Лекция №5

АВТОМАТИЧЕСКОЕ РЕФЕРИРОВАНИЕ И АННОТИРОВАНИЕ. СИСТЕМЫ МАШИННОГО ПЕРЕВОДА

Реферат и аннотация

Рефератом называют:

- **доклад на определенную тему, включающий обзор соответствующих литературных и других источников;**
- **изложение содержания научной работы, книги и т.д.**

Под **аннотацией** понимается **краткая характеристика произведения печати или рукописи.**

Обычно аннотация приводится после библиографического описания источника.

Аннотацию от реферата отличают:

- **существенно меньший объем;**
- **обязательная констатация назначения аннотируемого произведения.**

Автоматические реферирование и аннотирование

Автоматические реферирование и аннотирование получили значительную актуальность в связи с развитием Internet и каталогов информационных ресурсов. Для экономии времени поиска пользователям предлагаются каталоги аннотаций и рефератов источников.

Формирование рефератов и аннотаций вручную требует колоссальных человеческих ресурсов, поэтому и возникла задача создания методов автоматического реферирования и аннотирования.

Автоматическое реферирование и аннотирование — одно из направлений **компьютерной обработки естественно-языковых текстов (Natural Language Processing, NLP-системы)**.

И в этом качестве оно относится к фундаментальным технологиям ИИ.

Основные тенденции

Основные тенденции для данной области:

• аннотированные каталоги перерастают в гипертекстовые;

• на всех крупных сайтах Internet предусматривают оглавления (sitemap) и функции поиска по сайту;

• использование онтологических словарей-тезаурусов общего и специализированного назначения, а также методов ИИ.

Потребности в средствах автоматического реферирования и аннотирования испытывают: корпоративные системы документооборота; поисковые машины и каталоги ресурсов Internet; информационно-библиотечные системы; каналы вещания; службы рассылки новостей и др.

Методы автоматического реферирования и аннотирования подразделяются на **поверхностные** и **глубинные**. **Поверхностные методы** базируются на «экстрагировании» текста. **Глубинные методы**, развиваемые в настоящее время, базируются на применении тезаурусов и развитых механизмов синтаксического разбора текста.

Требования к реферату

Основные требования к реферату:

• сжатие (объем реферата должен составлять от 5 до 30 % от объема исходного документа);

• возможность использования нескольких источников;

• выражение всех основных мыслей оригинала.

Выделяют **три вида рефератов**:

1. повествовательные;
2. информационные;
3. критические (обзоры).

Подходы в теории автоматического реферирования

Построение реферата человеком включает следующие этапы:

- анализ источника;
- выделение в источнике наиболее важных и информативных фрагментов;
- формирование выводов.

В теории автоматического реферирования различают **три основных подхода**.

Первый из них не предполагает опору на знания, связанные с текстом на ЕЯ. В системах такого типа применяется универсальная база правил, независящая от предметной области и языка текста.

Второй подход предусматривает выделение различных уровней понимания текста, что требует использования наряду с универсальными правилами БЗ о предметной области и базы лингвистических правил, зависящих от языка.

Третий подход является гибридным. Он сочетает лучшие стороны первых двух.

Метод составления выдержек

В системах первого типа применяется **метод составления выдержек**. Он реализуется в два этапа.

На первом проводится сопоставление текста и фразовых шаблонов, в результате чего выделяются блоки наибольшей лексической и статистической релевантности.

На втором — путем соединения выделенных фрагментов формируется итоговый документ.

Для реализации первого этапа используют **модель линейных весовых коэффициентов**. В соответствии с ней каждому блоку U текста оригинала автоматически приписываются весовые коэффициенты:

- $\mathbf{к_1}$, зависящий от расположения блока U в оригинале;
- $\mathbf{к_2}$, зависящий от частоты появления блока в оригинале;
- $\mathbf{к_3}$, зависящий от частоты использования блока в ключевых предложениях;
- $\mathbf{к_4}$, отражающий показатели статистической значимости блока.

Метод составления выдержек

Затем по значениям $к_1$, $к_2$, $к_3$ и $к_4$ и коэффициентам настройки программы реферирования α_1, α_2, α_3 и α_4 вычисляется коэффициент важности блока

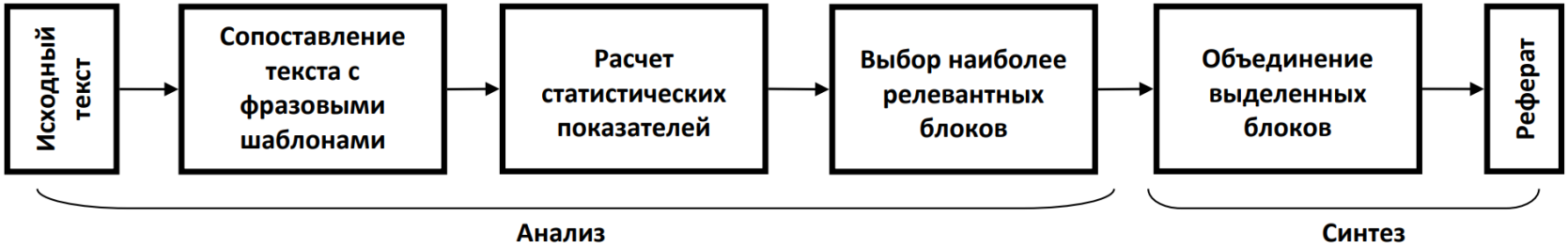
$$\mathbf{B(U) = \alpha_1 к_1 + \alpha_2 к_2 + \alpha_3 к_3 + \alpha_4 к_4.}$$

По коэффициентам важности выполняется отбор блоков в реферат.

Для вычисления каждого весового коэффициента используется своя группа правил. Для $к_1$ они учитывают расположение блока. Для $к_2$ правила учитывают результаты автоматической индексации документа. Для $к_3$ учитывается наличие в блоке таких ключевых фраз и выражений, как «в заключение...», «согласно результатам анализа...», «отличный от...», «малозначащий...» и т.п. Для $к_4$ правила учитывают вхождение термина в заголовки, колонтитулы, первый параграф текста, пользовательский профиль запроса и т.п.

Настройка с помощью коэффициентов α_1, α_2, α_3 и α_4 позволяет управлять степенью сжатия.

Обобщенная архитектура системы первого типа



Главное достоинство описанной модели линейных весовых коэффициентов заключается в простоте ее реализации, а главный недостаток связан с возможностью формирования бессвязных рефератов, не учитывающих контекст. Для его устранения вводится этап ручного редактирования результатов.

Реферирующие системы второго типа

Человеку, уловившему общий смысл информации, легче выделить главное и кратко изложить содержание.

Это и обусловливает создание **реферирующих систем второго типа**.

Для таких систем требуются:

- мощные вычислительные ресурсы;
- развитые грамматики и словари;
- развитые средства синтаксического разбора;
- средства генерации естественно-языковых конструкций;
- онтологические справочники.

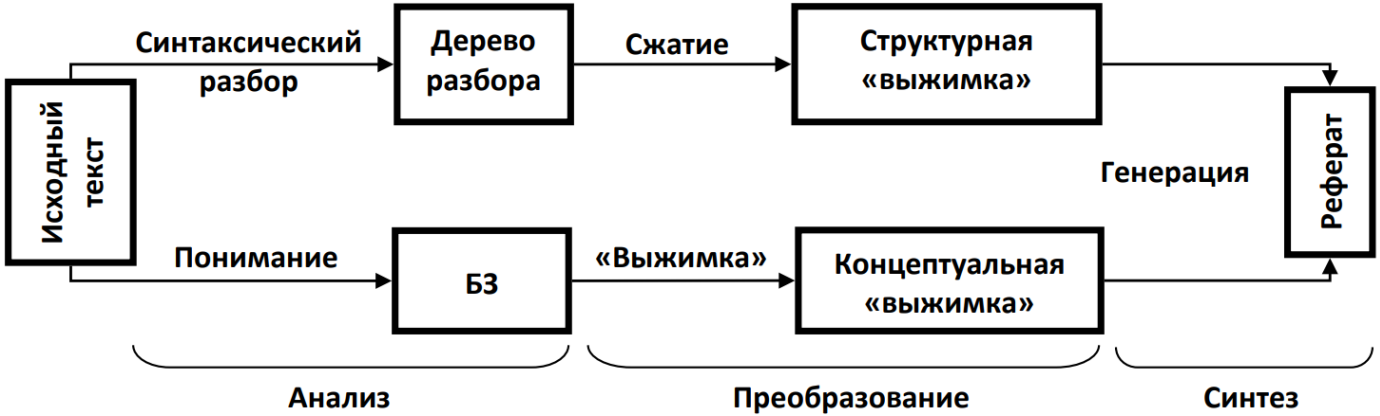
В этих системах реализуются три подхода:

1) традиционный метод синтаксического разбора;

2) подход с опорой на понимание ЕЯ;

3) комбинированный подход.

Формирование реферата в системах с опорой на знания



Стадии синтеза реферата в обоих подходах почти совпадают (используется генератор текста).

Для функционирования подобных систем необходимы:

- исчерпывающие словари (тезаурусы) типа WordNet;
- онтологические справочники типа Cyc и Penman Upper Model;
- большие объемы тестовых файлов с текстами (например, The Wall Street Journal или Perm Treebank от Linguistic Data Consortium).

Задачи, связанные с компьютерным реферированием

Отметим следующие задачи, связанные с компьютерным реферированием:

1. Создание одноязычных рефератов из источников на разных языках.

2. Построение рефератов по гибридным источникам, включающим как текстовые, так и числовые данные в разных формах (таблицы, диаграммы, графики и т.д.).

3. Создание рефератов на основе массивов документов. Например, построение единого реферата по сборнику тезисов докладов научной конференции. Одна из областей применения подобных средств — формирование новостных сообщений по газетным источникам.

4. Растущий объем мультимедийной информации обусловливает актуальность разработки средств ее автоматического реферирования. Методы извлечения семантики из мультимедийной информации находятся на начальных стадиях развития.

Средства автоматического аннотирования в целом аналогичны средствам автоматического реферирования. Однако требования к сжатию текста для них, как правило, на порядок более жесткие.

СИСТЕМЫ МАШИННОГО ПЕРЕВОДА

МАШИННЫЙ ПЕРЕВОД

Машинный перевод (МП) текстов с одних ЕЯ на другие — одна из наиболее ранних задач невычислительных приложений ЭВМ и ИИ.

Отметим два аспекта, определяющих актуальность задач МП и не снижающееся внимание к ним со стороны ученых и разработчиков ИС:

- все возрастающая потребность в переводах в науке, литературе, дипломатии, экономике и других областях деятельности, обусловливаемая повышением открытости границ, интернационализацией науки и экономики, взаимопроникновением культур и т.д.;
- для МП гораздо яснее критерии оценивания результатов, чем в задачах понимания текстов, организации диалога и др.

Создание систем МП требует совместной работы специалистов разного профиля: в первую очередь, лингвистов, математиков и программистов.

Системы МП

Системы МП различают по трем аспектам:

- **рабочим языкам;**
- **типам текста;**
- **ограничениям по ПрО.**

По количеству поддерживаемых рабочих языков различают **двуязычные** и **многоязычные системы МП**.

Язык исходного текста называется **входным**, а **язык перевода** (формируемого текста) — **выходным**.

В современных многоязычных системах МП поддерживаемые языки могут быть и входными, и выходными. Направление перевода определяет роли языков (входной, выходной).

Типы систем МП

По типу текста выделяются **системы для перевода письменного текста** и **устного диалога**.

Системы первого типа классифицируются по назначению для перевода:

- **деловой прозы** (научно-технических статей, заголовков и аннотаций, описаний изобретений, технической документации и др.);
- **художественной литературы**.

Системы для перевода устного диалога обычно **ориентированы на узкую тематику**:

- **резервирование мест в гостинице;**
- **определение маршрута проезда по городу и т.д.**

Такие системы интегрируются с системами анализа и синтеза устной речи.

Ограничения систем МП по предметной области обусловлены поддержкой в них лексики, соответствующей той или иной области знаний (медицины, информатики, математики и т.д.).

Системы МП

Системы МП бывают **автоматическими** и **автоматизированными**.

Автоматизированные системы МП реализуют **три схемы работы**:

- с постредактированием;
- с предредактированием;
- с пред- и постредактированием.

Выполняя перевод, человек уясняет смысл очередного фрагмента текста (фразы, абзаца) и выражает его на выходном языке, стараясь обеспечить структурную и смысловую близость к оригиналу.

При переводе человек использует как **лингвистические знания** о входном и выходном языках, так и **экстралингвистические знания** (знания о ПрО, общих закономерностях среды перевода, законах коммуникации).

Поколения систем МП

В соответствии с возможностями компьютерной реализации данных функций человека и разрабатывались поколения систем МП.

Выделяют **три поколения таких систем:**

1) П-системы – системы прямого перевода (direct systems);

2) Т-системы (от слова transfer – преобразование);

3) И-системы (от слова interlingua – язык-посредник).

Цикл работы П-системы состоит из трех этапов.

- **На первом этапе выполняется морфологический анализ входной фразы.**
- **На втором этапе выполняется перевод морфологического представления входной фразы в морфологическое представление выходной фразы.**
- **На третьем этапе выполняется морфологический синтез.**

Итоговый результат по качеству получается немного лучше подстрочного перевода.

Поколения систем МП

В **Т-системах** помимо **процедур морфологической обработки** реализуются **методы синтаксического анализа** и **синтеза**.

Работа Т-системы включает пять этапов:

- **На первом этапе** осуществляется **морфологический анализ входной фразы** (аналогично П-системам).
- **На втором этапе** по его результатам выполняется **синтаксический анализ**.
- **На третьем этапе** выполняется **переход от входного к выходному языку**.

Выделяются три уровня преобразования:

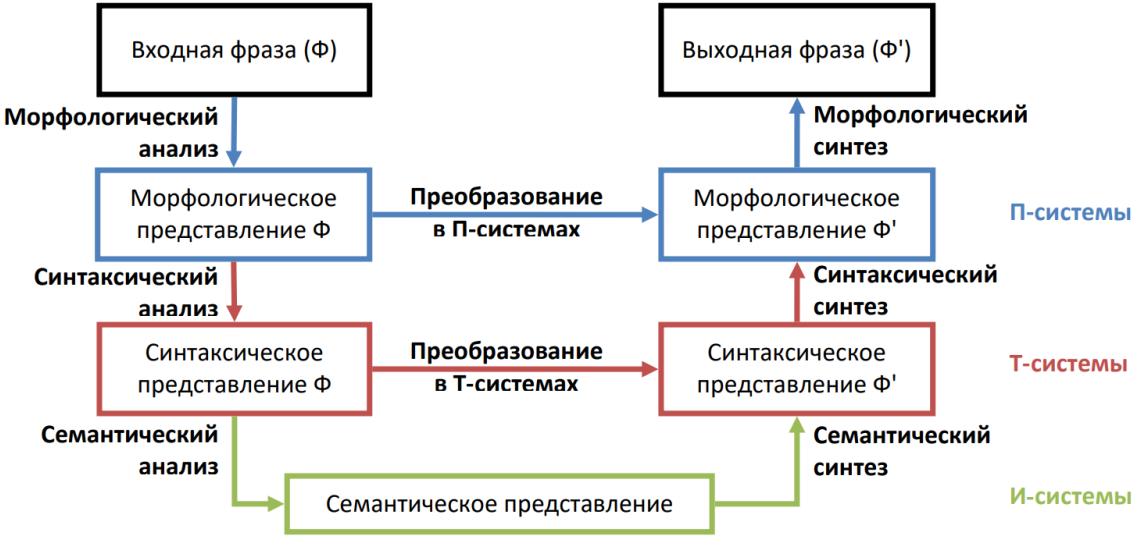
- **поверхностно-синтаксический;**
- **глубинно-синтаксический;**
- **синтактико-семантический.**

- **На четвертом этапе** проводится **синтаксический синтез**.
- **На пятом этапе**, как и в П-системах, осуществляется **морфологический синтез**.

Поколения систем МП

В **И-системах** наряду с **морфологией** и **синтаксисом** используются **экстралингвистические знания**, т.е. знания о **семантике** и **прагматике предметной области**.

Поэтому после **этапов морфологического** и **синтаксического анализа** входной фразы функционирование **И-системы** включает этап **семантического анализа**. Его результатом служат семантические представления входной и выходной фраз, эквивалентные с точностью до лексики.



ОБЕСПЕЧЕНИЕ СИСТЕМ МП

Системы МП представляют собой сложные программные комплексы с разными видами обеспечений.

К **лингвистическому обеспечению** систем МП относятся:

- **словари слов и словосочетаний с соответствующими признаками;**
- **морфологические таблицы суффиксов и окончаний;**
- **базы грамматических правил и др.**

Математическое обеспечение систем МП включает:

- **модели для представления лингвистической информации;**
- **алгоритмы их преобразования;**
- **правила логического вывода для уточнения обрабатываемого текста на основе экстралингвистических знаний.**

К **программному обеспечению** систем МП относятся:

- **программы выполнения перевода;**
- **ведения словарей;**
- **формирования базы правил и т.д.**

Информационное обеспечение систем МП представляет база экстралингвистических знаний о предметной области.