УДК 81’322.2’33

О.В. Лазаренко[1], Д.И. Панченко[2], Е.Ю. Айвас[3]
[1]ХГУ «НУА», г. Харьков, Украина, lazolvlad@gmail.com
[2] ХГУ «НУА», г. Харьков, Украина, panchenko.di2013@gmail.com
[3]ХГУ «НУА», г. Харьков, Украина, b.u.elena@mail.ru

МОДЕЛИРОВАНИЕ БАЗОВЫХ СОСТАВЛЯЮЩИХ ПРОЦЕССА ПОНИМАНИЯ ТЕКСТА В СИСТЕМЕ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ

В статье рассматривается процедура смыслового анализа текста с использованием концептуальных инвариантов текста, текстовых баз, описывающих главные смысловые аспекты текста, и прообразов рефератов для синтеза автоматических рефератов на уровне глубинной семантики. Предложенная процедура позволяет обеспечить универсализацию алгоритма смыслового анализа текстов различной тематики за счет создания ситуационных моделей при разработке системы автоматического реферирования.

АВТОМАТИЧЕСКОЕ РЕФЕРИРОВАНИЕ, СИТУАЦИОННАЯ МОДЕЛЬ, ТЕКСТОВАЯ БАЗА, КОНЦЕПТУАЛЬНЫЙ ИНВАРИАНТ ТЕКСТА, ПРООБРАЗ РЕФЕРАТА

Введение

Изучение и моделирование процесса понимания человеком текста, проводимые в различных областях современных научных исследований таких как исследование механизмов работы мозга (Дж. Хокинс [1]), разработка стратегий понимания дискурса (А. ван Дейк [2]), моделирование процесса реферирования (О.В. Лазаренко [3]) и др. прямо и косвенно подтверждают тот факт, что человек в процессе распознавания объектов и ситуаций использует наиболее важные их характеристики, хранящиеся в его памяти в виде инвариантных форм.

В своих исследованиях при разработке системы автоматического реферирования мы вышли на понимание этих механизмов через изучение особенностей смысловой структуры реферата в сравнении с первичным текстом и его заголовком [3,4,5,6 и др.]. В связи с чрезвычайной сложностью процесса реферирования, мы на начальном этапе наших исследований сознательно допустили теоретическую и эмпирическую неполноту, ограничившись изучением индикативных рефератов научных текстов. Оттолкнувшись, таким образом, от понимания особенностей и закономерностей наиболее структурированного объекта (индикативного реферата) и последовательно расширяя и углубляя область исследования, мы пришли к пониманию определенных механизмов сжатия смысла в цепочке «текст-реферат-заголовок». И оказалось, что в основе этих механизмов лежит использование инвариантных форм представления информации.

В ходе разработки процедуры семантического анализа текста с целью сжатия его смысла была построена семантико-контекстную модель реферирования, включающая модель заголовка и текстовую базу. Заголовок рассматривается нами как смысловой инвариант текста, а текстовая база как «информационное ядро» текста, содержащее информацию о ситуации, описанной в тексте. В текстовую базу входят предложения, отражающие основные смысловые аспекты исходного текста.

В своих дальнейших исследованиях процесса понимания текста [3] мы пришли к выводу о том, что процедура смыслового анализа текста с построением текстовых баз при выборе предложений, описывающих главные смысловые аспекты текста, позволяет обеспечить универсализацию алгоритма смыслового анализа текстов различной тематики и различных предметных областей. Инструментом такой универсализации стала ситуационная модель. В разрабатываемой нами системе ситуационная модель формируется в виде накопителя текстовых баз определенной тематики, автоматически извлекаемых из текста в процессе его смыслового анализа в соответствии с разработанным алгоритмом извлечения основных смысловых аспектов текста.

Предложенный подход к смысловому анализу текста позволяет обеспечить более качественный результат автоматического реферирования за счет:

1) выделения макроструктуры текста в виде главных смысловых аспектов и построения из них текстовой базы, являющейся «смысловым ядром» текста;

2) формирования в автоматическом режиме ситуационных моделей в виде накопителей текстовых баз с целью выявления инвариантных репрезентаций ситуаций;

3) использование инвариантных репрезентаций ситуаций для извлечения из текста информации, необходимой при построении реферата любого вида.

Основной целью наших исследований на данном этапе является разработка процедуры построения ситуационных моделей и инвариантной репрезентации ситуации, обеспечивающих анализ глубинной семантики текста.

1. Формирование прообраза реферата с использованием текстовой базы

На очередном этапе исследования процесса понимания текста с целью выделения необходимых смысловых аспектов для построения реферата мы пришли к выводу о том, что использование текстовых баз для выбора предложений, описывающих главные смысловые аспекты текста, позволяет более точно определить предложения-претенденты для формирования прообраза реферата. При разработке методики построения текстовой базы, представляющей «информационное ядро» текста, мы опирались на использование заголовка как смыслового инварианта текста и слов-указателей на необходимые для реферата смысловые аспекты: объект, результат, метод, область исследования, цель исследования [5]. В соответствии с этой методикой для каждого текста создавались две текстовые базы.

В первом случае в текстовую базу были выделены все предложения в тексте, содержащие слова из заголовка.

Во втором в текстовую базу выбирались предложения, содержащие слова из заголовка и слова-указатели на необходимые для реферата смысловые аспекты.

Анализ двух типов текстовых баз показал, с одной стороны, их схожесть в значительной степени, а с другой, более точный выбор предложений при использовании дополнительно слов-указателей.

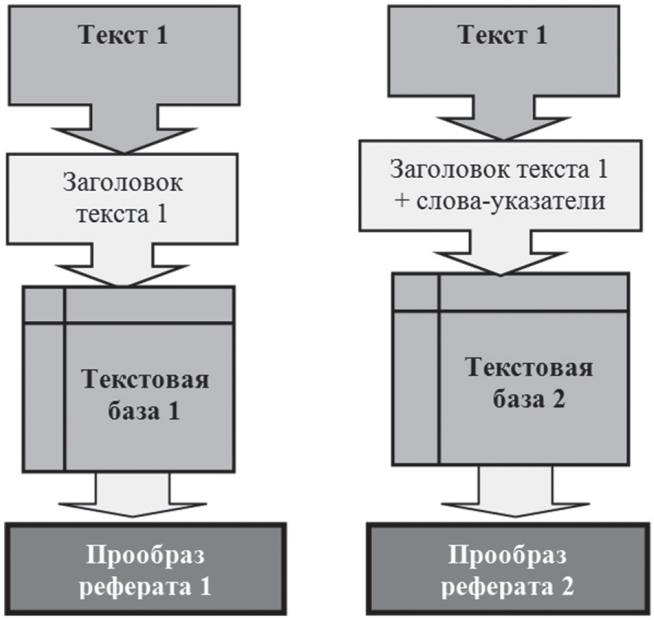

Рис. 1. Процедуры отбора предложений для создания прообраза реферата

По результатам анализа полученных текстовых баз для построения прообраза реферата были выделены первые три и последние два предложения из текстовой базы.

Пример 1.

Заголовок: Интенсификаторы в современном английском языке

Предложения из ТБ:

1. В данной статье в фокусе внимания находятся семантические и структурные свойства интенсификаторов современного английского языка, используемых в современных художественных, публицистических и газетных текстах.

2. Под интенсификаторами мы понимаем разноуровневые единицы языка, функционирующие как усилители признака в широком смысле.

3. Иначе говоря, интенсификатор – это слово или фраза, которая добавляет силу или эмфазу высказыванию.

...

33. В современной литературе, отражающей спонтанную разговорную речь, ощутимо присутствуют экспрессивные интенсификаторы-дисфемизмы, среди которых преобладают так называемые “four-letter-words”.

34. Избыточность представления интенсификации признака является стилистическим маркером экспрессивности, показателем особенностей индивидуального стиля автора или портретной / психологической характеристики героя.

Анализ выделенных предложений:

Все предложения достаточно информативны и описывают общий смысл статьи.

Пример 2.

Заголовок: Роль концептуальной метафоры “death” в экспликации концепта “vampire”

Предложения из ТБ

1. Согласно мнению американских ученых Дж. Лакоффа и М. Джонсона, метафоры являются концептуальными, поскольку одновременно существуют в двух концептах, сферах и тем самым показывают взаимосвязь между ними [Lakoff, Johnson, 1980, p. 37].

2. Понятие концепта является, на сегодняшний день, неоднозначным, однако, мы придерживаемся мнения Ю.С. Степанова, который предложил разделение концепта на слои: актуальный, этимологический и пассивный [Степанов, 2001, с.45].

3. Пассивный слой – особенности, типичные для определенных носителей данного концепта.

...

15. Таким образом, можно сделать вывод, что использование концептуальной метафоры “DEATH” для экспликации концепта “VAMPIRE” является очень распространенным средством, особенно в традиционном готическом романе.

16. Использование данной концептуальной метафоры увеличивает степень влияния на читателя и делает атмосферу романа еще более устрашающей.

Анализ выделенных предложений:

Выделенные предложения передают общий смысл статьи, лишним является третье предложение, т.к. оно содержит описание одного конкретного слоя концепта, что относится к деталям, не отражающимся в индикативных рефератах.

Пример 3.

Заголовок: Фразеологизмы с кулинарным компонентом в контексте гастрономического кода немецкой национальной культуры

Предложения из ТБ:

1. На протяжении столетий язык передаёт верования, заблуждения, национально-культурные установки, фиксируя таким образом национально-специфическое видение окружающего мира.

2. Собранные в этих словарях фразеологические единицы (далее ФЕ) включают в себя устойчивые словесные комплексы эпохи раннего средневековья, они снабжены примерами, авторскими комментариями, пометами об их распространённости в конкретной местности Германии, нередко приводятся эквиваленты из других языков.

3. Сравнительный анализ словарей XIX века и современных лексикографических источников позволяет сделать вывод о том, актуальна ли та или иная идиома в современном немецком языке.

...

36. При этом некоторое число аналогичных устойчивых словесных комплексов в других европейских языках свидетельствует о сходном отношении к пище, сложившемся, в частности, под влиянием христианской культуры.

37. Путём привлечения экстралингвистической информации – сведений о повседневной жизни, социально-культурных установках, религиозных представлениях народа – и соотнесения их с лингвистической формой выражения представляется возможным описать пока ещё недостаточно изученный сегмент гастрономического кода национальной культуры.

Анализ выделенных предложений:

Первые три предложения содержат общую информацию, в которой отсутствует указание на то, чему посвящена статья. Наиболее информативным является последнее предложение.

Общие выводы:

1. Некоторые прообразы реферата получились достаточно информативными (пример 1).

2. Первые три предложения текстовой базы могут содержать общую вводную информацию, которая не является полезной для реферата (пример 2).

3. Некоторые выбранные предложения содержат описание деталей, для понимания которых необходима дополнительная информация (пример 3).

Таким образом, прообразы рефератов не всегда получаются из выбранных первых трех и последних двух предложений из текстовой базы. Это подтвердило нашу гипотезу о том, что при выборе предложений для прообраза реферата, нужно ориентироваться на предложения в ТБ, выбранные из первых четырех абзацев текста, которые практически со 100% вероятностью позволяют найти в них указание на объект, область и цель исследования. Вместе с тем, последние два предложения из ТБ всегда представляют описание результата.

2. Оптимизация процедуры выбора предложений из текстовой базы для прообразов реферата

Очень часто тексты научных статей начинаются с общего описания проблемы, изучения и полученных к данному моменту результатов предшествовавших исследований. Иными словами, с описания истории вопроса. Естественно, что постановка задачи, цель и метод исследования при этом приводятся после вводной части. Вместе с тем во многих статьях такая информация либо полностью отсутствует, либо сведена к нескольким предложениям. Чтобы найти интересующие нас смысловые аспекты при любом варианте написания статьи, в нашем алгоритме поиска предложений в текстовой базе для прообраза реферата мы на начальном этапе отбираем предложения, выделенные в текстовую базу из первых четырех абзацев исходного текста. Экспериментальная проверка подтвердила достаточность этих предложений для выбора объекта (всегда), а также цели, метода и области исследования, если в статье есть на них указание.

Ниже приведен сравнительный анализ двух подходов для выбора предложений из текстовых баз для составления прообраза реферата:

$$R_{pr} \in TextBase_1 (P_1, P_2, P_3, P_{n-1}, P_n)$$

и

$$R_{pr} \in TextBase_2 (P_k, P_{k+1}, \dots, P_{n-1}, P_n),$$

где R_{pr} – прообраз реферата, TextBase – текстовая база, P_1, P_2, P_3 – первые три предложения из тестовой базы первого типа, P_k – первое из предложений, выбранных из первых четырех абзацев текста для текстовой базы второго типа, P_n – последнее предложение в текстовой базе любого типа.

Статья: Роль концептуальной метафоры “death” в экспликации концепта “vampire”

Rpr ∈ TextBase1 (P_1, P_2, P_3, P_{n-1}, P_n)

P_1-{}

P_2-{}

P_3-{}

P_{n-1}-{Таким образом, можно сделать вывод, что использование концептуальной метафоры “DEATH” для экспликации концепта “VAMPIRE”

является очень распространенным средством, особенно в традиционном готическом романе}

P_n - {Использование данной концептуальной метафоры увеличивает степень влияния на читателя и делает атмосферу романа еще более устрашающей}

Rpr ∈ TextBase2 (P_k, P_{k+1}, ... , P_{n-1}, P_n)

P_k - {В нашей статье внимание направлено на актуальный слой концепта “VAMPIRE”, а именно на использование концептуальной метафоры “DEATH” как репрезентанта данного концепта}

P_{k+1} - {Языковое воплощение лингвокультурологического концепта “VAMPIRE” имеет четкий ареал в литературе – это готические романы ужасов, романы о вампирах}

Pn-1 - {Таким образом, можно сделать вывод, что использование концептуальной метафоры “DEATH” для экспликации концепта “VAMPIRE” является очень распространенным средством, особенно в традиционном готическом романе}

Pn - {Использование данной концептуальной метафоры увеличивает степень влияния на читателя и делает атмосферу романа еще более устрашающей}

Приведенное сравнение демонстрирует тот факт, что использование второго типа текстовых баз позволяет создать более полный и точный прообраз реферата.

Как отмечалось в работе [3], в ходе наших исследований мы пришли к выводу о целесообразности поэтапного приближения к выбору необходимых предложений из текста. Поскольку предложений, указывающих на определенный смысловой аспект, может быть несколько, следует выделить их из текста в текстовую базу, которая представляет собой расширенное информационное ядро текста. А затем на основе анализа заголовка, слов-указателей на смысловые аспекты и предложений из текстовой базы выбрать необходимую информацию для прообраза реферата.

Проведенные исследования подтвердили эффективность такого подхода и соответствие полученных результатов концепции понимания, предложенной в работах голландского лингвиста ван Дейка [2], согласно которой для описания глобального содержания текста необходимо построение схемы, обеспечивающей «быстрый анализ поверхностных структур и выстраивание относительно простой и жесткой семантической конфигурации». В нашем случае это текстовые базы и прообразы рефератов. Такие структуры текста представляют собой обобщенное описание основного содержания дискурса, которое читатель строит в процессе понимания, и являются фактически рефератом или резюме. А это то, что и является конечной целью наших исследований.

В результате последних исследований мы вплотную подошли к моделированию процесса реферирования на уровне глубинной семантики текста. Следующим шагом на этом пути будет автоматическое построение ситуационной модели для создания инвариантных репрезентаций ситуаций, описываем в текстах, и представляющих собой набор наиболее важных признаков, выделенных на основе относительных характеристик ситуации, в которых возможны существенные упущения в сравнении с конкретной ситуацией, описываемой в конкретном тексте.

Выводы

В статье рассмотрена процедура построения прообразов рефератов на основе использования текстовых баз, позволяющая обеспечить более качественный результат автоматического реферирования.

Подтвержден тот факт, что для создания связного прообраза реферата необходимо использовать предложения из текстовой базы, выделенные из первых четырех абзацев текста, и последние два предложения текстовой базы.

Построенные таким образом прообразы рефератов являются хорошим базисом для создания инвариантных репрезентаций ситуаций в виде набора наиболее важных признаков обобщающего порядка в сравнении с конкретной ситуацией, описываемой в тексте.

Список литературы:

1. *Хокинс Дж., Блейкcли С.* Об интеллекте / Дж., Хокинс, Блейкcли С. – М.: Издательский дом “Вильямс”, 2007. – 240 с. **2.** *Дейк ван Т. А.* Стратегии понимания связного текста / Т. А. ван Дейк, В. Кинч // Новое в зарубежной лингвистике. – Вып. 23: Когнитивные аспекты языка. – М., 1988. – С. 153–211. **3.** *Лазаренко О.В.* Моделирование процесса понимания текста с использованием инвариантной репрезентации ситуаций в системе автореферирования /О.В. Лазаренко // // Біоніка інтелекту: навч.-техн. журнал. 2014. – Вип. 2(83). С. 15-19. **4.** *Лазаренко О. В.* Моделювання процесу узагальнення в системі автоматичного реферування / О. В. Лазаренко, А. А. Яковенко. – Х.: Изд-во НУА, 2007. – 136 с. **5.** *Лазаренко О.В.* Моделювання семантичних зв’язків «Текст-Реферат» в системах автоматичного реферування / О.В. Лазаренко, Д.І. Панченко. – Х.: Изд-во НУА, 2014. – 176 с. **6.** *Буряк Е.Ю., Лазаренко О. В., Панченко Д.И.* Разработка алгоритма смыслового анализа текста для синтеза реферата в системе автоматического реферирования/ Е.Ю. Буряк, О. В. Лазаренко, Д.И. Панченко / Бионика интеллекта: науч.-техн. журнал – Харьков : ХНУРЭ, 2015. – Вып. 2 (85) – С. 127 – 130.

Поступила в редколлегию 9.11.2016