УДК 004.82

О. В. Лазаренко – кандидат технических наук, доцент, заведующий кафедрой информационных технологий и математики;
Д. И. Панченко – старший преподаватель, аспирант Харьковского гуманитарного университета "Народная украинская академия"

Роль заголовка в построении системы автоматического реферирования

Робота выполнена на кафедре информационных технологий и математики ХГУ "Народная украинская академия"

В статье рассматривается моделирование процесса реферирования, его формализация и создание на основе разработанных формализмов системы автоматического реферирования. Представляется модель индикативного реферата в виде набора синтаксических конструкций со значениями "объекта" и "результата". Анализируется заголовок текста как реферат минимального объема со значением "результат" и "объект". Рассматривается возможность использования заголовка в качестве отправной точки для автоматического построения реферата.

Ключевые слова: система автоматического реферирования, индикативный реферат, заголовок, реферативная конструкция, онтология.

Лазаренко О. В., Панченко Д. І. Роль заголовка в побудові системи автоматичного реферування. У статті розглядається моделювання процесу реферування, його формалізація й створення на основі розроблених формалізмів системи автоматичного реферування. Подається модель індикативного реферату у вигляді набору синтаксичних конструкцій зі значеннями "об'єкта" й "результату". Аналізується заголовок тексту як реферат мінімального обсягу зі значенням "результат" і "об'єкт". Розглядається можливість використання заголовка як вихідної точки для автоматичної побудови реферату.

Ключові слова: система автоматичного реферування, індикативний реферат, заголовок, реферативна конструкція, онтологія.

Lazarenko O. V., Panchenko D. I. Heading Role in Building System OF Automatic Summarization. The article focuses on the summarization process modeling, its formalization and creation of a system of automatic summarization on the basis of the developed formalisms. The model of an indicative summary in the form of a set of syntactic constructions with the meaning "object" and "result" has been presented. The text heading has been analyzed as the minimum volume summary with the meaning "result" and "object". The possibility to use the heading as a starting point for automatic summarization has been under consideration.

Key words: system of automatic summarization, indicative summary, heading, summary construction, ontology.

Постановка научной проблемы и её значение. Широкий доступ к информации, накопленной человечеством, дает сегодня возможность быстро ее получать, обрабатывать и передавать дальше. Такие возможности диктуют довольно жесткие требования к человеку, а он, в свою очередь, обращается за помощью к системам, которые помогли бы оперативно и качественно выполнять уже не только рутинную (форматирование, корректировку и т. п.), но и содержательную обработку информации. В этой ситуации системы, способные анализировать информацию и обрабатывать ее на семантическом уровне, становятся острейшей потребностью современного общества.

К таким интеллектуальным системам относятся и системы автоматического реферирования, призванные существенным образом снизить напряжение при работе с большими информационными массивами. Поэтому потребность в автоматизации процесса создания качественных рефератов, сжато излагающих содержание документов, велика сегодня как никогда раньше.

В связи с этим особую актуальность приобретает задача моделирования процессов понимания, обобщения текстов и разработки интеллектуальных систем реферирования с опорой на знания. Все попытки автоматизации реферирования, которые предпринимались до последнего времени, привели к однозначному выводу – решение этой задачи невозможно без разработки соответствующих средств представления знаний. Статистические методы, использовавшиеся для анализа содержания текста, желаемых

результатов не дали. Поэтому сейчас во всем мире широким фронтом ведутся работы по созданию концептуальных моделей представления знаний и их использования для решения прикладных задач по обработке информации.

Целью наших исследований является моделирование процесса реферирования, его формализация и создание на основе разработанных формализмов системы автоматического реферирования (АР).

Изложение основного материала и обоснование полученных результатов иследования. Многие исследователи, работающие над созданием систем автоматического реферирования, стараются описать и формализовать процесс понимания. В идеале модель понимания для систем АР должна учитывать сложную иерархию качественно разных видов понимания – от языкового до психологического, и более того включать понимание текста каждым человеком соответственно его внутреннему миру, индивидуальному смысловому контексту. Такой подход, охватывающий все стороны изучаемого явления, пока далек от практической реализации.

Приступив к моделированию процесса реферирования, представляющего собой совокупность сложнейших процессов понимания и обобщения смысла, мы начали с изучения не самих процессов понимания, а с изучения их результата – с реферата. Причем не развернутого, информативного, а сжатого, индикативного. И не только потому, что он наиболее востребован сегодня в системах поиска в Интернете, но и, в первую очередь, потому, что рассматриваем его как отправную точку в исследовании этого вопроса, как объект наиболее простой по форме, но отражающий все особенности реферативного текста.

На сегодня проведенное нами исследование смысловой и синтаксической структур реферата позволило сделать выводы о том, что такое обобщение в реферировании, к каким особенностям в структуре реферативных предложений оно приводит, и на основании выявленных особенностей синтактико-семантической структуры этих предложений построить модель индикативного реферата [1].

Не углубляясь в детали этих исследований, остановимся на некоторых выводах, имеющих принципиальное значение для понимания специфики проведенного нами исследования заголовка [2].

Анализ синтаксической структуры реферативных предложений показал, что набор характерных для них синтаксических конструкций очень ограничен – это преимущественно простые предложения с неопределенно-личными бесподлежащными и страдательно-возвратными конструкциями. Он также целиком подтвердил распространенную в современной лингвистике вербоцентрическую концепцию, соответственно которой основной составляющей простого предложения является элементарная глагольная конструкция, состоящая из основного глагола (предикатного ядра) и зависимых от него обязательных элементов (актантов). В таких конструкциях глагол является основной единицей, которая определяет его смысловые связи с понятиями, актантами, заполняющимися существительными (именными группами).

Более того, общей оказалась семантика синтаксических структур этих предложений – “отношение между субъектом и его предикативным признаком – состоянием как результатом действия”.

Все это в полной мере отвечает основной функции реферата – краткому изложению того, что исследовано, открыто, описано. Т. е. сжатие в процессе реферирования имеет свою ярко выраженную специфику: отход от развернутых описаний, соображений, – и осуществляется на синтаксическом уровне за счет использования в тексте простых предложений с фиксированной семантикой синтаксических конструкций (“результат действия” или “направленность действия на достижение результата”).

Анализ большого корпуса индикативных рефератов показал, что сжатие на смысловом уровне происходит не только в рамках отдельного предложения, но и всей содержательной структуры реферата. Оказалось, что индикативный реферат, как правило, состоит из двух предложений со значениями – “объект исследования” и “результат исследования”. При этом первым предложением в реферате является предложение со значением “объекта”, а второе – со значением “результата”. Это разрешило представить модель реферата в виде набора синтаксических конструкций $\boldsymbol{СК^{o}_{1}}$ и $\boldsymbol{СК^{p}_{2}}$ со значениями *объект* и *результат*.

В отличие от реферата исходный текст содержит большой массив общеупотребительной и общенаучной лексики. Поэтому основная задача реферирования заключается в переходе от развернутой смысловой структуры текста к обобщенной смысловой структуре реферата.

Справиться с этой задачей может лишь система, которая способна анализировать содержание текстовых документов не только по формальным, но и, в первую очередь, смысловым признакам. Наш подход к решению этой задачи является одной из попыток создания такого рода системы.

Анализ содержания текста мы начали с анализа его заголовка, который рассматриваем как реферат минимального объема или как текст с максимальным уровнем обобщения смысла. Поскольку главной составляющей процедуры реферирования является обобщение, или точнее сжатие, нас также интересует, в чем заключается отличие обобщения при переходе от текста к заголовку, от текста к реферату и от реферата к заголовку, и в чем оно выражается. Обобщение при переходе от текста к реферату (обобщение 2) описано в работе [1]. Изучая смысловую структуру заголовка в сравнении с реферативной структурой, мы попробовали разобраться, как выражается обобщение при переходе от реферата к заголовку (обобщение 3).

В результате исследования смысловой и синтаксической структуры заголовка было обнаружено его сходство со структурой реферата. Также как и в индикативном реферате, смысловая структура заголовка состоит из двух метазначений – *объект* и *результат*. Но в отличие от реферата они являются элементами смысловой структуры одного предложения и употребляются в обратной последовательности: сначала – *результат*, потом *объект*. Такое сходство смысловых структур реферата и заголовка послужило основанием для изучения взаимосвязи текстов и их заголовков для того, чтобы с помощью информации, которая содержится в заголовке, выявить в тексте те лексические единицы, которые необходимы для семантического наполнения модели реферата данного текста.

В исследовании [1] была построена классификация лексем, принимающих участие в заполнении реферативных конструкций. В последующем была построена классификация лексем заголовка [2] и проведено ее сравнение с первой классификацией.

Как и в работе [1], все множество глаголов-предикатов было разбито на три класса со значением: $\mathbf{m_1}$ – *форма представления информации в тексте*, $\mathbf{m_2}$ – *этап работы*, $\mathbf{m_3}$ – *сравнительная оценка*. Лексические значения существительных, которые входят в состав именных групп, были разделены на шесть классов со значением: $\mathbf{m_4}$ – *объект*, $\mathbf{m_5}$ – *результат* (или *процесс, который стремится к достижению результата*), $\mathbf{m_6}$ – *свойство*, $\mathbf{m_7}$ – *цель*, $\mathbf{m_8}$ – *инструмент* (метод) и $\mathbf{m_9}$ – *место*.

Семантический анализ заголовков позволил сделать следующие выводы:

1. Основными компонентами структуры заголовка являются слова, которые принадлежат к классам со значениями *объект* и *результат*. При этом, в отличие от реферата в заголовочной конструкции, на первом месте, как правило, стоит актант *результат* (или *процесс, стремящийся к достижению результата*).

2. Отличительная особенность заголовков заключается в том, что практически во всех заголовках отсутствует предикат в чистом виде. В заголовке он, как правило, трансформируется в отглагольное существительное или существительное со значением действия. Это согласуется с известным в стилистике фактом, что замена отглагольных существительных глаголами способствует конкретизации текста, из чего следует, что трансформация глагола в отглагольное существительное повышает абстрактность, а, следовательно, и обобщенность текста.

3. Если рассматривать классы со значениями $\mathbf{m_1}$, $\mathbf{m_2}$ и $\mathbf{m_3}$ как отглагольные существительные, то можно сказать, что $\mathbf{m_1}$ как форма представления информации в тексте {*изложение, освещение, описание, введение, рассмотрение*} обычно не присутствует в заголовке, $\mathbf{m_3}$ демонстрирует субъективную оценку автора {*добавление, изменение, исправление, улучшение*}, что также встречается крайне редко, а $\mathbf{m_2}$ – трансформируется в $\mathbf{m_5}$, что выступает в роли *результата* и является одним из двух основных компонентов заголовочной структуры. Например, в реферате: *Анализируется инвести-*

ционная привлекательность различных регионов Украины. В заголовке: *Анализ инвестиционной привлекательности различных регионов Украины*.

4. Проведя сравнительный анализ с текстом реферата, было установлено, что все компоненты, присутствующие в заголовке, соответствуют компонентам реферата.

На основе проведенного исследования была построена общая модель заголовка:

СҐЗ = [ОБ(б) / К] [Sr] [V(m_5)] **A(m_4)** [A(m_7)] [A(m_9)] [A(m_8)].

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

Модель состоит из обязательных элементов – актантов **A(m_i)** и необязательных элементов – сирконстантов **Sr**. Наречия, заполняющие сирконстанты, являются носителями оценочной семантики, поэтому заполнение сирконстантов – исключительно интеллектуальный процесс. Актант "объект" в данной структуре является основным ее компонентом, актант "результат" присутствует, но не всегда, другие компоненты присутствуют в зависимости от их значимости в тексте, относящемуся к заголовку.

Следует отметить, что в полном виде данная модель встречается очень редко. Обычно в ней присутствуют от двух до четырех компонентов:

СҐЗ = V(m_5) A(m_4) A(m_7) A(m_9)

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

СҐЗ = V(m_5) A(m_4) [A(m_7)]

Формирование семантических признаков для математической модели префиксального словообразования.

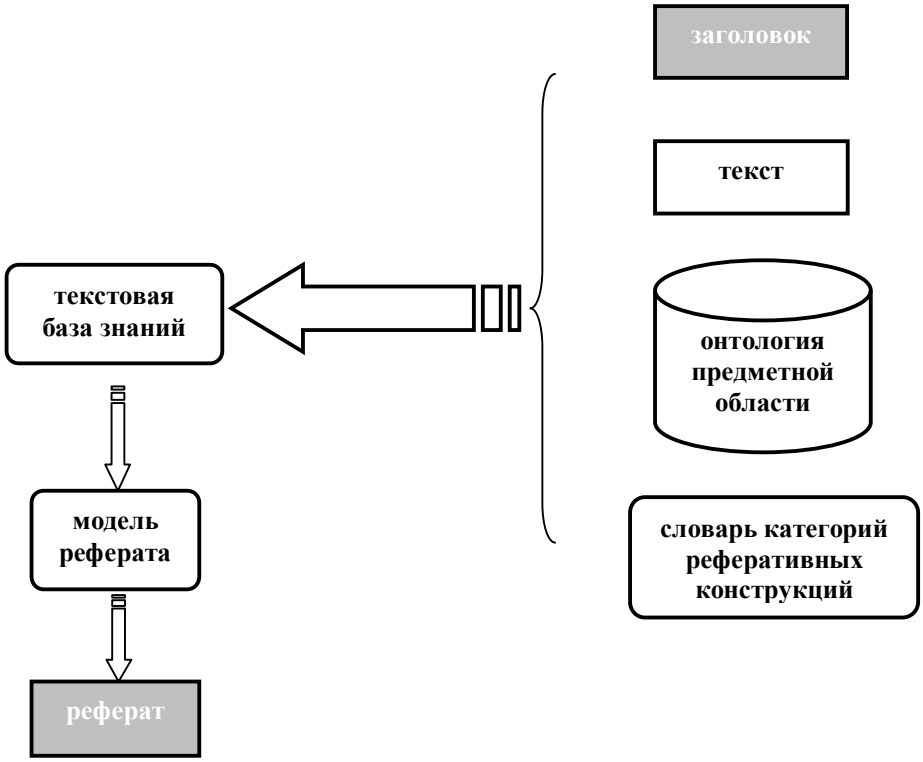
СҐЗ = Об Sr V(m_5) A(m_4) A(m_9)

Об одной реализации языка запросов в системе управления базой данных.

Проведенные исследования позволили нам увидеть четкое продолжение процедуры обобщения в заголовке в сравнении с процедурой обобщения в реферате как на семантическом, так и на синтаксическом уровнях. При сохранении тех же смысловых составляющих и в реферате, и в заголовке они имеют вместе с тем синтаксические и грамматические особенности их выражения, которые повышают уровень обобщения текста заголовка.

Наличие одних и тех же семантических компонентов в заголовке, реферате и тексте, являющихся разными формами выражения одного и того же понятия, позволяет описать смысловые структуры словосочетаний на разных уровнях обобщения информации. А это, в свою очередь, позволило построить модель реферирования в виде процедуры перехода от заголовка к тексту и дальше к реферату в ходе его содержательного конструирования.

Для оптимизации процедуры автоматического извлечения именных групп из текста в нашей системе необходимо использование онтологий верхнего уровня и предметных областей. Онтология верхнего уровня представляет собой вырожденную онтологию в виде словаря категорий реферативных конструкций – *объект*, *результат*, *цель*, *инструмент*. Онтология предметной области представляется в виде таксономии понятий конкретной области знаний [3]. Для построения текстовой базы знаний мы отталкивались от понятий, содержащихся в заголовке документа, по которым отыскиваются соответствующие им именные группы в тексте. В результате сопоставления терминов из заголовка с имеющимися онтологиями формируются цепочки именных групп для реферативных конструкций.



Выводы. Анализ смысловой и синтаксической структур заголовка показал, что заголовок является аналогом индикативного реферата в максимально сжатом виде, что позволило нам рассматривать заголовок в качестве отправной точки в разработке системы автоматического реферирования.

Литература

1. Лазаренко О. В., Яковенко А. А. Моделювання процессу узагальнення в системі автоматичного реферування : монографія / О. В. Лазаренко, А. А. Яковенко. – Х. : Вид-во НУА, 2007. – 124 с.
2. Лазаренко О. В., Попова Т. В. Аналіз смислової структури заголовка як тексту з максимальним рівнем узагальнення : сб. науч. трудов / О. В. Лазаренко, Т. В. Попова // Вып. 12 "Проблеми семантики слова, речення та тексту". – К. : КНЛУ, 2004. – С. 143–149.
3. Лазаренко О. В., Панченко Д. И. Роль онтологий при обработке знаний в "Семантическом Web" // Лінгвістичні студії : зб. наук. пр. Вип. 18 / укл. : Анатолій Загнітко (наук. ред.) та ін. – Донецьк : ДонНУ, 2009. – С. 258–262.

Статья сдана в редколлегию
17.06.2009 г.