МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего профессионального образования

«Волгоградский государственный социально-педагогический университет»

**Лингвистические информационные ресурсы и их применение для задач компьютерной обработки естественного языка**

*Реферат*

 **Выполнила:**

 студентка филологического факультета,

 гр. ФЛ-РАБ-11

 Голыгина Ольга Алексеевна

 **Научный руководитель:**

Махонина Анжела Анатольевна

Волгоград, 2014

**Содержание**

Введение……………………………………………………………………………...3

ГЛАВА 1. ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА…………………………...4

* Обработка естественного языка: основные понятия………………………4
* Основные задачи обработки естественного языка…………………………6
* Трудности, возникающие при выполнении задач обработки естественного языка…………………………………………………………………………..9
* Программное обеспечение, предназначенное для обработки естественного языка…………………………………………………………10

ГЛАВА 2. ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ…………………………………12

2.1. Компьютерные словари……………………………………………………….13

2.2. Компьютерные базы словосочетаний………………………………………..16

2.3. Компьютерные грамматики…………………………………………………..17

2.4. Составление лингвистических информационных ресурсов и автоматизация этого процесса………………………………………………………………………18

Заключение………………………………………………………………………….20

Список использованной литературы…………………………………………...…21

**Введение.**

 В настоящее время основными проблемами лингвистики является изучение лексики и семантики, быстрый автоматизированный перевод. В этих исследованиях невозможно обойтись без работы со словарями, архивами. Но, к сожалению, у учёных не всегда существует возможность доступа к необходимым информационным ресурсам. Помочь в этом современным лингвистам может такая отрасль науки, как компьютерная, прикладная лингвистика, которая занимается созданием разнообразных систем по обработке естественного языка. Но эта обработка невозможна без наличия лингвистических информационных ресурсов. Что же это за ресурса и как ими пользоваться? Этому и посвящено наше исследование.

 *Цель работы* – исследовать лингвистические информационные ресурсы.

 Поставленная цель обусловила следующие *задачи*:

* определить понятие «обработка естественного языка» и сопутствующую ему терминологию;
* выявить основные задачи обработки естественного языка;
* выявить трудности, возникающие при выполнении задач обработки естественного языка;
* описать программное обеспечение, предназначенное для обработки естественного языка;
* описать и классифицировать лингвистические ресурсы;
* описать процесс составления лингвистических информационных ресурсов.

**ГЛАВА 1.**

**ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА**

* **Обработка естественного языка: основные понятия**

 Для начала следует выяснить, что понимается под термином обработка естественного языка.

 *Обработка естественного языка* — одно из направлений искусственного интеллекта и математической лингвистики, которое занимается изучением проблем компьютерного анализа и синтеза естественных языков.

 Расшифруем некоторые понятия, входящие в состав данного определения.

 *Естественный язык* – это хранящаяся в сознании человека сложная система правил, в соответствии с которыми происходит речевая деятельность, т.е. порождение и понимание текстов.

 *Искусственный интеллект* – это область компьютерной науки (информатики), специализирующаяся на моделировании интеллектуальных и сенсорных способностей человека с помощью вычислительных устройств.

 *Математическая лингвистика* – математическая дисциплина, предметом которой является разработка и изучение понятий, образующих основу формального аппарата для описания строения естественных языков (т. е. метаязыка лингвистики).

 *Анализ* как форма и метод деятельности искусственного интеллекта трактуется как понимание языка. Под *синтезом* же в сфере искусственного интеллекта понимают генерацию грамотного текста.

* **Основные задачи обработки естественного языка**

 Понимание, распознавание естественного языка – ключевая задача, поскольку узнавание и распознавание языка живого требует колоссальных знаний языковой системы, языкового строя, их особенностей и закономерностей.

 Определим основные, наиболее актуальные *задачи обработки естественного языка*. Это:

* распознавание речи;
* синтез речи;
* анализ текста;
* синтез текста;
* машинный перевод;
* создание вопросно-ответных систем;
* информационный поиск;
* извлечение информации;
* анализ тональности текста
* реферирование текста.

 Итак, рассмотри подробнее сущность названных задач.

* В первую очередь следует рассмотреть одну из наиболее важных задач, а именно *распознавание речи*. Под этим процессом подразумевается процесс, ведущий к преобразованию речевого сигнала человеческого голоса в цифровую информацию. Такая возможность может быть использована людьми, лишёнными способности набора текста с помощью рук либо для упрощения и ускорения этого процесса.
* *Синтез речи* – формирование по печатному тексту сигналов речи, то есть искусственное производство человеческой речи. Выполнением этой задачи занимается такая отрасль современной информатики, компьютерной лингвистики, информационных технологий, как искусственный интеллект.

 Эта задача в основном предназначена для использования в информационных и справочных системах, диспетчерских службах, для выдачи информационных запросов о технологических процессах, для помощи людям с ограниченными возможностями зрения и речи.

* *Анализ текста* – процесс извлечения содержательной, высокого качества информации из текста на естественном языке для автоматизации процесса извлечения и анализа данных.
* *Синтез текста* мы определяли выше, в предыдущем пункте данной главы. Ему присуще наличие некоторых специфических подзадач. К примеру, многоязыковая генерация. Так называют автоматическое составление специальных документов на нескольких языках (патентных формул, инструкций по эксплуатации технических изделий или программных систем). Для решения этого типа задач применяются подробные языковые модели.
* *Машинный*, или *автоматический перевод*. Под данной задачей обработки естественного языка подразумевается процесс перевода устных текстов, написанных на естественном языке, на другой, тоже естественный, язык при помощи электронно-вычислительных машин в предназначенных для данного типа задач компьютерных программах.
* *Создание вопросно-ответных систем*, то есть таких информационных систем, которые способны принимать, распознавать, классифицировать вопросы и давать ответы на них на естественном языке.

 Эта задача осуществляется по следующему алгоритму:

* определения типа вопроса;
* поиском текстов, потенциально содержащих ответ на этот вопрос;
* извлечением ответа из этих источников.

 Такие системы можно классифицировать на:

* те, которые предназначены для работы с текстами и темами определённой тематики,
* те, которые способны работать с информацией, относящейся к различным сферам знания.
* *Информационный поиск* – процесс выявления информации в документах, содержащихся в доступных системе поиска базах данных, которые соответствуют заданному запросу по тематике.

 Исполнение этой задачи подразумевает под собой исполнение следующей последовательности операций:

* формулирование информационного запроса;
* поиск потенциальных обладателей соответствующей информации;
* извлечение информации из найденных документов;
* ознакомление с результатами произведённого поиска и выборка наиболее подходящих условиям запроса источниками.

 Существует 3 разновидности информационного поиска:

* поиск по содержанию всего документа;
* поиск по названию документа, дате его создания, автору, размеру и т.п. данным;
* поиск по тематике изображения, предмету, присутствующему на нём.
* *Извлечение информации* – задачи обработки естественного языка, выполняющая автоматическое извлечение необходимых данных из источника информации, текста (как правило, неструктурированного). Эта задача непосредственно связана с предыдущей.
* *Анализ тональности текста* – анализ лексем текста, оценка их эмоциональной окрашенности и классификация по принадлежности к нейтральному, позитивному или негативному лексическому слою языка.
* *Реферирование* – сокращение объёма текста за счёт выделения основных тезисов путём поиска соответствий заданным в поиске ключевым словам и его краткое изложение.
* **Трудности, возникающие при выполнении задач**

**обработки естественного языка**

 Когда эти условия будут выполняться, тогда можно будет говорить о создании весьма удобной формы взаимодействия человека с компьютером – более простой, доступной, понятной, продуктивной, эргономичной.

 Но, к сожалению, в процессе выполнения этой задачи возникают препятствия, создаваемые теми или иными особенностями естественного языка. Например, на качество понимания текста могут повлиять такие факторы, как:

* непосредственно сам язык, его отнесённость к той или иной языковой семье, группе;
* порядок речи (прямой, обратный или свободный – чаще всего при обработке русского языка трудности связаны именно с ним);
* характерные особенности национальной культуры носителей естественного языка;
* говорящий, источник речи, а именно:
* логический строй речи;
* синтаксическое построение речи;
* грамматический строй речи;
* грамотность;
* фонетические особенности речи;
* полисемичность языка;
* наличие омонимов в данном естественном языке;
* способы словообразования, присущие определённому языку;
* неологизмы, окказионализмы;
* фразеологические обороты и устойчивые выражения.
* **Программное обеспечение, предназначенное для обработки естественного языка**

 Осуществление названных задач требует наличия определённого программного обеспечения. Укажем наиболее популярные и востребованные пакеты программного обеспечения, способного выполнить тот или иной тип обработки естественного языка:

* General Architecture for Text Engineering (GATE)
* [ModularHYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework" HYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework"AudioHYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework" HYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework"RecognitionHYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework" HYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework"Framework](http://ru.wikipedia.org/wiki/Modular_Audio_Recognition_Framework)
* [NaturalHYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit" HYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit"LanguageHYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit" HYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit"Toolkit](http://ru.wikipedia.org/wiki/Natural_Language_Toolkit) (NLTK).
* *General Architecture for Text Engineering (GATE)* – одна из систем обработки естественного языка с открытым исходным кодом. Её работа основана на наборах компонентов, созданных на языке [Java](http://ru.wikipedia.org/wiki/Java). Данная система начала своё развитие в 1995 году.

 С помощью GATE стала возможна реализация таких задач, в которых требуется определить смысловое содержание некого текста и кодировать его путем добавления аннотаций к отдельным сегментам текста в структурированном виде. Эта система применяется для выполнения задач по извлечению информации, по составлению как ручной, так и автоматической семантической аннотации текста, работы с онтологиями, машинного обучения, анализа потока сообщений в блогах.

 Система GATE получила распространение в различных сферах научных знаний. Нас интересует только её применение в компьютерной лингвистике, обработке естественного языка, а также в моделировании языковых процессов.

* [*ModularHYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework" HYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework"AudioHYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework" HYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework"RecognitionHYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework" HYPERLINK "http://ru.wikipedia.org/wiki/Modular\_Audio\_Recognition\_Framework"Framework*](http://ru.wikipedia.org/wiki/Modular_Audio_Recognition_Framework) *(MARF)* – модульная библиотека распознавания аудио. Данное программное обеспечение предназначено для выборки алгоритмов для обработки и распознавания текста, речи, звука, то есть выполнения непосредственных потребностей поставленных задач обработки естественного языка. Это приложение также способно осуществлять дополнение алгоритмов и модулей. Система MARF способна работать как библиотека в приложениях или может использоваться как источник для изучения и расширения лингвистических баз данных.
* [*NaturalHYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit" HYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit"LanguageHYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit" HYPERLINK "http://ru.wikipedia.org/wiki/Natural\_Language\_Toolkit"Toolkit*](http://ru.wikipedia.org/wiki/Natural_Language_Toolkit)*(NLTK)* – библиотеки и программ для символьной и статистической обработки естественного языка.

 NLTK предназначена для студентов, изучающих и занимающихся компьютерной лингвистикой, эмпирическая лингвистикой, когнитивистикой, искусственным интеллектом, информационным поиском.

 Все перечисленные программные обеспечения предназначены для создания лингвистических информационных ресурсов. Рассмотрим несколько аспектов, касающихся этих ресурсов.

**ГЛАВА 2. ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ**

 Разработка лингвистических процессоров (к примеру тех, о которых говорилось в предыдущей главе) требует широкого и наиболее полного представления лингвистической информации о естественном языке, который требуется обработать. Такая информация отображается в таких лингвистических информационных ресурсах, как:

* компьютерные словари;
* базы словосочетаний;
* компьютерные грамматики.

 Рассмотрим подробнее типы этих ресурсов.

**2.1. Компьютерные словари**

 Именно словари являются наиболее традиционной формой представления лексической информации. Так же и в компьютерной лингвистике – наиболее распространённой и широко используемой являются компьютерные словари.

 Словари различаются:

* единицами (как правило, это слова или словосочетания);
* структурой;
* охватом лексики (к примеру, словари терминов конкретной узкой области, словари общей лексики и т.п.).

 Для устранения дальнейших сложностей в понимании определим, что же такое словарная единица.

 Единицей словаря называется словарная статья. В словарной статье представлена наиболее полная информация о лексеме. Также следует отметить, что лексические омонимы обычно представляются в разных словарных статьях, так как они содержат в себе разное смысловое наполнение и не могут являться одной и той же лексической и словарной единицей.

 Здесь следует назвать некоторые виды компьютерных словарей:

* морфологические словари;
* толково-комбинаторные словари;
* словари синонимов;
* словари паронимов;
* тезаурусы;
* онтологии.

 Определим функции и общее назначение каждого из этих типов.

* *Морфологические словари* наиболее распространены в компьютерной лингвистике. Используются они для морфологического анализа. В словарной статье представляется морфологическая информация об определённом слове, а именно такие характеристики слова, как:
* часть речи;
* словоизменительный класс (только для языков флективного типа);
* перечень значений слова и т.п.

 Иногда в морфологический словарь может быть добавлена и грамматическая информация (например, модели управления слова). Это обстоятельство зависит от организации и возможностей лингвистического процессора.

* *Толково-комбинаторный словарь*. В нём представлена более полная и широкая информация о словах. В словарной статье в этом случае кроме информации морфологической, синтаксической и семантической (синтаксические и семантические валентности) предлагаются сведения о лексической сочетаемости этого слова. На толково-комбинаторный словарь опирается лингвистическая модель «Смысл⇔Текст». Такая модель призвана обеспечить получение связных синтаксических структур для всех предложений текстов, подвергаемых обработке, причём независимо от степени их сложности, и переработку текстов на естественном языке без смысловых потерь.
* В ряде лингвистических процессоров используются *словари синонимов.* В структуру такой словарной статьи входят:
* заглавное слово;
* антоним к заглавному слову;
* вариант заглавного слова;
* синоним;
* вариант синонима.
* *Словарь паронимов* – это довольно новый вид не только компьютерных, но и словарей вообще. Паронимы – это внешне схожие слова, различающиеся по смыслу, например, чужой и чуждый, правка и справка. Эти словари предназначены для исправления ошибок в тексте и обучения правильному употреблению паронимов в разных контекстах.
* *Тезаурус* – это семантический словарь. В нём представлены смысловые связи слов – синонимические, родовидовые отношения, отношения части и целого, ассоциативные отношения. Применение и составление тезаурусов связано с решением задач информационного поиска.
* *Онтология* – это набор понятий и сущностей определенной области, сферы знаний, который ориентирован на многократное использование для различных задач. Такие словари могут создаваться на базе существующей в языке лексики – в этом случае они называются лингвистическими.

 Такой лингвистической онтологией считается система WordNet – объёмный лексический ресурс, в котором собраны слова английского языка: существительные, прилагательные, глаголы и наречия. Там же представлены их смысловые связи нескольких типов. Для каждой из названных частей речи слова сгруппированы в синонимические ряды (синсеты), между которыми устанавливаются отношения антонимии, гипонимии (отношение род-вид), меронимии (отношение часть-целое).

Общую онтологию формирует верхний уровень иерархии – система основных понятий о мире.

 По аналогичной схеме были построены лексические ресурсы для других языков Европы, объединенные под общим названием EuroWordNet.

**2.2. Компьютерные базы словосочетаний**

 В базы словосочетаний отбираются наиболее типичные, устойчивые словосочетания, фразеологические обороты конкретного языка. Такая база словосочетаний русского языка, содержащая около миллиона единиц, составляет ядро системы КроссЛексика.

 *Система КроссЛексика* – это большой электронный словарь сочетаний и смысловых связей русских слов.

 Словосочетания вносятся в компьютерную базу в соответствии со следующими принципами:

* высокая частотность;
* высокая степень ассоциации;
* синонимичность лексической единице;
* значительная многозначность компонентов;
* обозначение типа объекта.

**2.3. Компьютерные грамматики**

 Ещё один вид лингвистических ресурсов – это *грамматики естественного языка*. Их тип зависит от используемой в лингвистическом процессоре модели синтаксиса. Компьютерная грамматика представляет собою набор правил, которые выражают общие синтаксические свойства слов и групп слов. Общее количество правил компьютерной грамматики зависит и от модели синтаксиса, изменяясь от нескольких десятков до нескольких сотен.

 Но здесь существует такая проблема, как соотношение в модели языка грамматики и лексики: чем больше информации представлено в словаре, тем короче может быть грамматика и наоборот.

**2.4. Составление лингвистических информационных ресурсов и автоматизация этого процесса**

 Составление компьютерных словарей, тезаурусов и грамматик – это весьма объемная и трудоемкая работа. Чаще разработка лингвистической модели и соответствующего процессора оказывается проще, нежели составление лингвистических информационных ресурсов. Следовательно, одной из подзадач компьютерной лингвистики является автоматизация построения лингвистических ресурсов.

 Компьютерные словари часто формируются конвертацией обычных текстовых словарей. Но при построении словарей и тезаурусов для быстро развивающися научных областей требуется намного более кропотливая и сложная работа. Для извлечения необходимой лингвистической информации исходным материалом являются коллекции и корпуса текстов.

 *Корпус текстов* – это коллекция текстов, собранная по определенному принципу (например, по жанру, авторской принадлежности и т.п.), в которой все тексты снабжены некой лингвистической разметкой (или аннотациями) – морфологической, акцентной, синтаксической и т.д. На данный момент существует более сотни различных корпусов – для разных естественных языков и с различной разметкой. В России наиболее известным корпусом текстов является Национальный корпус русского языка.

 Размеченные корпуса создаются лингвистами и используются для:

* выполнения целей и задач, которые преследуются в лингвистических исследованиях;
* настройки используемых в компьютерной лингвистике моделей и процессоров с помощью известных математических методов машинного обучения.

 В последнее время чаще в качестве более полного лингвистического ресурса рассматриваются тексты сети Интернет, так как корпусы и коллекции текстов обязательно ограничены по представленным в них языковым явлениям. Ведь, бесспорно, Интернет на сегодняшний день – самый представительный источник образцов современной речи. Тем не менее, его использование как корпуса требует разработки специальных технологий.

**Заключение**

 На сегодняшний день компьютерная лингвистика достигла довольно высоких результатов в создании различных приложений по автоматической обработке текстов на естественном языке. Дальнейшее ее развитие зависит и от разработки новых приложений, и от независимой разработки различных моделей языка. Наиболее отработанными являются модели морфологического анализа и синтеза. Модели синтаксиса еще не доведены до уровня устойчиво и эффективно работающих модулей, несмотря на большое количество существующих формализмов и методов. Еще менее изучены и формализованы модели уровня семантики и прагматики, хотя такая потребность на данный момент возникла. Решить многие проблемы и вопросы, возникающие в ходе этой работы, может развитие лингвистических информационных ресурсов.

**Список использованной литературы**

* Большакова Е.И., Клышинский Э.С., Ландэ Д.Э., Носков А.А., Пескова О.В., Ягунова Е.В., Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. – М.: МИЭМ, 2011. – 272 с.
* Естественно-языковые системы: курс лекций. – Улан-Удэ: Изд-во ВСГТУ, 2006. – 92 с.: илл.
* Колодяжная Л.И., Поликарпов А.А., Некоторые количественные характеристики русской синонимии на материале "Словаря синонимов" под редакцией А.П. Евгеньевой (Л.: Наука, 1975) [Электронный ресурс] – <http://ns2.philol.msu.ru/~humlang/articles/polkolsyn.htm>
* Лукашевич Н.В., Добров Б.В., Чуйко Д.С., Отбор словосочетаний для словаря системы автоматической обработки текстов [Электронный ресурс] – <http://www.cir.ru/docs/ips/publications/2008_dialog_extraction.pdf>