. На вход лексического анализа подаются предложения из текста поочередно, а на выходе проверенный набор слов и знаков препинания. На вход морфологического анализатора поступает массив «слов», знаков препинания и чисел, выделенных из входного текста на этапе лексического анализа. Предложение разбивается на лексемы. Для каждой лексемы, производится поиск по словарю коротких и служебных

слов (поиск так, называемых стоп-слов – информационно считающихся пустыми). Если слово найдено в словаре его анализ на этом завершается (запоминается только часть речи и само слово). Если слово в словаре не найдено производится разбиение слова на основу и аффикс – окончание. Заполняются статические и динамические данные лексемы.

Среди массива динамической информации лексем производится поиск именных словосочетаний. Формируется массив словосочетаний и незадействованных в словосочетаниях существительных. То же самое повторяется для следующего предложения в тексте.

После преобразования всех предложений в тексте, для каждого найденного существительного проверяется частота его встречаемости в тексте. Если частота больше порогового значения, то проверяется наличие в базе знаний идентичных понятий. Если таковое не было найдено, понятие добавляется в онтологию и запоминается ссылка на источник данных.

Данный морфологический анализатор был разработан в среде MS Visual Studio 2010 на языке C#. Для реализации было создано несколько модулей: модуль взаимодействия с БД, модуль морфологического анализа, модуль лексического анализа, модуль оценки важности данного понятия для предметной области и модуль взаимодействия с пользователем.

Исследования работы морфологического анализатора

Для оценки работы морфологического анализаторы был проведен сравнительный анализ данных полученных экспертом и данных полученных при автоматическом разборе текста.

Сбор данных от эксперта производился с помощью формы, которая показана на рис. 2.

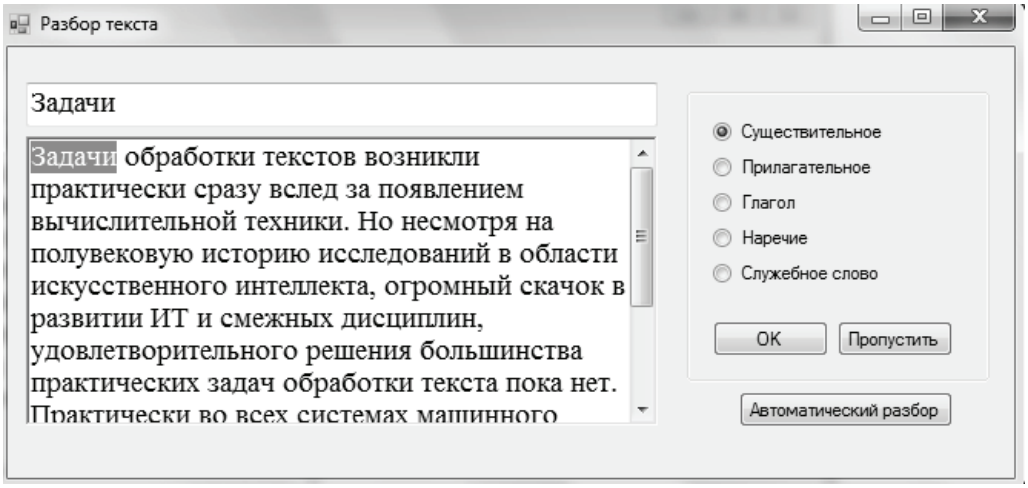

Рисунок 2. Экранная форма взаимодействия с экспертом

Гистограмма, показанная на рисунке 3, отображает состояние заполнения каждого из словарей при разборе текста экспертом и созданным морфологическим анализатором.

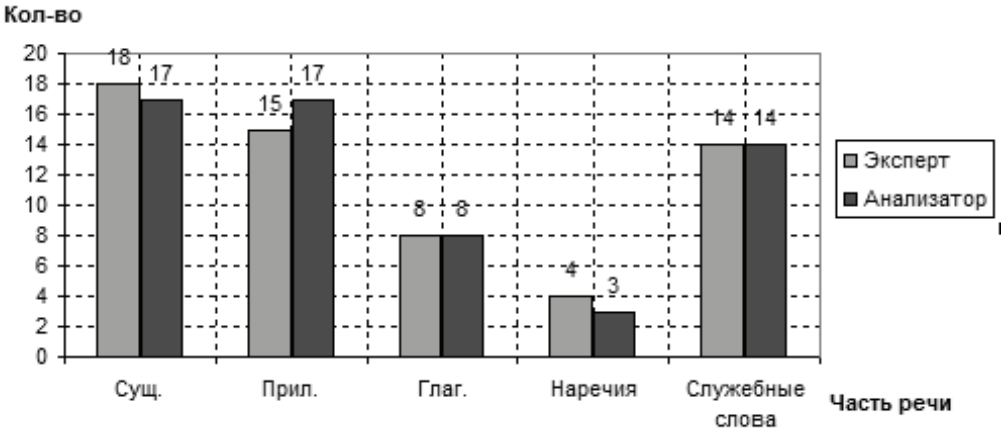

Рисунок 3. Гистограмма заполнения словарей

Исходя из графика видно, что погрешность между данными не превышает 2%, а значит, созданный морфологический анализатор, справляется с поставленной задачей.

Выводы

В данной статье рассматривался морфологический разбор текста для построения понятийного аппарата электронной библиотеки кафедры АСУ, который может быть использован в проектировании онтологической модели.

Для автоматизации процессов построения онтологической модели необходимо иметь словарь терминов и понятий, которые являются основными компонентами онтологии. Для выявления понятий следует использовать морфологический разбор текста, потому как определение терминов вручную практически невозможо. После морфологического разбора понятия текста необходимо привести в нормальную форму. Важным аспектом является выделение из текста, понятий выраженных словосочетаниями. Для этого используется перечень именных словосочетаний. Предлагается реализовывать выявление данных структур из текстов с помощью продукционных правил.

Литература

[1] Найханова Л.В. Основные аспекты построения онтологий верхнего уровня и предметной области: Монография. – Улан-Удэ: Изд-во БНЦ СО РАН, 2008, – 244 с.

[2] Андреев А.М., Березкин Д.В., Брик А.В. Лингвистический процессор для информационно-поисковой системы – М: МГУ

[3] Королёв А.Н. Лингвистическое обеспечение информаци-онно-поисковой системы Excalibur RetrievalWare: Аналитический аспект.