ИСПОЛЬЗОВАНИЕ ЗАГОЛОВКА ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА СМЫСЛОВОГО АНАЛИЗА ТЕКСТА В СИСТЕМЕ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ

Буряк Е. Ю.
Харьковский гуманитарный университет
«Народная украинская академия»,
г. Харьков, ул. Лермонтовская, 27, тел. 714-20-07,
e-mail: b.u.elena@mail.ru

Целью нашего исследования является анализ влияния заголовка на качество смыслового анализа текста в процессе автоматического реферирования текста. Эта тема возникла как результат научной дискуссии по вопросу целесообразности использования заголовка в системах автоматической обработки текстов. Безусловно, заголовок научного текста является сложным объектом для анализа, так как представляет собой предложение, содержащее в максимально сжатом виде информацию об объекте исследования и полученном в ходе исследования результате. То есть заголовки пишутся лаконично, в них опущены все семантически второстепенные элементы. Однако именно благодаря этому заголовок следует использовать не как источник информации, а как камертон, позволяющий точно выбрать в тексте информацию, отражающую основную идею статьи. По сути, заголовок представляет собой концептуальный инвариант, представленный в различной степени детализации в статье, рефератах этой статьи и других вторичных документах, отражающих ее содержание.

Для того чтобы можно было использовать заголовок в процессе автоматической обработки текста, необходимо детально изучить и описать все варианты смыслового сжатия текста и способы его представления в заголовке, что и было сделано в работе [1]. В заголовке была обнаружена аналогия с компрессией, имеющей место в индикативном реферате. Вследствие чего заголовок можно рассматривать как реферат минимального объема или как текст с максимальным уровнем сжатия смысла.

На основе проведенного исследования была разработана общая модель заголовка:

$$\textbf{СКЗ :}\ О(б)\ /\ К - \mathrm{Sr} - \mathrm{V}(m_5) - \mathbf{A}(\mathbf{m_4}) - \mathrm{A}(m_7) - \mathrm{A}(m_9) - \mathrm{A}(m_8).$$

Жирным шрифтом выделен обязательный элемент заголовка, наличие всех других – возможна.

Модель состоит из обязательных элементов – актантов $A(m_i)$ и необязательных элементов – сирконстантов *Sr*. Актант *объект* является основным в данной структуре, актант *результат* присутствует часто, но не всегда, остальные компоненты присутствуют в зависимости от их значимости в тексте.

В полном виде эта модель встречается редко, так как включает в себя все в принципе возможные варианты. Обычно в ней присутствует от двух до четырех компонентов:

$$\textbf{\textit{СКЗ :}}\ V(m_5) - \mathbf{A(m_4)} - A(m_7) - A(m_9)$$

Процедура дедуктивного вывода для планирования работы системы управления в динамической среде.

В результате исследования содержательной и синтаксической структуры заголовка выявилось его сходство со структурой реферата. Как и в индикативном реферате, содержательная структура заголовка состоит из двух метазначений - *объект* и *результат*. Впрочем, в отличие от реферата они являются элементами содержательной структуры одного предложения и используются в обратной сравнительно с содержательной структурой реферата (*объект - результат*) последовательности: *результат - объект*. Такое сходство смысловых структур реферата и заголовка, содержащих одинаковые смысловые аспекты, стала основанием для изучения взаимосвязи текстов и заголовков, чтобы с помощью информации, содержащейся в заголовке, обнаружить в тексте те лексические единицы, которые необходимы для семантического наполнения модели реферата данного текста.

Задача нашего исследования состояла в анализе заголовка текста как концептуального инварианта, смысл которого разворачивается в тексте статьи, рефератах и интеллект-карте, и доказательств эффективности использования его в процессе автоматического реферирования.

В ходе исследования была рассмотрена статья с заголовком «Что такое Data Mining?», особенностью которого является то, что значение *результат* выражено в заголовке в виде вопроса. Т.е. возникли трудности именно с *результатом*, т.к. необходимо было проследить наличие и повторение в статье и рефератах различных интерпретаций вопроса *Что такое*. Ответом на этот вопрос является описание самого *понятия* Data Mining и его *сути*.

Данный заголовок можно представить в следующем виде:

Что такое = *результат*, Data Mining = *объект*.

Для проведения исследования были написаны информативный и индикативный рефераты в традиционном текстовом виде и в виде интеллект-карты (Mind Map), в которой заголовок был выбран в качестве базового образа.

Было проведено сравнение заголовка с текстом статьи, индикативным и информативным рефератами и ИК. Результатом сравнения стали выводы о том, что данный заголовок конкретно и точно описывает суть статьи, в то время как ИК, индикативный и информативный рефераты подробно и глубоко раскрывают эту суть. Наибольшую схожесть заголовок имеет с индикативным рефератом, т.к. сам, по сути, является рефератом минимального объема и выражен в наиболее сжатой форме.

Фрагмент статьи**:** Что такое Data Mining?

…Термин Data Mining получил свое *название из двух понятий*: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining)... *Термин Data Mining* часто *переводится к*ак добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний…

Фрагмент информативного реферата

В статье рассматривается технология Data Mining, *суть* данной технологии и предпосылки ее появления… Автор описывает *значение термина* Data Mining, его происхождение и приводит примеры различных переводов данного термина…

Индикативный реферат

Описаны *суть*, развитие и *понятие* инновационной технологии Data Mining. Отмечена эффективность использования методов Data Mining при работе с большими базами данных.

Проведенное исследование показывает, что заголовок, являясь концептуальным инвариантом всех текстов (и в статье, и в рефератах, и в ИК можно проследить наличие и повторение описания в различных интерпретациях *объекта* и *результата*, содержащихся в заголовке), позволяет осуществить правильный выбор необходимой информации, раскрывающей и позволяющий понять смысл статьи. Этот факт свидетельствует о целесообразности использования его при разработке алгоритма автоматического реферирования.

Это подтверждается и мнением голландского лингвиста ван Дейка, занимающегося исследованием стратегий понимания дискурса человеком: «Языковому пользователю нет необходимости дожидаться конца абзаца, главы или целого текста, чтобы понять, о чем идет речь в тексте или в его фрагменте, … пользователь языка может догадаться о теме текста уже после минимума текстовой информации из первых пропозиций. Догадку может подтвердить самая различная информация: заглавие, тематические слова, тематические первые предложения…» [1].

Список литературы

1. *Лазаренко О. В.* Анализ смысловой структуры заголовка как текста с максимальным уровнем обобщения / О. В. Лазаренко, Т. В. Попова // Проблеми семантики слова, речення та тексту: Збірник наукових праць. – К.: КНЛУ, 2004. – Вип. 12. – С. 143–149.

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.