Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

И. Г. ГОМОНОВА, И. В. СЕРИКОВА

КОМПЬЮТЕРНАЯ ФИЛОЛОГИЯ

Практическое руководство

для студентов филологического факультета
специальности 1-21 05 02-02
«Русская филология (компьютерное обеспечение)»

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Гомель
ГГУ им. Ф. Скорины
2020

УДК 80:004.9(076)
ББК 80с515я73
Г646

Рецензенты:

доктор филологических наук О. А. Лещинская,
кандидат филологических наук М. М. Козловская

Рекомендовано к изданию научно-методическим советом учреждения образования «Гомельский государственный университет имени Франциска Скорины»

Гомонова, И. Г.

Г646 Компьютерная филология : практическое руководство / И. Г. Гомонова, И. В. Серикова ; Гомельский гос. ун-т им. Ф. Скорины. – Гомель : ГГУ им. Ф. Скорины, 2020. – 43 с.

ISBN 978-985-577-642-1

В практическое руководство вошли материалы обобщающего характера по основным темам учебных дисциплин направления «Компьютерное обеспечение», а также вопросы для самоконтроля, способствующие более качественному усвоению предложенной информации.

Издание рекомендуется студентам выпускного курса филологического факультета специальности 1-21 05 02-02 «Русская филология (компьютерное обеспечение)» для подготовки к государственному экзамену.

УДК 80:004.9(076)
ББК 80с515я73

ISBN 978-985-577-642-1

© Гомонова И. Г., Серикова И. В., 2020
© Учреждение образования «Гомельский государственный университет имени Франциска Скорины», 2020

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

ОГЛАВЛЕНИЕ

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

ПРЕДИСЛОВИЕ

Компьютерная филология − комплекс учебных дисциплин, изучаемых студентами в рамках направления специальности 1-21 05 02-02 «Русская филология (компьютерное обеспечение)», обеспечивающего получение квалификации «Филолог. Преподаватель русского языка и литературы. Специалист по компьютерной филологии». К стандартному набору дисциплин по направлению «Компьютерное обеспечение», предусмотренному для студентов университетов филологических специальностей, относятся «Введение в компьютерную филологию», «Инженерия знаний», «Методы автоматической обработки текстов», «Корпусная лингвистика», «Формализация языка в экспертных системах». Цель данных дисциплин − сформировать у студентов системное представление об использовании компьютерных технологий в профессиональной деятельности филолога.

При подготовке к государственному экзамену по специальности и направлению специальности компьютерная филология вызывает у студентов трудности, связанные с необходимостью обобщения и систематизации большого объема информации, включающей материал всех дисциплин направления «Компьютерное обеспечение».

Данное практическое руководство охватывает основные темы, связанные с подготовкой к итоговой аттестации студентов по специальности 1-21 05 02-02 «Русская филология (компьютерное обеспечение)», содержит краткие ответы на экзаменационные вопросы, а также задания для самоконтроля по каждой из выделенных тем, что способствует более качественному усвоению предложенной информации. Содержание предлагаемых в практическом руководстве ответов на вопросы направлено на закрепление сформированности знаний, умений и навыков.

Практическое руководство поможет студентам-выпускникам систематизировать и обобщить знания, приобретенные в процессе изучения дисциплин направления «Компьютерное обеспечение», сосредоточить свое внимание на основных понятиях, самостоятельно определить структуру ответов на экзаменационные вопросы.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

1. КОМПЬЮТЕРНАЯ ФИЛОЛОГИЯ КАК НАУЧНАЯ И УЧЕБНАЯ ДИСЦИПЛИНА. ОСНОВНЫЕ ТИПЫ ЭЛЕКТРОННЫХ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ

Компьютерная филология как научная и учебная дисциплина возникла на рубеже 50−60-х годов прошлого века. Ее предметом является вся сфера применения компьютерных моделей языка. Компьютерная филология рассматривается также как направление прикладной лингвистики, ориентированное на использование компьютерных инструментов для моделирования функционирования языка в тех или иных условиях.

Компьютерная филология как прикладная дисциплина выделяется прежде всего по инструменту, т. е. по использованию компьютерных средств обработки языковых данных.

Важнейшие направления компьютерной филологии:

− создание и использование систем обработки естественного языка (например, систем обработки связного текста);

− разработка и использование информационно-поисковых систем;

− создание гипертекстовых систем (множества текстов со связывающими их отношениями);

− разработка и использование компьютерных технологий составления и эксплуатации словарей.

− машинный перевод.

Направления компьютерной филологии соотносятся с базовыми функциями языка: коммуникативной (быть средством общения), когнитивной (служить средством формирования и выражения мысли, деятельности сознания), аккумулятивной (служить средством приобретения, накопления и сохранения знаний).

Коммуникативная функция оптимизируется в компьютерной филологии:

1) при создании диалоговых систем;

2) при создании систем обработки естественного языка;

3) при решении проблем машинного перевода и т. д.

Когнитивная функция оптимизируется в компьютерной филологии:

1) при реконструкции способов когнитивного моделирования и применении их в системах искусственного интеллекта;

2) при создании программ, позволяющих осуществлять моделирование и анализ текста.

Аккумулятивная функция оптимизируется в компьютерной филологии при создании баз данных: от электронных библиотек (баз данных текстов) до автоматизированных словарей (баз данных лексических единиц).

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Лингвистические ресурсы – это компьютерные средства поддержки работы лингвиста. Таких средств создано много, и они разнообразны. Это текстовые редакторы, звуковые анализаторы, компьютерные программы получения конкордансов, различные лингвистические базы данных с соответствующими средствами управления этими базами (например, электронные словари), корпуса текстов, компьютерные программы для обучения иностранным языкам и многое другое.

Типология лингвистических ресурсов может опираться на различные основания, но для лингвистов существенно рассмотрение ресурсов именно в лингвистическом аспекте. С лингвистической точки зрения различаются ресурсы лингвистических данных и средства обработки лингвистического материала, а среди ресурсов лингвистических данных – ресурсы первичных данных и ресурсы вторичных данных.

Вопросы для самоконтроля

1. Что является предметом компьютерной филологии?
2. Каковы важнейшие направления компьютерной филологии и как они соотносятся с базовыми функциями языка?
3. Как направления компьютерной филологии соотносятся с базовыми функциями языка?
4. Что такое электронные лингвистические ресурсы и на какие типы они делятся?

2. КОМПЬЮТЕРНАЯ ЛЕКСИКОГРАФИЯ. ЭЛЕКТРОННЫЕ СЛОВАРИ И ПРИНЦИПЫ ИХ ПОСТРОЕНИЯ

Компьютерная лексикография − прикладная научная дисциплина, которая изучает методы использования компьютерной техники для составления словарей, а также сами электронные словари. В рамках компьютерной лексикографии разрабатываются компьютерные технологии составления и эксплуатации словарей. Специальные программы − базы данных, компьютерные картотеки, программы обработки текста − позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и обрабатывать ее.

Множество различных компьютерных лексикографических программ делится на две большие группы: программы поддержки лексикографических работ и автоматические словари различных типов, включающие лек-

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

сикографические базы данных. Программы поддержки лексикографических работ – это компьютерные программы, призванные тем или иным образом облегчить труд лексикографа. Вместо обычной картотеки в компьютерных средах используются записи в базы данных.

Электронный словарь – это словарь в компьютере или другом электронном устройстве. С технической точки зрения это компьютерная база данных, которая содержит особым образом закодированные словарные статьи, позволяющие осуществлять быстрый поиск нужных слов (словосочетаний, фраз).

Электронные словари по адресату делятся на человекоориентированные и машиноориентированные, предназначенные для использования программами.

Автоматические словари, предназначенные для пользователя, по интерфейсу и структуре словарной статьи существенно отличаются от автоматических словарей, включенных в системы машинного перевода, системы автоматического реферирования, информационного поиска и т. д. Чаще всего они являются компьютерными версиями бумажных словарей.

Автоматические словари для программ обработки текста являются автоматическими словарями в полном смысле. Они, как правило, не предназначены для обычного пользователя, а особенности их структуры и сфера охвата словарного материала задаются теми программами, которые с ними взаимодействуют.

Электронные словари обладают рядом существенных преимуществ по сравнению с традиционными словарями: бóльший объем; автоматический поиск; скорость поиска; возможность поиска единицы в любой форме; одновременный поиск не только по названию словарной статьи, но и по всему объему словарей; возможность запоминать ранее открытые страницы и возвращаться к ним; возможность помещать нужные слова в «блокноты» или «ставить закладки»; возможность выполнять функцию текстового редактора; возможность выполнять функцию «гипертекст»; возможность звукового сопровождения.

Разрешение многозначности и снятие омонимии – главные проблемы компьютерной лексикографии.

Кроме того, электронные словари сохранили некоторые недостатки бумажных словарей. В основном это проблема неполноты словаря и поддержания его в актуальном состоянии, которая решается за счет привлечения большего числа сотрудников для пополнения словарей; предоставления возможности пользователям самим пополнять и редактировать словари; импорта данных из уже существующих словарей (в том числе из отсканированных копий бумажных словарей). К недостаткам можно отнести

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

и то, что многие словари требуют наличия определенной совокупности программных средств. Однако этот недостаток постепенно устраняется вследствие возрастающих темпов компьютеризации.

Компьютерная лексикография развивается двумя путями: оцифровывание традиционных словарей и создание специальных онлайновых словарей.

Вопросы для самоконтроля

1. Что изучает компьютерная лексикография как прикладная научная дисциплина?
2. Какие электронные ресурсы относятся к компьютерным технологиям составления и эксплуатации словарей?
3. Что представляет собой электронный словарь с технической точки зрения?
4. На какие типы делятся электронные словари по адресату?
5. В чем заключаются преимущества электронных словарей по сравнению с традиционными словарями?
6. Каковы главные проблемы компьютерной лексикографии?

3. АВТОМАТИЧЕСКОЕ ЧТЕНИЕ ТЕКСТА: ПОНЯТИЕ, ОСОБЕННОСТИ И ПРИМЕРЫ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Для быстрого и качественного ввода текстовой информации в компьютер широко используются сканеры. Для того чтобы «понять» содержание текста, т. е. перевести графическое изображение символов в пригодную для дальнейшей обработки (редактирования, реферирования, перевода и т. д.) текстовую форму, необходима система автоматического чтения текста или оптического распознавания символов (OCR-система – Optical Character Recognition).

Система автоматического чтения текста − это компьютерная программа, позволяющая преобразовать текст с бумажного носителя в электронный текстовый файл, который может быть прочитан средствами обработки текстов.

Основные принципы работы системы: целостность (объект описывается как целое с помощью значимых элементов и отношений между ними); целенаправленность (распознавание строится как процесс

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

выдвижения и целенаправленной проверки гипотез); адаптивность (способность компьютерной системы к самообучению).

В процессе сканирования и распознавания текста документа OCR-системы автоматически подбирают яркость сканирования, фрагментируют каждую страницу, выделяя в ней области графических иллюстраций и таблиц, распознают символы текста, проверяют орфографию распознанных слов и показывают окончательный результат в текстовом редакторе. OCR-системы позволяют распознавать печатные символы почти двух сотен языков. Хорошо распознаются рукопечатные символы, т. е. символы, написанные от руки печатными буквами с небольшим интервалом между ними.

Наряду со сплошными текстами (без таблиц и иллюстраций) программы автоматического чтения текста хорошо распознают: тексты с графикой, подписями, логотипами; таблицы; тексты, напечатанные на цветном фоне; тексты разноформатных документов (например, чертежей). OCR-системы узнают все используемые в тексте документа шрифты без предварительного обучения, хорошо воспринимают полужирный, курсивный, слипшийся, подчеркнутый и многоколоночный текст.

Изначально в мире преобладали системы автоматического чтения текста, требующие обучения каждому новому шрифту (новой гарнитуре, стилю, размеру и т. д.). Такие системы называются мультифонтовыми (от англ. *font* − шрифт). Противоположным классом OCR-систем являются омнифонтовые программы. Их не нужно обучать, эти программы распознают разные стилевые начертания одной и той же буквы, так как знают топологию (правила начертания) этой буквы.

OCR-системы поддерживают все модели сканеров и любые графические форматы. Точность распознавания OCR-систем на текстах хорошего и среднего качества достигает 97−99 %.

Развитие программ автоматического чтения текстов идет в направлении повышения точности распознавания текстов; выделения текстовой информации на фоне шумов, а также интеграции OCR-систем с различными программами обработки информации (системами машинного перевода, системами автоматического аннотирования и реферирования текстов, электронными архивами и т. д.).

Вопросы для самоконтроля

1. Что представляет собой система автоматического чтения текста (OCR-система)?
2. Каковы основные принципы работы OCR-системы?

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

3. Каковы возможности системы автоматического чтения текста?

4. В каком направлении идет развитие программ автоматического чтения текстов?

4. АВТОМАТИЗАЦИЯ ПРОЦЕССОВ АННОТИРОВАНИЯ И РЕФЕРИРОВАНИЯ ТЕКСТОВ

Рефераты и аннотации представляют собой вторичные документы, содержащие сведения о первичных (первичные документы − это книги, статьи, патенты и т. п.). Сущность аннотирования и реферирования заключается в максимальном сокращении объема источника информации при существенном сохранении его основного содержания. Аннотация служит только для информирования о существовании документа определенного содержания и характера. Реферат служит для передачи основного содержания документа.

Составление реферата или аннотации текста с помощью компьютерных программ называется автоматическим реферированием или аннотированием.

При выполнении работы по составлению реферата или аннотации человеком обычно выделяют три этапа: подготовительный (определяется тематическая направленность текста и осмысливается документ в целом); аналитический (текст делится на фрагменты; в каждом фрагменте выделяются основные смысловые единицы (предложения, словосочетания, слова); составляется план будущих реферата или аннотации); этап непосредственного построения реферата или аннотации (из выделенных смысловых единиц составляется вторичный текст в соответствии с планом).

В качестве основных смысловых единиц, выделяемых из исходного текста на втором этапе, выступают ключевые слова – слова, относящиеся к основному содержанию текста и повторяющиеся в нем несколько раз (с учетом всех возможных синонимов); ключевые словосочетания − сочетания слов, среди которых есть одно или несколько ключевых; ключевые предложения − предложения, содержащие два и более ключевых слов или словосочетаний.

Если поручить составление реферата или аннотации компьютеру, то его надо научить выполнять те же действия, которые осуществляет человек. Компьютер должен уметь: находить в тексте ключевые слова, словосочетания и предложения; находить в тексте менее значимые единицы; составлять из текстовых единиц двух первых типов смысловые единицы

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

реферата или аннотации; составлять из таких единиц текст реферата или аннотации.

Практически во всех существующих системах автоматического реферирования в качестве основных смысловых единиц реферата выступают ключевые предложения или ключевые словосочетания и слова исходного текста. Первые в том порядке, в котором они идут в исходном тексте, образуют текст реферата. Второй тип смысловых единиц (ключевые словосочетания и слова) используется компьютером для построения табличных рефератов.

При составлении с помощью компьютера аннотации также используются как ключевые предложения (в том виде, что и при составлении реферата), так и ключевые слова и словосочетания. Последние перечисляются вслед за реляторами типа «В статье рассматриваются следующие вопросы...», «Книга посвящена следующим проблемам...», «Статья раскрывает следующие понятия...» и т. д.

По способам выделения из исходных текстов ключевых единиц различают несколько методов автоматического реферирования и аннотирования текстов. Наиболее известны следующие три группы методов: статистические, позиционные, логико-семантические.

При статистических методах ключевыми словами считаются такие знаменательные слова текста, которые с учетом всех синонимов встречаются в тексте наибольшее число раз; ключевым предложением считается предложение текста, которое имеет несколько ключевых слов и содержит ключевые слова на небольшом расстоянии друг от друга. Принадлежность слова, словосочетания или предложения к числу ключевых определяется специальными статистическими коэффициентами.

При позиционных методах ключевым предложением считается предложение, входящее в заголовок, подзаголовок, начало или конец какой-то части текста или всего текста. Такие предложения, как правило, содержат существенную информацию о целях, выводах и результатах исследования, описанного в первичном документе.

Логико-семантические методы опираются на исследование структуры и семантики текстов. Существует несколько вариантов этих методов, но цель их одна − выделить из конкретного текста предложения с наибольшим функциональным весом, который зависит от многих факторов: наличия в исследуемом предложении семантически значимых слов, связи этого предложения с другими предложениями текста, синтаксического типа самого предложения и т. д.

Современные системы автоматического аннотирования и реферирования основаны на статистических методах.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Вопросы для самоконтроля

1. Что представляют собой процессы автоматического реферирования и аннотирования?
2. Каковы этапы работы по составлению реферата и аннотации?
3. Какие текстовые единицы выступают в качестве основных смысловых единиц реферата и аннотации?
4. На какие типы делятся методы автоматического реферирования и аннотирования по способам выделения из исходных текстов ключевых единиц?
5. В чем сущность каждого метода?
6. Какой метод является основным для современных систем автоматического аннотирования и реферирования?

5. МАШИННЫЙ ПЕРЕВОД ТЕКСТОВ

Машинный перевод − процесс перевода текстов с одного естественного языка на другой с помощью специальной компьютерной программы, а также направление научных исследований, связанных с построением таких систем.

Первая публичная демонстрация машинного перевода (Джорджтаунский эксперимент) состоялась в 1954 году. Несмотря на примитивность системы (словарь в 250 слов, грамматика из 6 правил, перевод нескольких простых фраз), эксперимент получил широкий резонанс: в разных странах начались исследования в области машинного перевода. Однако в то время он не получил широкого распространения из-за недостаточного развития компьютерных технологий. В начале 80-х годов в связи с широкой компьютеризацией машинный перевод стал экономически выгодным, наступило время широкого практического использования переводческих систем, сложился рынок коммерческих разработок по этой теме. В эти и последующие годы стали более совершенными программы компьютерного перевода.

Однако некоторые проблемы машинного перевода не решены до сих пор. Трудности связаны с особенностями функционирования языка: не всегда учитываются значения, которые может иметь слово в разных стилях речи; возникают ошибки в переводе слов в составе устойчивых сочетаний; не учитываются дополнительные смыслы, которые возникают при изменении порядка слов; не определяется изменение значения слова в контексте и др.

Репозиторий ГГУ имени Ф. Скорины

Современные компьютерные переводные программы постоянно совершенствуются: включают в себя элементы искусственного интеллекта и средства редактирования текстов.

В процессе машинного перевода текста выполняется следующая последовательность действий: морфологический анализ каждого слова предложения исходного языка; синтаксический анализ каждого предложения текста исходного языка; синтаксический синтез каждого предложения переводного языка; морфологический синтез каждого слова предложения переводного языка.

В ходе морфологического анализа слов предложения исходного языка каждое слово получает наборы лексико-грамматических признаков (часть речи, род, число, падеж, время, лицо и т. д.). Компьютер может сформировать такие наборы либо по формальным признакам слов, либо с опорой на специальный автоматический словарь. В нем каждой словоформе приписаны соответствующие лексико-грамматические признаки, и в процессе морфологического анализа слова компьютер берет их из словаря в готовом виде.

Синтаксический анализ предложения исходного языка сводится к поиску основных членов предложения. Синтаксический синтез предложения переводного языка заключается в создании предложения синтаксической структуры, определяемой правилами переводного языка и синтаксической структурой предложения исходного языка. Чтобы компьютер мог выполнить это задание, он должен иметь в памяти сведения о синтаксических структурах исходного языка, переводного языка и их соответствиях друг другу. Еще одна задача этапа синтаксического синтеза связана с заменой слов исходного языка их переводными эквивалентами из словаря переводного языка.

Морфологический синтез каждого слова предложения переводного языка сводится к постановке слов в нужной форме. Для этого компьютер должен владеть информацией о лексико-грамматических признаках каждого слова переводного языка, которая берется из автоматического словаря.

Эффективность системы машинного перевода во многом определяет автоматический словарь этой системы.

Существуют два принципиально разных подхода к построению алгоритмов машинного перевода: основанный на правилах (rule-based) и статистический, или основанный на статистике (statistical-based).

Первый подход является традиционным и используется большинством разработчиков систем машинного перевода (например, ПРОМТ). Такие системы состоят из двуязычных словарей и грамматик, охватывающих основные семантические, морфологические, синтаксические законо-

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

мерности каждого языка. На основе этих данных исходный текст последовательно, по предложениям, преобразуется в текст перевода.

Ко второму типу относятся популярные сервисы Яндекс.Перевод, Переводчик Google и др. Статистический машинный перевод основан на сравнении больших объемов языковых пар. Языковые пары − тексты, содержащие предложения на одном языке и соответствующие им предложения на втором, которые могут быть как вариантами написания двух предложений человеком − носителем двух языков, так и набором предложений и их переводов, выполненных человеком. Чем больше языковых пар и чем точнее они соответствуют друг другу, тем лучше результат статистического машинного перевода. Статистический машинный перевод обладает свойством «самообучения».

Качество перевода зависит от тематики и стиля исходного текста, а также грамматической и лексической родственности исходного и переводного языков. Так, машинный перевод художественных текстов практически всегда оказывается неудовлетворительного качества. При переводе технических текстов, если правильно выбрать словарь по специальности, к которой относится текст, получается вполне удовлетворительный результат. Этот перевод нуждается лишь в небольшой редакторской корректировке. Чем более формализован стиль исходного документа, тем лучшего качества перевода можно ожидать.

Вопросы для самоконтроля

1. В чем заключается сущность процесса машинного перевода?
2. Какова последовательность действий в процессе машинного перевода текста?
3. В чем состоит особенность каждого этапа данного процесса?
4. Какие существуют подходы к построению алгоритмов машинного перевода?
5. С чем связаны трудности машинного перевода и от чего зависит его качество?

6. ПРЕДМЕТ КОРПУСНОЙ ЛИНГВИСТИКИ. ИСТОРИЯ СОЗДАНИЯ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ

Корпусная лингвистика – раздел лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

корпусов (корпусов текстов) с применением компьютерных технологий. В отличие от других разделов науки о языке корпусная лингвистика является более широким понятием, методологией, которую можно применить ко многим аспектам языковых исследований.

Предмет корпусной лингвистики − теоретические основы и практические механизмы создания и использования корпусов текстов. Двойственный характер корпусной лингвистики (нацеленность как на создание, так и на использование корпусов текстов) обусловливается двойственным характером ее объекта – корпуса текстов, который, с одной стороны, представляет собой исходный материал для корпусной лингвистики; с другой стороны, является результатом деятельности корпусной лингвистики.

Под лингвистическим корпусом понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач; совокупность текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой.

Система управления текстовыми и лингвистическими данными корпуса называется корпусным менеджером. Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления пользователю результатов в удобной форме (в виде конкорданса – списка всех фиксаций искомой языковой единицы в контекстах со ссылками на источник).

Корпуса текстов используются для получения разнообразных справок и статистических данных о языковых единицах. В частности, на основе корпусов можно получить данные о частоте словоформ, лексем, грамматических категорий; проследить изменение частот в различные периоды; получить данные о совместной встречаемости лексических единиц; изучать динамику процессов изменения лексического состава языка; проводить анализ лексико-грамматических характеристик в разных жанрах и у разных авторов и т. д. Корпуса служат также источником и инструментом многоаспектных лексикографических работ. Данные корпусов используются для создания и уточнения грамматик и в целях обучения языку.

Первые корпуса текстов были созданы в 1960-е годы. Брауновский корпус (1963 г., Брауновский университет (США)) включает 500 текстов из американских книг, газет, журналов, впервые опубликованных в США в 1961 году. Каждый текст имеет объем 2 000 словоупотреблений, и все собрание включает 1 млн слов. Тексты в Брауновском корпусе принадлежат 15 наиболее массовым жанрам англоязычной печатной прозы США. Корпус сопровождается большим количеством материалов статистической обработки

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

(например, частотным и алфавитно-частотным словарём). Цель создания Брауновского корпуса – обеспечить системное изучение отдельных жанров письменного английского языка. Появление Брауновского корпуса вызвало всеобщий интерес и оживленные дискуссии (по поводу принципов отбора текстов и состава потенциально решаемых задач).

Позднее европейские исследователи составили корпус текстов, впервые опубликованных в Великобритании в 1961 году, следуя тем же принципам: 15 жанров, 500 текстов по 2 000 словоупотреблений. Корпус Ланкастер-Осло-Берген (по названиям британского и двух норвежских университетов) включает 1 млн слов британского варианта английского языка.

Таким образом, два самых ранних больших корпуса – это корпуса письменной речи американского и британского вариантов английского языка. Оба корпуса остаются полезными и сейчас, на них основываются многочисленные исследования английского языка.

Брауновский корпус задал стандарт в 1 млн словоупотреблений для создания корпусов текстов на других языках. По аналогичной модели был построен и первый русский корпус, созданный в 1980-е годы в Университете Уппсалы, Швеция − Уппсальский корпус русского языка.

Однако размер в 1 млн слов достаточен для лексикографического описания только самых частотных слов. По этой причине, а также в связи с ростом компьютерных мощностей, способных работать с большими объемами текстов, в 1980-е годы в мире было предпринято несколько попыток создать корпуса большего размера. В Великобритании такими проектами были лингвистический Банк английского языка в Бирмингемском Университете, Британский Национальный Корпус, Международный корпус английского языка, Корпус современного американского английского. В СССР таким проектом был Машинный Фонд русского языка, создававшийся по инициативе А. П. Ершова.

К 1990 году было зафиксировано уже более 600 компьютерных корпусов. В последующие годы количество и многообразие создаваемых корпусов росли. В настоящее время корпуса созданы для многих языков мира. Некоторые из них содержат миллиарды словоупотреблений.

В первой половине 1990-х годов корпусная лингвистика окончательно сформировалась как отдельное направление науки о языке.

Вопросы для самоконтроля

1. Что является предметом корпусной лингвистики?
2. Что такое лингвистический корпус?
3. Каковы задачи, решаемые с помощью лингвистических корпусов?

4. Когда и где были созданы первые корпуса текстов? Каковы их основные характеристики?

5. Когда корпусная лингвистика сформировалась как отдельный раздел науки о языке?

7. КОРПУС ТЕКСТОВ КАК ОСОБЫЙ ЛИНГВИСТИЧЕСКИЙ РЕСУРС. ТИПОЛОГИЯ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ

Лингвистический корпус – это совокупность текстов, собранных в соответствии с определенными принципами, представленных в электронном виде, унифицированных, размеченных по определенному стандарту, обеспеченных специализированной поисковой системой. Основными характеристиками корпуса текстов являются размер, репрезентативность и разметка.

Первые корпуса состояли из 1 млн словоупотреблений, однако такой объем не позволяет отражать язык во всем его многообразии. В настоящее время считается, что общеязыковой (национальный) корпус должен включать не менее 100 млн словоупотреблений. Объем современных корпусов исчисляется сотнями миллионов или миллиардами.

С течением времени объем корпуса может меняться, однако эти изменения должны либо не менять репрезентативность корпуса, либо менять ее обоснованно. Репрезентативность − это представительность корпуса, пропорциональное соотношение его отдельных частей (по разным характеристикам – время, жанры, стили, авторы и др.). Репрезентативность корпуса определяет достоверность полученных на его материале результатов.

С точки зрения репрезентативности корпуса текстов делятся на два типа: корпуса первого типа универсальны, они отражают все многообразие речевой деятельности; корпуса второго типа отражают бытование определенного лингвистического или культурного феномена в общественной речевой практике, критерий отбора текстов для такого корпуса задает его создатель, исходя из целей своей практической или научной деятельности.

С точки зрения отбора текстов в корпус различают сбалансированные и мониторные корпуса. В сбалансированные корпуса в соответствии с установленными пропорциями включаются разнообразные по жанрам, стилям и тематике тексты. Пополнение таких корпусов происходит только после тщательной процедуры отбора новых текстов. Мониторные корпуса постоянно пополняются новыми текстами на данном языке, при этом

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

баланс текстов разных стилей и жанров не соблюдается. Создатели мониторных корпусов считают, что статистическая обоснованность данных, полученных из корпуса, будет достигнута за счет его объема, исчисляемого миллиардами.

Для решения различных лингвистических задач мало наличия массива текстов. Тексты должны содержать дополнительную лингвистическую и экстралингвистическую информацию − разметку. Набор этих данных во многом определяет возможности, предоставляемые корпусами исследователям.

Разметка заключается в приписывании текстам и их компонентам специальных тэгов: собственно лингвистических, описывающих лексические и грамматические характеристики элементов текста, и экстралингвистических (сведения об авторе и о тексте).

Существует большое число разных типов лингвистических корпусов, что определяется многообразием исследовательских и прикладных задач, для решения которых они создаются, и различными основаниями для классификации.

По цели создания корпуса делятся на многоцелевые и специализированные. По назначению выделяют исследовательские и иллюстративные корпуса. По типу языковых данных корпуса делятся на письменные, устные и смешанные. По критерию параллельности − на одноязычные, двуязычные и многоязычные.

Важным критерием для пользователей корпуса является его доступность. В этом отношении корпуса бывают свободно доступными, коммерческими и закрытыми.

Несмотря на разнообразие корпусов текстов, есть два основных способа деления их на типы: противопоставление корпусов, относящихся ко всему языку, корпусам, относящимся к какому-либо подъязыку, и деление корпусов по виду лингвистической разметки.

Вопросы для самоконтроля

1. Что представляет собой корпус текстов как особый тип электронного ресурса?
2. Каковы основные характеристики корпуса текстов?
3. Что такое репрезентативность корпуса?
4. На какие типы делятся корпуса с точки зрения репрезентативности и отбора текстов?
5. В чем заключается сущность разметки корпуса?
6. Каковы основные способы деления корпусов текстов на типы?

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

8. ПОНЯТИЕ РАЗМЕТКИ. ВИДЫ РАЗМЕТКИ В КОРПУСЕ ТЕКСТОВ

Под разметкой корпуса текстов понимается наличие в корпусе данных, не являющихся частью текста, но несущих информацию о нем.

Процесс разметки состоит в приписывании текстам и их компонентам специальных тэгов: экстралингвистических и собственно лингвистических. Соответственно, в корпусах текстов представлена экстралингвистическая и лингвистическая разметка. Экстралингвистическая разметка (метаразметка) включает в себя сведения об авторе и о тексте. Собственно лингвистическая разметка содержит морфологические, синтаксические, семантические и другие характеристики элементов текста.

В 1980-е годы был принят стандарт разметки электронных текстов под названием SGML (Standard Generalized Markup Language). Он был разработан внутри типографской индустрии, но быстро распространился на другие отрасли. Смысл SGML в том, что документы, набранные в разных текстовых процессорах, можно редактировать, анализировать и изменять в любом из них.

SGML ввел концепцию тэгов. Тэги (от англ. *tag* – ярлык, метка) − это служебные пометки в тексте, содержащие информацию о самом тексте. Для каждого случая можно определять собственные тэги и таким образом создавать диалекты языка SGML. Программа, отображающая размеченный текст, интерпретирует тэги в соответствии с заложенными в нее правилами и показывает пользователю текст, оформленный согласно им.

Язык разметки SGML очень сложен и используется довольно редко. Но он послужил базой для создания таких широко известных языков разметки, как HTML и XML.

Язык HTML (Hyper-Text Markup Language) создан из SGML путем выделения ограниченного набора тэгов, в основном относящихся к оформлению, а не к содержанию документа. Второе широко известное подмножество SGML − расширяемый язык разметки XML (eXtensible Markup Language), который применяется для хранения любых структурированных данных, в том числе и текстов в корпусах. XML − это свод синтаксических правил для описания структуры данных.

В корпусной лингвистике большое внимание уделяется стандартизации разметки. Специально для разметки корпусов текстов в 1987−1989 годах была разработана система, описывающая, какие именно параметры текстов нужно размечать. Эта система использует XML и называется Text Encoding Initiative Guidelines (TEI Guidelines, Инициатива по Кодированию Текстов). TEI – один из самых известных стандартов разметки тек-

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

стов в формате XML, широко используемый в проектах по созданию электронных коллекций для гуманитарных наук.

Преимущества разметки в стандарте TEI: опора на тщательно разработанную теорию структуры текста; легкость адаптации к конкретному материалу; независимость от конкретного программного продукта. В настоящее время практически все проекты по созданию корпусов в той или иной мере следуют рекомендациям TEI.

Вопросы для самоконтроля

1. Что понимается под разметкой корпуса текстов?
2. Какие виды разметки представлены в корпусах текстов?
3. Какой язык разметки применяется для описания текстов и их компонентов в корпусах?
4. Как называется стандарт разметки текстов в формате XML, широко используемый в корпусах, и каковы его преимущества?

9. МЕТАРАЗМЕТКА И ЕЕ ФУНКЦИИ

Значимой частью поискового аппарата корпуса является метаразметка текстов, входящих в него.

Метаразметка включает: «внешнюю» разметку, содержащую библиографические, типологические, тематические и социологические характеристики; формальную (структурную) разметку, отражающую деление текста на разделы, главы, части, абзацы, предложения; технико-технологическую разметку, содержащую кодировку, даты обработки, источник электронной версии и т. п.

Метаразметка выполняет несколько функций: служит для создания архитектуры корпуса; позволяет контролировать процесс наполнения корпуса; позволяет установить корреляцию между метатекстовыми параметрами и языковыми особенностями текста; обеспечивает возможность поиска и отбора текстов пользователем для составления подкорпусов по заданным параметрам (основная).

Чем больше набор параметров, по которым характеризуется каждый текст, тем шире возможности, предоставляемые корпусами исследователям.

К метаразметке корпуса предъявляется требование унификации, обусловленное следующими факторами: многократное использование многими пользователями; совместимость с другими корпусами; совместимость

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

с общепринятыми научными теориями и классификациями; возможность применения стандартных программных средств.

При осуществлении метаразметки учитывают два класса факторов, влияющих на язык текстов (по классификации Дж. Синклера): внешние, внеязыковые факторы (E-external) и внутренние факторы (I-internal).

Дж. Синклер выделяет три группы E-факторов: E1 (origin) – факторы, относящиеся к созданию текста автором; E2 (state) – факторы, относящиеся к внешним признакам текста; E3 (aims) – факторы, относящиеся к причинам создания текста и его влиянию на аудиторию, и две группы I-факторов: I1 (topic) – предметная область текста; I2 (style) – стилистические особенности текста. Эта классификация основана преимущественно на логических свойствах коммуникации и поэтому может быть применена к описанию дискурса на любом языке.

При разработке базы метаданных Национального корпуса русского языка (НКРЯ) учитывался как зарубежный опыт создания корпусов, так и принципы описания, разработанные в отечественной типологии текстов и лингвостилистике. В основу метаразметки текстов НКРЯ положен стандарт EAGLES, принятый во многих современных системах автоматической обработки текстов (рекомендации Дж. Синклера). Эти рекомендации были адаптированы к русскому материалу С. А. Шаровым и составили вариант метаразметки под названием «вариант Синклера-Шарова». Окончательный вариант метаразметки вырабатывался с учетом опыта отечественной стилистики и типологии текстов, а также возможных запросов будущих пользователей корпуса, в результате чего список обязательных параметров описания текстов был расширен за счет стилистических категорий.

В НКРЯ используется сравнительно простая система метаразметки, предназначенная не для специалистов по корпусной лингвистике, работающих с универсальной международной классификацией, а для рядового пользователя. При метаразметке каждый текст описывается по 24 параметрам: 9 параметров характеризуют сам текст (название; дата создания; размер в словах; сфера функционирования; тема, к которой можно отнести содержание текста (для нехудожественных текстов); хронотоп, или место и время описываемых событий (для художественных текстов и мемуаров); тип текста; жанр художественной литературы; стиль текста); 3 параметра характеризуют автора (имя, пол, дата рождения); 3 параметра характеризуют возможную аудиторию (возраст, уровень образования, размер аудитории); 4 параметра содержат библиографические данные о тексте; 5 параметров представляют служебную информацию, необходимую для учета и организации текстовых файлов в составе корпуса.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Вопросы для самоконтроля

1. Что включает в себя метаразметка корпуса?
2. Какие функции выполняет метаразметка?
3. Чем обусловлено требование унификации, предъявляемое к метаразметке корпуса?
4. Какие два класса факторов учитывают при осуществлении метаразметки?
5. Какие метапараметры включает в себя экстралингвистическая разметка Национального корпуса русского языка?

10. ОСНОВНЫЕ ПРИНЦИПЫ И РАЗНОВИДНОСТИ ЛИНГВИСТИЧЕСКОЙ РАЗМЕТКИ

Лингвистическая разметка подразумевает присвоение словам особых кодов, каждому из которых соответствует определенный набор признаков, характеризующих данное слово. Эти коды называются тэгами (от англ. *tag* – ярлык, метка), а сам процесс приписывания словам тэгов называется тэггингом (от англ. *tagging*).

Выделяются следующие разновидности лингвистической разметки: морфологическая, синтаксическая, семантическая, словообразовательная и др. Все они осуществляются в соответствии с определенными принципами: схема разметки должна быть обоснованной и теоретически нейтральной (традиционной); система лингвистических понятий должна быть общепринятой; введение параметров должно быть мотивированным; разметка должна соответствовать международным стандартам.

Схема разметки предполагает наличие, во-первых, набора тэгов; во-вторых, описания того, что каждый из них означает; в-третьих, правил присвоения тэгов единицам текста.

Морфологическая разметка (англ. *part-of-speech tagging* – частеречная разметка) – основная разновидность лингвистической разметки. Это подтверждается тем, что большинство крупных корпусов являются морфологически размеченными; морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического; успехи в компьютерной морфологии позволяют автоматически с большой степенью правильности размечать корпуса больших размеров. Морфологическая разметка включает лемму, признаки части речи, признаки грамматических категорий. Морфологический стандарт Национального корпуса русского языка опирается на морфологическую модель, представлен-

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

ную в «Грамматическом словаре русского языка» А. А. Зализняка, с отступлениями, продиктованными спецификой корпуса.

Синтаксическая разметка является результатом синтаксического анализа (парсинга), выполняемого на основе данных морфологического анализа. Синтаксическая разметка описывает синтаксические конструкции: синтаксические связи и отношения. В отличие от морфологии, способы представления синтаксической структуры и синтаксических отношений не столь унифицированы, наблюдается разнообразие синтаксических теорий. На синтаксическом уровне, как и на морфологическом, проявляется тенденция к меньшей детализации разметки для увеличения скорости и последовательности анализа текста.

Существующие немногочисленные корпуса с синтаксической разметкой опираются либо на общепринятые классификации традиционной грамматики (например, разметка в Хельсинкском аннотированном корпусе русских текстов), либо на доступные узкому кругу специалистов и требующие детального предварительного знакомства классификации (например, разметка в терминах деревьев зависимостей и синтаксических отношений в Национальном корпусе русского языка).

Проблемы семантической разметки обусловлены тем, что, в отличие от грамматических классов, семантические классы не имеют формальных показателей; грамматические классы общеприняты, а общепринятой семантической классификации нет. Семантические тэги обозначают семантические категории, к которым относится данное слово, и более узкие подкатегории, специфицирующие его значение. В процессе словообразовательной разметки каждому слову приписывается последовательность словообразовательных тэгов, отражающих его морфемное членение.

Число тэгов, применяемых в разных корпусах, варьируется. Чем больше набор тэгов, тем более детальный анализ текста осуществим с его помощью. Однако по мере увеличения объема корпусов наметилась тенденция к сокращению числа тэгов. Упрощенная система кодировки помогает избежать ошибки, устранить непоследовательность, уйти от неоднозначности и ускорить разметку больших массивов текста, содержащих миллионы слов.

Вопросы для самоконтроля

1. Какие разновидности лингвистической разметки представлены в корпусах текстов?
2. Каковы основные принципы лингвистической разметки?
3. Что включает в себя схема разметки?

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

4. Почему морфологическая разметка является основной разновидностью лингвистической разметки в корпусах текстов?

5. В чем заключается специфика синтаксической разметки?

6. Чем обусловлены проблемы семантической разметки?

7. Какая тенденция в развитии разметки корпусов наблюдается в настоящее время?

11. ПОНЯТИЕ НАЦИОНАЛЬНОГО КОРПУСА. ИНТЕРНЕТ КАК КОРПУС

Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всем многообразии жанров, стилей, территориальных и социальных вариантов и т. п.

Национальный корпус имеет две важные особенности. Во-первых, он характеризуется репрезентативностью (представительностью, или сбалансированным составом текстов). Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Хорошая представительность достигается только при значительном объеме корпуса (десятки и сотни миллионов словоупотреблений).

Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов − разметку. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. Национальный корпус предназначен в первую очередь для обеспечения научных исследований лексики и грамматики языка, а также процессов языковых изменений, происходящих в языке на протяжении сравнительно небольших периодов − от одного до двух столетий. Другая задача корпуса – предоставление всевозможных справок, относящихся к лексике, грамматике, акцентологии, истории языка.

Большинство крупных языков мира имеет свои национальные корпуса (различающиеся по полноте и уровню научной обработки текстов).

Общепризнанным образцом является, в частности, Британский национальный корпус (BNC – British National Corpus); на него ориентированы многие современные корпуса. Объем корпуса − 100 млн слов. Разработан BNC в первой половине 90-х годов XX века в Оксфордском университете при участии Ланкастерского университета и Британской библиотеки. Бри-

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

танский национальный корпус включает метатекстовую и морфологическую разметку. В корпусе представлены письменная речь – 90 %, включающая самые разнообразные по жанру, стилю и тематике тексты, и устная речь – 10 %. Подкорпус устной речи включает в себя речь добровольно вызвавшихся участвовать в проекте людей различных возрастов, проживающих в разных частях Великобритании и принадлежащих к различным социальным классам. BNC − это синхронный корпус общего назначения. Он отражает состояние британского варианта английского языка конца XX − начала XXI века.

Одним из наиболее известных национальных корпусов славянских языков является Чешский национальный корпус (Český národní korpus), созданный в середине 1990-х годов в Карловом университете Праги и представляющий собой сбалансированный представительный корпус. Он включает в себя несколько независимых корпусов, из которых свободно доступен в сети Интернет корпус PUBLIC (объем − 20 млн слов). Чешский национальный корпус имеет метаразметку и морфологическую разметку. В корпусе возможен поиск по словоформе, по лемме, по грамматической информации. Доступна информация о частотном распределении языковых единиц, а также информация о коллокациях. Массив текстов в корпусе делится на синхроническую и диахроническую части. Синхроническая часть включает в себя письменные тексты, разговорные тексты и диалектную речь.

Национальный корпус русского языка (НКРЯ), созданный в начале XXI века (свободный доступ к корпусу в сети Интернет открыт в 2004 г.), − сбалансированный представительный корпус, имеющий в настоящее время общий объем более 600 млн слов, снабженный метаразметкой и лингвистической разметкой разных типов. НКРЯ охватывает прежде всего период от середины XVIII до начала XXI века: этот период представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах –литературном, просторечном, диалектном. В корпус входят произведения разных стилей и жанров.

В настоящее время НКРЯ включает следующие подкорпуса: основной корпус, в который входят прозаические письменные тексты XVIII − начала XXI века; синтаксический корпус, в котором для каждого предложения построена полная морфологическая и синтаксическая структура (дерево зависимостей); газетный корпус, в котором представлены статьи из средств массовой информации 1990−2000-х годов; параллельные корпуса, в которых текстам на одном языке сопоставлен перевод этих текстов на другой язык; корпус диалектных текстов; корпус поэтических текстов, в котором возможен поиск не только по лексическим и грамматическим, но и по специфическим для стиха признакам; обучающий корпус, разметка которого ориентирована на школьную программу русского языка; кор-

пус устной речи; акцентологический корпус (корпус истории русского ударения); мультимедийный корпус, куда входят снабженные видео- и аудиорядом фрагменты кинофильмов 1930−2000-х годов; исторический корпус. НКРЯ предоставляет пользователю широкий круг поисковых возможностей.

В качестве огромного многоязычного корпуса может рассматриваться сеть Интернет (веб-пространство). Язык представлен в Интернете в большом объеме и разнообразии и непосредственно доступен для машинной обработки. Ни один корпус текстов не может сравниться по размеру и репрезентативности языкового материала с веб-пространством. При этом встает вопрос о сбалансированности веб-корпуса, так как в Интернете определенные типы речевых произведений представлены чаще, чем в языке вообще. При использовании веб-пространства как корпуса роль корпусных менеджеров выполняют поисковые системы. Основным средством поиска информации в сети в настоящее время являются глобальные поисковые системы вербального типа, индексирующие все Интернет-пространство (например, Google, Fast Search (AllTheWeb), AltaVista, Yandex, Rambler и др.). Индексы вербальных систем – это своего рода конкордансы к текстам.

Использование Интернета как корпуса ограничено изучением лексического материала (в этой сфере возможности очень велики). Грамматические исследования на базе Интернета минимальны.

Вопросы для самоконтроля

1. Какими особенностями характеризуется национальный корпус?
2. Для решения каких основных задач предназначен национальный корпус?
3. Приведите краткую характеристику двух–трех национальных корпусов.
4. Какими преимуществами и недостатками характеризуется веб как корпус?

12. МОДЕЛИРОВАНИЕ. ОСНОВНЫЕ ЭТАПЫ МОДЕЛИРОВАНИЯ

Моделирование – это способ познания действительности, используемый различными науками. К моделированию обращаются лингвисты, желая понять систему языка. Свои мысленные представления о системе языка они овеществляют в схемах, чертежах, графиках, диаграммах разного рода.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Основные цели моделирования:

− понять, как устроен конкретный объект, какова его структура, основные свойства, законы развития и взаимодействия с окружающим миром;

− научиться управлять объектом (процессом) и определить наилучшие способы управления при заданных целях и критериях;

− прогнозировать прямые и косвенные последствия реализации заданных способов и форм воздействия на объект.

Объект моделирования – широкое понятие, включающее объекты живой или неживой природы, процессы и явления действительности. Каждый объект, для которого создается модель, называют оригиналом или прототипом.

Любая модель не является абсолютной копией своего оригинала, она лишь отражает некоторые его качества и свойства, наиболее существенные для выбранной цели исследования. В процессе моделирования осуществляется выбор наиболее близкой к оригиналу модели и перенос результатов исследования на оригинал.

Познавая объективный мир, человек стремится рассмотреть изучаемый предмет снаружи и изнутри, потрогать его руками, а если нужно, то и послушать, понюхать, попробовать на вкус, словом, воспринять всеми пятью органами чувств. Но существует множество объектов внешнего мира, которые по своим размерам, расположению, характеру функционирования недоступны прямому восприятию человека. Моделирование (эксперимент) в таких случаях может быть незаменимо. Создавая мысленные образы таких предметов, догадываясь об их виде и устройстве, человек все так же стремится увидеть их, сделать обозримыми, находящимися в его поле зрения. Для этого он создает их вещественные копии – модели.

Процесс моделирования включает три элемента: субъект исследования; объект исследования; модель, отражающую отношения познающего субъекта и познаваемого объекта.

Первым этапом любого исследования является постановка задачи, которая определяется заданной целью. Решение любой практической задачи всегда связано с преобразованием некоторого объекта (материального или информационного) или управления им. На первом этапе построения модели предполагается наличие некоторых знаний об объекте-оригинале. Познавательные возможности модели обусловливаются тем, что модель отображает (воспроизводит, имитирует) какие-либо существенные черты объекта-оригинала.

Второй этап – анализ объекта. Результат анализа объекта – выявление его составляющих (элементарных объектов) и определение связей между ними. Простой пример подчиненных связей объектов − разбор

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

предложения. Сначала выделяются главные члены, затем второстепенные члены, относящиеся к главным, затем слова, относящиеся к второстепенным.

Третий этап – разработка информационной модели объекта. Построение модели должно быть связано с целью моделирования. Каждый объект имеет большое количество различных свойств. В процессе построения модели выделяются главные, наиболее существенные, свойства, состояния, действия и другие характеристики элементарных объектов в любой форме: устно, в виде схем, таблиц. Формируется представление об элементарных объектах, составляющих исходный объект, т. е. информационная модель. Например, строится схема предложения.

Для одного объекта один субъект может построить несколько моделей, если он решает разные задачи, приводящие к разным целям моделирования. Для одного объекта разные субъекты могут построить разные модели, даже если задача моделирования у них одна. Разные объекты могут иметь одинаковые по виду модели, даже если их строили разные субъекты исходя из разных целей моделирования.

Вопросы для самоконтроля

1. Что такое моделирование?
2. Каковы основные цели моделирования?
3. Что является объектом моделирования?
4. Какие элементы включает в себя процесс моделирования?
5. Каковы основные этапы моделирования?

13. ПОНЯТИЯ «МОДЕЛЬ» И «ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ». АЛГОРИТМ, ЗАДАЧА И МОДЕЛЬ

Модель (франц. *modèle*, от лат. *modulus* – мера) – это формализованное описание объекта, системы нескольких объектов, процесса или явления, выраженное конечным набором предложений какого-либо языка, математическими формулами, таблицами, графиками, специальными знаками или схемами. Модель должна отражать наиболее существенные черты реального объекта, процесса или явления, которые важны для проводимого в данный момент процесса моделирования.

Модель в языкознании:

1) искусственно созданное лингвистом реальное или мысленное устройство, воспроизводящее, имитирующее своим поведением (обычно

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

в упрощенном виде) поведение какого-либо другого устройства (оригинала) в лингвистических целях;

2) образец, служащий стандартом (эталоном) для массового воспроизведения; то же, что тип, схема, парадигма и т. п. (например, модель спряжения или склонения, словообразовательная модель, модель предложения и т. п.).

Понятие «лингвистическая модель» возникло в структурной лингвистике, но вошло в активный научный обиход в 60−70-е годы XX века с возникновением математической лингвистики и проникновением в языкознание идей и методов кибернетики.

Если к модели поставить вопрос и добавить дополнительные условия в виде исходных данных (связь с другими объектами, начальные условия, ограничения), то она трансформируется в задачу. Таким образом, модель + вопрос + дополнительные условия = задача. Чтобы решить задачу, необходимо знать ее начальные условия, а также метод или способ ее решения.

Пример задачи: найти среди приведенных русских глаголов те, которые употреблены в форме инфинитива. Для выделения из группы глаголов инфинитивных форм необходимо, чтобы среди анализируемых глаголов были эти формы (начальные условия). А способ решения сводится к следующей проверке: оканчивается ли соответствующий глагол на *-ть, -чь, -ти*.

Процесс решения задачи может быть представлен алгоритмом. Алгоритм − один из способов представления (отражения) мысли, процесса, явления в искусственной вычислительной среде, которой является компьютер. Специфика алгоритма состоит в отражении последовательности действий. Примеры алгоритмов в природе неизвестны, они порождение человеческого разума, способного к установлению плана. Собственно алгоритм − это и есть план, развернутый в последовательность действий.

Какова разница между алгоритмом и моделью? Алгоритм − это процесс решения задачи путем реализации последовательности шагов, модель − совокупность потенциальных свойств объекта.

Вопросы для самоконтроля

1. Что такое модель?
2. В чем заключается сущность понятия «лингвистическая модель»?
3. Что такое задача и что необходимо знать для ее решения?
4. В чем заключается разница между алгоритмом и моделью?

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

14. ВИДЫ ИНФОРМАЦИОННЫХ МОДЕЛЕЙ

Информационная модель – это совокупность информации об объекте, описывающая свойства и состояние объекта, процесса или явления, а также связи и отношения объекта с окружающим миром.

Информационные модели отражают разные типы систем объектов, в которых реализуются различные структуры взаимодействия и взаимосвязи между элементами системы. Для отражения систем с различными структурами используются следующие виды информационных моделей: табличные, иерархические и сетевые.

В табличной информационной модели перечень однотипных объектов или свойств размещен в первом столбце (или строке) таблицы, а значения их свойств размещаются в следующих столбцах (или строках) таблицы. Представление объектов и их свойств в форме таблицы часто используется в научных исследованиях. Табличные информационные модели проще всего строить на компьютере с помощью электронных таблиц (приложение Excel) и систем управления базами данных (приложение Access). Визуализировать табличные модели можно путем построения различных типов диаграмм или графиков.

Группа объектов, обладающих одинаковыми общими свойствами, называется классом объектов. Внутри класса объектов могут быть выделены подклассы, объекты которых обладают некоторыми особенными свойствами, в свою очередь подклассы могут делиться на еще более мелкие группы и т. д. Такой процесс систематизации объектов называется классификацией. В процессе классификации объектов часто строятся информационные модели, которые имеют иерархическую структуру.

В иерархической структуре элементы распределяются по уровням. На первом уровне может располагаться только один элемент, который является «вершиной» иерархической структуры. Основное отношение между уровнями состоит в том, что элемент более высокого уровня может состоять из нескольких элементов нижнего уровня, при этом каждый элемент нижнего уровня может входить в состав только одного элемента верхнего уровня. Иерархические модели могут быть статическими и динамическими. Статическая модель − одномоментный срез информации об объекте. Динамическая модель позволяет увидеть изменения объекта во времени.

Сетевые информационные модели применяются для отражения систем со сложной структурой, в которых связи между элементами имеют произвольный характер.

По степени формализации информационные модели бывают образными, образно-знаковыми и знаковыми.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Образные модели (рисунки, фотографии и др.) представляют собой зрительные образы объектов, зафиксированные на каком-либо носителе информации (бумаге, фото- и кинопленке и др.). Широко используются образные информационные модели в образовании (учебные плакаты по различным предметам) и науках, где требуется классификация объектов по их внешним признакам.

Знаковые информационные модели строятся с использованием различных языков (знаковых систем). Знаковая информационная модель может быть представлена в форме текста (например, программы на языке программирования), формулы, таблицы.

Образно-знаковые модели делятся на геометрические (рисунок, пиктограмма, чертеж, карта, план, объемное изображение); структурные (таблица, граф, схема, диаграмма); словесные (описание естественными языками); алгоритмические (нумерованный список, пошаговое перечисление, блок-схема).

Вопросы для самоконтроля

1. Что такое информационная модель?
2. На какие виды делятся информационные модели по структуре?
3. Что понимается под табличными информационными моделями?
4. Что понимается под иерархическими информационными моделями?
5. С какой целью применяются сетевые информационные модели?
6. На какие виды делятся информационные модели по степени формализации?

15. ИНЖЕНЕРИЯ ЗНАНИЙ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ. ПРЕДМЕТ И ЗАДАЧИ ИНЖЕНЕРИИ ЗНАНИЙ

Инженерия знаний представляет собой совокупность моделей, методов и технических приемов, нацеленных на создание систем, предназначенных для решения проблем с использованием знаний. Инженерия знаний – это теория, методология и технология, которые охватывают методы добычи, анализа, представления и обработки знаний экспертов. Работу по оснащению программ специальными экспертными знаниями из проблемной области, выполняемую человеком либо компьютером (программой), также можно назвать инженерией знаний.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Представление знаний, их обработка и использование, рассматриваемые применительно к конкретной прикладной области, являются предметом инженерии знаний.

С областью инженерии знаний тесно связано понятие искусственного интеллекта. Сущностью искусственного интеллекта можно считать научный анализ и автоматизацию интеллектуальных функций человека.

Задачи инженерии знаний.

1. Анализ предметной и проблемной областей.

Предметная область – сфера человеческой деятельности, выделенная и описанная согласно установленным критериям. В описание должны входить сведения об элементах, явлениях, отношениях и процессах, отражающих различные аспекты этой деятельности. В описании предметной области должны присутствовать характеристики возможных воздействий окружающей среды на элементы и явления предметной области, а также обратные воздействия этих элементов и явлений на среду.

Проблемная область – комплексное понятие, включающее предметную область, решаемые задачи, цели, возможные стратегии и эвристики. Предметную область можно определить как объект или, например, производственную систему со всем комплексом понятий и знаний о ее функционировании. При исследовании проблемной области необходимы знания о задачах, решаемых в производственной системе, и стоящих перед ней целях.

2. Приобретение знаний, которое реализуется путем получения информации извне и ее систематизации.

3. Представление знаний – ключевая задача для систем управления знаниями.

Модели представления знаний можно условно разделить на декларативные и процедурные. Декларативная модель представления знаний основывается на предположении, что проблема представления предметной области решается независимо от того, как эти знания потом будут использоваться. В процедурном представлении знания содержатся в небольших программах, которые определяют, как выполнять специфичные действия (как поступать в специфичных ситуациях).

4. Поиск и хранение знаний.

Поиск и хранение необходимых знаний связаны с понятием корпоративной памяти, которая хранит неоднородную информацию из различных источников и делает ее доступной пользователям для решения корпоративных задач.

В настоящее время актуальна разработка модели представления знаний, которая обеспечивала бы автоматизированную обработку информации на семантическом уровне в системах управления знаниями.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Вопросы для самоконтроля

1. Что представляет собой инженерия знаний?
2. Что является предметом инженерии знаний?
3. В чем заключается сущность искусственного интеллекта?
4. Каковы задачи инженерии знаний?

16. ПОНЯТИЕ «ЗНАНИЯ» В ИСКУССТВЕННОМ ИНТЕЛЛЕКТЕ, ТИПЫ ЗНАНИЙ. ДАННЫЕ И ЗНАНИЯ

Знания – это совокупность сведений о сущностях (объектах, предметах) реального мира, их свойствах и отношениях между ними в определенной предметной области. Иными словами, знания – это выявленные закономерности предметной области (принципы, связи, законы), позволяющие решать задачи в этой области. С точки зрения искусственного интеллекта знания можно определить как формализованную информацию, на которую ссылаются в процессе логического вывода. В этом случае под предметной областью понимается область человеческих знаний, в терминах которой формулируются задачи и в рамках которой они решаются. Предметная область представляется описанием части реального мира, которое в силу своей приближенности рассматривается как ее информационная модель.

Общепризнанного определения знания, как и определения искусственного интеллекта, не существует. Наиболее общее определение трактует знание как всю совокупность данных (информации), необходимую для решения задачи. В этом определении подчеркивается, что данные в привычном понимании также являются знаниями. Однако знания в информационном плане не ограничиваются рамками данных.

Информация, содержащаяся в знаниях, должна включать сведения о системе понятий предметной области, в которой решаются задачи; системе понятий формальных моделей, на основе которых решаются задачи; соответствии систем понятий, упомянутых выше; методах решения задачи; текущем состоянии предметной области. Из перечисленных компонентов только последний соответствует понятию «данные».

Знания можно разделить на процедурные и декларативные.

Декларативные знания – это совокупность сведений о качественных и количественных характеристиках конкретных объектов, явлений и их элементов, представленных в виде фактов и эвристик. Традиционно такие знания

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

накапливались в виде разнообразных таблиц и справочников, а с появлением ЭВМ приобрели форму информационных массивов (файлов) и баз данных.

Процедурные знания – это методы, алгоритмы, программы решения различных задач, последовательности действий (в выбранной проблемной области) – они составляют ядро баз знаний.

Метазнания – это знания об организации всех остальных типов знаний. Иначе они называются специальными. Метазнания содержат признаки декларативных и процедурных знаний.

Работа со знаниями, иначе называемая обработкой знаний, лежит в основе всего современного периода развития искусственного интеллекта.

В памяти ЭВМ знания представляются в виде знаковой системы. С понятием «знак» связываются понятия «экстенсионал» (конкретное значение или класс допустимых значений знака) и «интенсионал» (характеристика содержания знака).

Соответственно различают два типа знаний: экстенсиональные и интенсиональные. Экстенсиональные знания – это набор количественных и качественных характеристик различных конкретных объектов; это данные, хранящиеся в базах данных. Экстенсиональные знания называют также предметными (фактографическими). Интенсиональные знания – это совокупность основных терминов, применяемых в проблемной области, и правил, позволяющих получать новые знания. Например, понятие «персональный компьютер». Его интенсионал: «Персональный компьютер − это дружественная ЭВМ, которую можно поставить на стол и купить менее чем за $2000−3000». Экстенсионал этого понятия: «Персональный компьютер − это Mac, IBM PC, Sinkler...».

Кроме того, знания могут быть поверхностными и глубинными. Поверхностные – знания о видимых взаимосвязях между отдельными событиями и фактами в предметной области. Глубинные – абстракции, аналогии, схемы, отображающие структуру и природу процессов, протекающих в предметной области. Эти знания объясняют явления и могут использоваться для прогнозирования поведения объектов. Например, поверхностные знания: «Если нажать на кнопку звонка, раздастся звук», «Если болит голова, следует принять таблетку»; глубинные знания: «Принципиальная электрическая схема звонка и проводки», «Знания врачей высокой квалификации о причинах, видах головных болей и методах их лечения».

Вопросы для самоконтроля

1. Что такое знания?
2. Как понимаются знания с точки зрения искусственного интеллекта?

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

3. Какие выделяются типы знаний?
4. Что такое данные?
5. Как соотносятся данные и знания?

17. ЭКСПЕРТНЫЕ СИСТЕМЫ: ПОНЯТИЕ, НАЗНАЧЕНИЕ И ОСНОВНЫЕ СВОЙСТВА

Экспертные системы − это наиболее распространенный класс интеллектуальных систем, ориентированный на тиражирование опыта высококвалифицированных специалистов в областях, где качество принятия решений традиционно зависит от уровня экспертизы, например, медицина, юриспруденция, геология, экономика, военное дело и др.

Современные экспертные системы − это сложные программные комплексы, аккумулирующие знания специалистов в конкретных предметных областях и распространяющие этот эмпирический опыт для консультирования менее квалифицированных пользователей. Разработка экспертных систем, как активно развивающаяся ветвь информатики, направлена на использование ЭВМ для обработки информации в тех областях науки и техники, где традиционные математические методы моделирования малопригодны. В этих областях важна смысловая и логическая обработка информации, важен опыт экспертов.

Наибольшие трудности в разработке экспертных систем вызывает не процесс машинной реализации систем, а домашинный этап анализа знаний и проектирования базы знаний. Этим занимается специальная наука − инженерия знаний. Процесс создания экспертной системы часто называют инженерией знаний.

Особенности экспертных систем, отличающие их от других видов программ из области искусственного интеллекта:

1. Экспертные системы имеют дело с предметами реального мира, операции с которыми обычно требуют наличия значительного опыта, накопленного человеком. Экспертные системы имеют ярко выраженную практическую направленность в научной или коммерческой области. Множество программ из области искусственного интеллекта являются сугубо исследовательскими, и основное внимание в них уделяется абстрактным проблемам или упрощенным вариантам реальных проблем.

2. Одной из основных характеристик экспертной системы является ее производительность, т. е. скорость получения результата и его достоверность. Экспертная система должна за приемлемое время найти решение, которое было бы не хуже, чем то, которое может предложить специалист

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

в этой предметной области. Исследовательские программы искусственного интеллекта могут и не быть очень быстрыми – это инструмент исследования, а не программный продукт.

3. Экспертная система должна обладать способностью объяснить, почему предложено именно такое решение, и доказать его обоснованность. Пользователь должен получить всю информацию, необходимую ему для того, чтобы быть уверенным, что решение принято «не с потолка». Экспертная система проектируется в расчете на взаимодействие с разными пользователями, для которых ее работа должна быть по возможности прозрачной. Исследовательские программы «общаются» только со своим создателем, который и так знает, на чем основывается ее результат.

Таким образом, экспертная система содержит знания в определенной предметной области, накопленные в результате практической деятельности человека (или человечества), и использует их для решения проблем, специфичных для этой области. Это очень важное отличие экспертных систем от прочих, «традиционных» систем, в которых предпочтение отдается более общим и менее связанным с предметной областью теоретическим методам, чаще всего математическим.

Вопросы для самоконтроля

1. Что такое экспертные системы?
2. В чем заключается особенность современных экспертных систем?
3. Как называют процесс создания экспертных систем?
4. В чем заключаются особенности экспертных систем, отличающие их от других видов программ из области искусственного интеллекта?

18. АССОЦИАЦИЯ И ЕЕ ВИДЫ. АССОЦИАТИВНЫЕ СЛОВАРИ.

Среди всех типов мышления (наглядно-образное, наглядно-действенное, ассоциативное, теоретическое, практическое и т. д.) наиболее важно для эффективности восприятия мира ассоциативное мышление, хотя в реальности, размышляя, человек не только устанавливает ассоциации, но и производит большое количество других мыслительных операций.

Ассоциация – связь между элементами мыслительного процесса (ощущениями, восприятиями, представлениями, идеями), заключающаяся в том, что появление при определенных условиях одного элемента влечет за собой

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

появление другого или нескольких элементов. В научный и философский язык термин «ассоциация» ввел Джон Локк.

Связи между явлениями и предметами могут затрагивать те или иные их стороны. В связи с этим выделяются ассоциации различных видов: по сходству (предметы внешне похожи друг на друга: *лампочка – груша*, *ложка – лопата*); по контрасту (предметы имеют противоположные признаки: *мокрый – сухой*, *тяжелый – легкий*), эти ассоциации считаются самым сложным видом; по смежности во времени или пространстве (оба предмета расположены близко друг к другу в пространстве или времени: *лето – отпуск*, *стол – стул*); причинно-следственные (предметы являются причиной и следствием друг друга: *лампочка – свет*, *дождь – лужа*); по смыслу (такие взаимосвязи между предметами, которые отражают личный опыт человека и могут быть непонятны для непосвященных: *самолет – любовь*, *дождь – насморк*).

По связи с конкретным органом чувств ассоциации делятся на визуальные, звуковые, осязательные, вкусовые и обонятельные.

Словесные (вербальные) ассоциации могут возникать в ответ на раздражение любого из органов чувств.

Вербальные ассоциации, следуя лингвистической традиции, принято делить на два основных вида: синтагматические (*небо – голубое*) и парадигматические (*стол − стул*). В последнее время выделяют третий вид вербальных ассоциаций − тематические (*друг − дорога*).

Синтагматические ассоциации связывают пары слов, в которых содержание одного члена входит в содержание второго члена в качестве одного из признаков этого содержания (*бабушка − старая*, *бабушка − вяжет*).

Парадигматические ассоциации связывают пары слов, имеющих в своих содержаниях как минимум один общий существенный признак. Они достаточно разнообразны и включают ассоциативные пары, соотносимые с членами различных лексико-семантических, тематических и т. п. полей и групп (*бабушка − дедушка*, *бабушка – старушка* и др.).

Тематические ассоциации связывают пары слов, которые не имеют общих существенных признаков в своем содержании (*бабушка − блины*).

На вербальных ассоциативных связях основан материал ассоциативных словарей.

Ассоциативный словарь – лексикографическое издание, в котором отражена информация о том, как говорящие соединяют слова-реакции с определенными словами-стимулами, что отражает семантические парадигматические и синтагматические связи. В ассоциативном словаре соединяются идеографический и сочетаемостный принципы построения словаря.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Объектом описания ассоциативных словарей является слово. Систематизация слов в ассоциативных словарях основана на психологических ассоциациях предметов или понятий, называемых словом. Лексические единицы группируются в поля, в центре каждого из которых стоит слово, объединяющее другие слова, в той или иной степени близкие ему по значению или ассоциирующиеся с ним по смыслу.

Ассоциативные словари, как и все словари, выполняют несколько главных функций: информативную (позволяют кратчайшим путем приобщиться к знаниям), коммуникативную (дают знания о словарном составе родного или чужого языка как о средстве общения) и нормативную (фиксируют значения слов, закрепляют языковую норму).

Применение специальных компьютерных программ, предназначенных для исследования материалов ассоциативного словаря, позволяет выявить наиболее вероятные прямые и обратные связи между словами, а также установить силу такой связи и судить о близости значений слов.

Вопросы для самоконтроля

1. Что такое ассоциация?
2. Каковы основные виды ассоциаций?
3. На какие виды делятся вербальные ассоциации?
4. Что такое ассоциативный словарь?
5. Какие функции выполняют ассоциативные словари?
6. Какую роль в разработке ассоциативных словарей играют компьютерные программы?

19. СЕМАНТИЧЕСКАЯ СЕТЬ КАК МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ. СТРУКТУРА И КЛАССИФИКАЦИЯ СЕМАНТИЧЕСКИХ СЕТЕЙ

Семантическая сеть – модель предметной области, которая отражает семантику предметной области в виде понятий и отношений.

Изначально семантическая сеть была задумана как модель представления долговременной памяти в психологии, но впоследствии стала одним из способов представления знаний в экспертной системе.

В названии соединены термины двух наук – математики и лингвистики: семантика в языкознании изучает смысл единиц языка, а сеть в мате-

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

матике представляет собой разновидность графа – набора вершин, соединенных дугами, которым присвоено некоторое число.

Основной формой представления семантической сети является граф, вершины которого соответствуют объектам предметной области, а дуги задают отношения между ними. Объектами могут быть понятия, события, свойства, процессы.

В качестве отношений наиболее часто используются следующие: таксономические («класс – подкласс – экземпляр», «множество – подмножество – элемент» и т. п.); структурные («часть – целое»); функциональные (определяемые обычно глаголами *производит*, *влияет* и т. п.); количественные; пространственные; временные; логические; казуальные (причинно-следственные) и др.

Семантические сети классифицируются:

− по количеству типов отношений: однородные с единственным типом отношений и неоднородные с различными типами отношений;

− по назначению (обычно совпадает с преобладающим типом отношений): классифицирующие − позволяют описывать различные иерархические отношения между понятиями; функциональные − вычислительные модели, позволяющие описывать процедуры вычислений одних информационных единиц через другие; сценарии − используются для описания казуальных отношений (причинно-следственных или устанавливающих влияние одних явлений или фактов на другие), а также отношений типа «средство − результат», «орудие − действие» и др.

Достоинства семантических сетей: универсальность, достигаемая за счет выбора соответствующего набора отношений; наглядность системы знаний, представленной графически; близость структуры сети, представляющей систему знаний, семантической структуре фраз на естественном языке; соответствие современным представлениям об организации долговременной памяти человека.

Недостатки семантических сетей: сетевая модель не содержит ясного представления о структуре предметной области; сетевые модели представляют собой пассивные структуры, для обработки которых необходим специальный аппарат формального вывода; проблема поиска решения в семантической сети сводится к задаче поиска фрагмента сети, соответствующего подсети, отражающей поставленный запрос (это обусловливает сложность поиска решения в семантических сетях).

Для реализации семантических сетей в экспертных системах существуют специальные сетевые языки. Систематизация отношений конкретной семантической сети зависит от специфики знаний предметной области и является сложной задачей.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Особого внимания заслуживают общезначимые отношения, присутствующие во многих предметных областях. Именно на таких отношениях основана концепция семантической сети.

Семантические сети широко используются в экспертных системах в качестве языка представления знаний, в системах распознавания речи и понимания естественного языка.

Вопросы для самоконтроля

1. Что такое семантическая сеть?
2. Что является основной формой представления семантической сети?
3. Как семантические сети классифицируются по количеству типов отношений?
4. Как семантические сети классифицируются по назначению?
5. Каковы достоинства и недостатки семантических сетей?
6. Каковы основные сферы применения семантических сетей?

20. ГИПЕРТЕКСТ: МОДЕЛЬ И СТРУКТУРА. КЛАССИФИКАЦИЯ ГИПЕРТЕКСТОВЫХ СИСТЕМ

Гипертекст − совокупность электронных документов, связанных между собой специальными ссылками (гиперссылками) для быстрого перехода от одного документа в заданное место другого и произвольных перемещений внутри документов; технология построения совокупностей связанных гиперссылками документов, применяемая при разработке вебсайтов, электронных энциклопедий, словарей и др.

Гиперссылка – это выделенная (цветом, подчеркиванием или другими средствами экранной визуализации) часть электронного документа, представленная фрагментом текста и скрытым от пользователя адресом, указывающим (программе, интерпретирующей документ) место перехода (в другом или том же документе). Чтобы сделать переход, пользователю достаточно активировать гиперссылку.

Гипертекстовый документ – электронный документ с гиперссылками. Гипертекстовые документы создают с помощью редакторов, имеющих встроенные интерпретаторы языков разметки (HTML (HyperText Markup Language – язык разметки гипертекста), XML (eXtensible Markup Language – расширяемый язык разметки) и др.). Язык разметки позволяет описать структуру документа, размещение и формат вывода заголовков, фрагментов текста, изображений и других составляющих, задать гиперссылки и др.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Разметка делается с помощью специальных меток, названных тэгами. Интерпретируя тэги, браузер или другая программа формирует отображение документа, соответствующее устройству вывода, например, дисплею ноутбука, смартфона или др.

По мере развития интернета в гипертекстовые документы стали включать не только текстовые и графические, но и аудио- и видеосоставляющие. Такие документы и технологии их разработки получили название гипермедийных.

Гипертекстовая технология реализуется в конкретной гипертекстовой системе, которая состоит из двух частей: гипертекста (базы данных) и гипертекстовой оболочки (системы управления гипертекстом). Гипертекстовая оболочка выполняет следующие функции: поддержка ссылочных связей; создание, редактирование и наращивание гипертекста; прямой доступ; просмотр, или браузинг; выделение виртуальных структур.

Принцип автоматизации процессов создания совокупностей связанных между собой документов и их использования для сохранения и накопления знаний сформулировал американский инженер Ванневар Буш (1945). Он предложил концепцию машины коллективной памяти, предназначенной сделать более доступными накопленные знания. Термин «гипертекст» ввел в 1965 году американский первооткрыватель в области информационных технологий Теодор Нельсон. Создание британским физиком Тимом Бернерсом-Ли языка HTML (1990-е гг.) и сделанное им же изобретение Веба (1989) стали важнейшими событиями в развитии гипертекстовой технологии.

Вопросы для самоконтроля

1. Что такое гипертекст и гипертекстовый документ?
2. В чем заключается сущность гипертекстовой технологии?
3. Каковы важнейшие события в развитии гипертекстовой технологии?

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. Скорины

ЛИТЕРАТУРА

1. Баранов, А. Н. Введение в прикладную лингвистику : учеб. пособие / А. Н. Баранов. – М. : Эдиториал УРСС, 2001. – 360 с.

2. Беляева, Л. Н. Лингвистические автоматы в современных гуманитарных технологиях : учеб. пособие / Л. Н. Беляева. – СПб. : ООО «Книжный дом», 2007. – 192 с.

3. Лингвистические ресурсы автоматизированного рабочего места филолога : коллективная монография / Л. Н. Беляева [и др.]. – СПб. : Инфо-Да, 2004. – 184 с.

4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Е. И. Большакова [и др.]. − М. : МИЭМ, 2011. − 272 с.

5. Боярский, К. К. Введение в компьютерную лингвистику : учеб. пособие / К. К. Боярский. – СПб. : НИУ ИТМО, 2013. – 72 с.

6. Введение в электронные лингвистические ресурсы : учеб. пособие для студ. вузов / сост. В. Е. Гольдин, О. Ю. Крючкова. – Саратов, 2011. – 63 с.

7. Всеволодова, А. В. Компьютерная обработка лингвистических данных / А. В. Всеволодова. – М. : Флинта : Наука, 2007. – 96 с.

8. Гаврилова, Т. А. Базы знаний интеллектуальных систем : учебник / Т. А. Гаврилова, В. Ф. Хорошевский. – СПб. : Питер, 2000. – 384 с.

9. Джексон, П. Введение в экспертные системы : учеб. пособие / П. Джексон ; пер. с англ. – М. : Издательский дом «Вильямс», 2001. − 624 с.

10. Захаров, В. П. Информационно-поисковые системы : учебно-метод. пособие / В. П. Захаров. − СПб. : СПбГУ, 2005. − 48 с.

11. Захаров, В. П. Корпусная лингвистика : учебник для студентов гуманитарных вузов / В. П. Захаров. – Иркутск : ИГЛУ, 2011. – 161 с.

12. Зубов, А. В. Информационные технологии в лингвистике : учеб. пособие / А. В. Зубов, И. И. Зубова. – М. : Академия, 2004. – 208 с.

13. Зубов, А. В. Основы искусственного интеллекта для лингвистов / А. В. Зубов, И. И. Зубова. – М. : «Логос», 2007. – 320 с.

14. Леонтьева, Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы : учеб. пособие для студ. лингв. фак-тов вузов / Н. Н. Леонтьева. − М. : Академия, 2006. – 304 с.

15. Марчук, Ю. Н. Компьютерная лингвистика : учеб. пособие / Ю. Н. Марчук. – М. : АСТ : Восток – Запад, 2007. – 317 с.

16. Национальный корпус русского языка: 2003–2005: результаты и перспективы : сб. ст. / Ин-т рус. яз. им. В. В. Виноградова Рос. Акад. наук. – М. : Индрик, 2005. – 343 с.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

17. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / отв. ред. В. А. Плунгян. – СПб. : Нестор-История, 2009.– 502 с.

18. Потапова, Р. К. Новые информационные технологии и лингвистика : учеб. пособие / Р. К. Потапова. – 4-е изд. – М. : КомКнига, 2005. – 368 с.

19. Романенко, В. Н. Сетевой информационный поиск: Информация в Интернете. Поисковые машины. Электронные каталоги библиотек. Как формулировать запросы : практич. пособие / В. Н. Романенко, Г. В. Никитина. – СПб., 2003. – 285 с.

20. Хроленко, А. Т. Современные информационные технологии для гуманитария : практич. руководство / А. Т. Хроленко, А. В. Денисов. – М. : Флинта : Наука, 2007. – 128 с.

21. Автоматическая обработка текстов [Электронный ресурс]. − Режим доступа : http://aot.ru.

22. Беларускі N-корпус [Электронный ресурс]. – Режим доступа : http://bnkorpus.info/.

23. Вавилонская башня – собрание ЛБД и других электронных языковых ресурсов, инициированное С. А. Старостиным [Электронный ресурс]. – Режим доступа : http://starling.rinet.ru.

24. ДИАЛОГ: Международная конференция по компьютерной лингвистике [Электронный ресурс] – Режим доступа : http://www.dialog-21.ru.

25. Интеллектуальные информационные системы [Электронный ресурс]. – Режим доступа : http://www.pd-web.net/intellektualnye-informacionnyesistemy/.

26. Информационно-поисковые системы по словарям и базам данных русского языка [Электронный ресурс]. – Режим доступа : http://lexrus.ru.

27. Информационные технологии в обучении языку. Ресурсный центр учебно-научной лаборатории прикладной лингвистики и информационных образовательных технологий Института дистанционного образования Новосибирского государственного технического университета [Электронный ресурс]. – Режим доступа : http://www.itlt.edu.nstu.ru.

28. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа : http://ruscorpora.ru/.

29. Портал искусственного интеллекта [Электронный ресурс]. − Режим доступа : http://www.aiportal.rU/articles/knowledge-models/l/.

30. Проект ВААЛ [Электронный ресурс]. – Режим доступа : http://www.vaal.ru.

31. Филология и лингвистика [Электронный ресурс]. – Режим доступа : http://filologia.su/freymy/.

32. Языкознание.ру [Электронный ресурс]. – Режим доступа : http://yazykoznanie.ru.

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Производственно-практическое издание

Гомонова Инна Геннадьевна,
Серикова Ирина Валерьевна

КОМПЬЮТЕРНАЯ ФИЛОЛОГИЯ

Практическое руководство

Редактор В. И. Шкредова
Корректор В. В. Калугина

Подписано в печать 27.05.2020. Формат 60x84 1/16.
Бумага офсетная. Ризография.
Усл. печ. л. 2,56. Уч.-изд. л. 2,8.
Тираж 25 экз. Заказ 238.

Издатель и полиграфическое исполнение:
учреждение образования
«Гомельский государственный университет имени Франциска Скорины».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 3/1452 о т 17.04.2017 .
Специальное разрешение (лицензия) № 02330 / 450 от 18.12.2013.
Ул. Советская, 104, 246019, Гомель

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

И. Г. ГОМОНОВА, И. В. СЕРИКОВА

КОМПЬЮТЕРНАЯ ФИЛОЛОГИЯ

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Гомель
2020