Модуль I. Лекция 3
Тема: Компьютерная лингвистика

1. Работа с текстом
2. Автоматическое аннотирование и реферирование текста
3. Аппаратное и программное обеспечение информационных технологий в лингвистике.

1. Работа с текстом

В условиях все возрастающего количества текстов в окружающем человека мире возникает проблема: как в этом море информации найти нужные документы и познакомиться с их содержанием? Решению данной проблемы может помочь составление рефератов и аннотаций полнотекстовых документов. Они дают читателю представление о содержании исходных документов и позволяют оценить степень необходимости обращения к полным текстам каждой работы. Кроме того, рефераты и аннотации акцентируют внимание читателя на новых сведениях, т.е. позволяют за небольшой промежуток времени узнать много новой информации.

Рефераты и аннотации составляются вручную, например самим автором исходного текста или библиографическим работником, или автоматически, с помощью специальных компьютерных программ. Наиболее качественным является первый вид рефератов и аннотаций, поскольку в этом случае создается новый текст, называющий основную мысль высказывания и отличающийся связным характером. Но для обработки большого массива текстов за минимальное количество времени требуется привлечение автоматических средств для решения задачи реферирования и аннотирования текстов.

Реферат определяется как связный текст, который кратко выражает

• центральную тему,

• предмет

• цель,

• методы,

• результаты исследования.

Рефераты обычно составляют к научно-техническим документам научным монографиям, статьям, патентам на изобретение и др. В зависимости от жанра исходного текста (монография, статья, патенты др.) и от предметной области (медицина, химия, лингвистика и т.д.) заданные элементы реферата могут различаться. Так, для научных рефератов дополнительно к названным выше элементам реферата прибавляется краткое изложение сути, практической апробации перспектив исследования.

Различают следующие виды рефератов:

• связный текст — новое текстовое образование, порождаемое на основе логико-смыслового анализа исходного текста;

• реферат-клише — модификация заданной клишированной структуры, пустые ячейки которой заполняются после анализа заданного текста;

• квазиреферат — перечень наиболее информативных предложений текста.

Очевидно, что для автоматического создания рефератов — связных текстов требуются более сложные компьютерные программы, чем для создания рефератов-клише и квазирефератов.

Некоторые исследователи считают реферат и аннотацию синонимами, а некоторые предлагают разводить эти понятия, определяя аннотацию как краткое изложение содержания документа, дающее общее представление о его теме. Согласно этому определению в отличие от реферата, знакомящего читателя с сутью излагаемого в документе содержания, аннотация выполняет лишь сигнальную функцию.

2. Автоматическое аннотирование и реферирование текста

В большинстве программ, направленных на автоматическое составление краткого содержания текста, можно задать разную степень компрессии

текста, т.е. одна и та же программа создает как развернутые рефераты, так и краткие аннотации. В связи с этим в отношении автоматического процесса составления краткого содержания текста обычно используется двойное обозначение: автоматическое реферирование и аннотирование текста.

Создаваемые в процессе реферирования и аннотирования аннотации рефераты представляют собой вторичные документы. Первичными (или исходными) документами являются сами книги, статьи, патенты и др.

Программы автоматического аннотирования и реферирования ориентированы на то, как это делает человек. Для человека этот процесс включает следующие этапы:

1) подготовительный: определение темы текста, его понимание;

2) аналитический: деление текста на фрагменты (абзацы и т.п.) и выделение в каждом фрагменте главных смысловых слов, которые составляют план будущего реферата;

3) непосредственное составление реферата или аннотации: соединение выделенных смысловых единиц в связный текст.

Главными смысловыми единицами исходного текста выступают ключевые слова, ключевые словосочетания и ключевые предложения.

Ключевое слово — знаменательное слово, относящееся к основному содержанию текста и повторяющееся в нем несколько раз. Ключевое словосочетание — сочетание слов, среди которых есть одно или несколько ключевых. Ключевое предложение — предложение, которое содержит несколько (два и более) ключевых слов.

По способам выделения из исходных текстов ключевых словосочетаний и предложений различаются следующие методы автоматического реферирования и аннотирования текстов:

1) статистические,

2) позиционные,

3) логико-семантические.

При статистическом методе принадлежность слова к категории ключевых определяется его статистическими характеристиками: ключевое слово согласно этому методу встречается среди знаменательных слов текста наибольшее количество раз. Ключевое предложение, соответственно, содержит несколько ключевых слов, которые располагаются на небольшом расстоянии друг от друга.

В позиционном методе принцип отнесения предложения к ключевым опирается на его местонахождение в тексте: ключевые предложения входят в заголовок, подзаголовок, находятся в начале и конце текста.

Целью логико-семантического метода, при котором учитывается структура и семантика текста, является выделение предложений с наибольшим функциональным весом. Такими предложениями считаются те, которые содержат семантически значимые слова, особым образом связаны с другими предложениями, имеют определенный синтаксический тип предложения и т.п.

Наиболее простыми системами автоматического реферирования и аннотирования является функция AutoSummarize в MS Word, системыIntelligent Text Miner, Oracle Context и Inxight Summarizer (компонент поискового механизма AltaVista) (IBM). Правда, возможности этих программ ограничены выбором оригинальных фрагментов из исходного документа и их соединением в короткий текст.

Кроме того, можно привести примеры следующих систем автоматического реферирования и аннотирования текстов:

• ОРФО 5.0 (компания “Информатик”): программа включает функцию автоматического аннотирования русских текстов;

• “Либретто” (компания “МедиаЛингва”): программа встраивается в Word и обеспечивает автоматическое реферирование и аннотирование русских и английских текстов;

• поисковая система “Следопыт”, которая включает средства автоматического реферирования и аннотирования документов;

• программы Extractor и TextAnalyst (компания “Медиасистемы”),которые выдают последовательности именных групп, выделенных с помощью синтаксических анализаторов.

В целом можно констатировать, что автоматические рефераты и аннотации представляют собой, по сути, квазирефераты, т.е. результатом автоматической компрессии текста в большинстве случаев становится либо набор ключевых слов, либо перечень ключевых предложений, что, впрочем, в значительной степени помогает решить задачу аннотирования и реферирования большого объема текстов в малые сроки.

3. Аппаратное и программное обеспечение информационных технологий в лингвистике

Для выполнения объемных расчетов над лингвистическими данными, а также для лингвистического моделирования удобно использовать электронные вычислительные машины (или компьютеры).

Компьютер — это электронное устройство, служащее для автоматического создания, обработки, передачи и воспроизводства информации по созданным человеком алгоритмам (программам), написанным на понятном для машины языке. Как следует из приведенного определения, в использовании компьютеров сочетается аппаратное (hardware) и программное обеспечение (software) информационных технологий.

К аппаратному обеспечению относится сам компьютер (стационарный или переносной), а также периферийные устройства, служащие для ввода/вывода информации в компьютер пользователем (клавиатура, мышь, монитор, принтер и т.д.) или для соединения компьютера с другими устройствами (например, модем).

Программное обеспечение — это компьютерные программы, представляющие собой последовательность написанных на машинном языке команд, служащие для управления аппаратными средствами или для

выполнения различных операций над информацией, и соответствующая документация.

В зависимости от назначения программных средств различают системное и прикладное программное обеспечение. Системные программы служат управлению работой аппаратных средств и включают операционные системы, утилиты, драйверы и некоторые другие виды программ. Прикладные программы предназначены для конечного пользователя и позволяют ему выполнять различные операции над информацией: создавать и обрабатывать текст (текстовые редакторы), обрабатывать графические изображения (графические редакторы), работать над звуковой и видеоинформацией (мультимедийные программы), создавать электронные таблицы для обработки статистических данных (электронные таблицы) и т.д. Для лингвиста особенно полезными являются такие виды прикладных программ, как электронные переводчики и словари, а также мультимедийные обучающие программы.

Наряду с аппаратным и программным обеспечением (ПО) информационных технологий некоторые исследователи используют также понятие lingware (или linguware), которым обозначаются все лингвистические компьютерные ресурсы (грамматические справочники, словари, энциклопедии, лингвистические базы данных и т.п.).

Совокупность аппаратных, программных и лингвистических средств, необходимых для автоматической обработки лингвистических данных, обозначим понятием автоматическое рабочее место (АРМ) лингвиста. АРМ лингвиста будет включать сам компьютер, операционное и базовое прикладное ПО, а также всевозможные лингвистические компьютерные ресурсы, касающиеся родного и изучаемых иностранных языков.

В зависимости от специализации АРМ лингвиста может дополняться прикладными программами и лингвистическими ресурсами, связанными с переводом или обучением иностранному языку. Задачей обучающихся является постоянная актуализация своего АРМ, включающая поддержание

современного состояния аппаратного и программного обеспечения, а также постоянное пополнение собственной лингвистической ресурсной базы, т.е. поиск, сохранение, приобретение или создание лингвистических справочников, словарей и баз данных.

Вопросы для обсуждения

1. Что такое реферат?
2. Назовите виды реферата. Приведите примеры
3. Что является главными смысловыми единицами исходного текста?
4. Назовите цели логико-семантического метода
5. Какие системы автоматического реферирования и аннотирования текстов вы знаете?

Рекомендуемая литература

1. Башмаков И.А, Башмаков А.И. Интеллектуальные информационные системы. М.: МГТУ им. Н.Э. Баумана, 2005. С. 77—90.

2. Беляева Л.Н. Лингвистические автоматы в современных гуманитарных технологиях: учеб. пособие. СПб.: Книжный Дом, 2007. С. 87—101.

3. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике:учеб. пособие. М.: Академия, 2004. С. 55—75.

4. Батура Т. В., Бакиева А. М. Методы и системы автоматического реферирования текстов — Новосибирск: 2019. — 110 с. — 100 экз. — ISBN 978-5-4437-0974-1

5. Седова Е. П. Автоматическое реферирование научных публикаций средствами синтаксического анализа на материале современных статей по компьютерному си — СПб.: 2018. — 49 с.

6. Осминин П. Г. Современные подходы к автоматическому реферированию и аннотированию // Вестник Южно-Уральского государственного университета — Челябинск: 2012. — вып. 25. — С. 134–135.