УДК 81’322.2’33

О.В. Лазаренко[1], Д.И. Панченко [2], Е.Ю. Буряк[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

РАЗРАБОТКА АЛГОРИТМА СМЫСЛОВОГО АНАЛИЗА ТЕКСТА ДЛЯ СИНТЕЗА РЕФЕРАТА В СИСТЕМЕ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ

Предложена процедура синтеза реферата с использованием текстовой базы и модели заголовка в процессе смыслового анализа текста в системе автоматического реферирования. Описывается алгоритм синтеза реферата на смысловом уровне. Предлагается процедура создания текстовой базы для синтеза реферативных конструкций.

АВТОМАТИЧЕСКОЕ РЕФЕРИРОВАНИЕ, СИНТЕЗ РЕФЕРАТА, ТЕКСТОВАЯ БАЗА, МОДЕЛЬ ЗАГОЛОВКА, МОДЕЛЬ РЕФЕРАТА

Введение

Разработка интеллектуальной системы автоматического реферирования предполагает исследование различных аспектов понимания текста. Как отмечалось в наших работах [1, 2, 3], прежде чем приступить к исследованию этих вопросов, мы рассмотрели наиболее общие закономерности реферирования, выраженные в конечном продукте – реферате, сознательно, в связи с чрезвычайной сложностью процесса реферирования, допустив на начальном этапе теоретическую и эмпирическую неполноту. С тем, чтобы в дальнейшем, оттолкнувшись от понимания изученных закономерностей, расширять область исследования.

В связи с этим нами были изучены особенности синтаксической и семантической структур реферата, построена модель компрессии первичного текста на всех уровнях, сформулированы правила порождения реферативных предложений и на их основе разработана модель унифицированного реферата в виде типовых для индикативных рефератов семантико-синтаксических конструкций.

Для наполнения модели реферата соответствующей семантикой была разработана процедура семантического анализа текста с целью сжатия его смысла, что позволило построить семантико-контекстную модель реферирования, включающую модель заголовка и текстовую базу. Текстовая база является «информационным ядром» текста, и содержит информацию о ситуации, описанной в тексте. В текстовую базу входят предложения, отражающие основные смысловые аспекты исходного текста.

В рамках нашего подхода заголовок рассматривается как реферат минимального объема или как текст с максимальным уровнем компрессии содержания. Сравнительный анализ компрессии в заголовке и реферате показал продолжение процедуры свертывания (компрессии) смысла текста в заголовке в сравнении с процедурой компрессии в реферате как на семантическом, так и на синтаксическом уровнях. При сохранении тех же смысловых составляющих в реферате и в заголовке они имеют, однако, свои синтаксические и грамматические особенности их выражения.

На базе выявленных аналогий была предложена модель заголовка, отражающая сходство его с рефератом, с целью более эффективного использования этой модели в процессе автоматического реферирования. По этим результатам был разработан алгоритм наполнения модели реферата соответствующей семантикой с использованием промежуточных элементов – текстовых баз, необходимых для отбора предложений, содержащих указания на объект, результат, цель, метод и место, т.е. смысловые аспекты, образующие смысловую структуру реферата.

В основе построения текстовой базы лежит анализ заголовка и текста и выявление на его основе предложений для текстовой базы.

Основной целью наших исследований на этом этапе стала разработка алгоритма построения текстовой базы с опорой на семантико-контекстную модель текста, модель заголовка и модель реферата.

1. Разработка алгоритма анализа заголовка для построения текстовой базы

При разработке алгоритма выделения в тексте предложений для текстовой базы мы исходили из того, что анализ содержания текста должен включать анализ заголовка. Заголовки пишутся максимально сжато, лаконично, в них опущены все семантически второстепенные элементы. По сути, они являются смысловым инвариантом текста.

Поэтому для выбора предложений в текстовую базу в начале проводится анализ заголовка.

В заголовке выделяются слова, описывающие смысловые аспекты текста – *результат*, *объект* и, если есть, – *метод*, *цель*, *место*.

Общая синтаксическая конструкция заголовка(СКЗ) имеет вид:

$$\textbf{\textit{СКЗ:}}\ Pp - V(m_5) - A(m_4) - A(m_7) - A(m_9) - A(m_8),$$

где *Pp* – выражения с предлогами: *Об одном..., К вопросу о..., Еще раз о..., Ученые о...*и т.п. (аналог сирконстанта в модели реферата), $V(m_5)$ –отглагольное

существительное(аналог предиката) со значением *результат*, $A(m_4)$ – актант со значением *объект*, $A(m_7)$ – актант со значением *цель*, $A(m_9)$ – актант со значением *место*, $A(m_8)$ – актант со значением *метод*.

Все обозначения подробно описаны в [1].

Примеры заголовков:

СКЗ : $V(m_5) - A(m_4) - A(m_7) - A(m_9)$. *результат + объект + цель + место*

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

СКЗ : $Pp - V(m_5) - A(m_4) - A(m_7)$. *результат + объект + цель*

О формировании семантических признаков для математической модели префиксального словообразования.

СКЗ : $Pp - V(m_5) - A(m_4) - A(m_9)$. *результат + объект + место*

Об одной реализации языка запросов в системе управления базой данных.

Как видно из приведенных примеров конкретных реализаций общей структуры заголовка 5-4-7-9-8, возможны различные комбинации смысловых аспектов в заголовках: 5-4-7-9-0, 5-4-7-0-0, 5-4-0-9-0 и др.

Описание обобщенной структуры заголовка приведено в табл. 1.

В соответствии с этой структурой заголовок разбивается на составные элементы.

1. Выделяются выражения с предлогом в начале заголовка (если такие есть). Для этого используется составленный в процессе анализа заголовков словарь выражений, встречающихся в начале заголовка.

2. Определяется существительное в соответствующем падеже (родительном, дательном или предложном, если есть соответствующее выражение с предлогами, или именительном (если такого выражения нет) и стоящие слева от него слова в таком же падеже. Выделенные слова соответствуют смысловому аспекту m_5 в модели заголовка со значением *результат*.

3. Ищется следующее существительное в родительном падеже и предшествующие ему слова в таком же падеже. Выделенные слова соответствуют смысловому аспекту m_4 в модели заголовка со значением – *объект*.

Аналогичным образом выделяются в заголовке все имеющиеся в нем смысловые аспекты.

2. Разработка процедуры смыслового анализа текста для построения текстовой базы

В процессе анализа текста с целью выбора необходимой для построения реферата информации нам необходимо отобрать предложения, которые содержат указания на объект, результат, цель, метод и место, т.е. смысловые аспекты, образующие смысловую структуру реферата. Выбор этих предложений представляет определенные трудности в силу разнообразия описания этих аспектов в исходном тексте.

Более того, предложений, указывающих на определенный смысловой аспект, может быть несколько, поэтому целесообразно выделить их из текста, опираясь на результаты анализа заголовка, слова-указатели на смысловые аспекты в тексте и глаголы, указывающие, в первую очередь, на *объект* описания и *результат* исследования.

Анализ заголовка мы рассмотрели. Теперь рассмотрим, какие слова-указатели на интересующие нас смысловые аспекты встречаются в текстах. Чаще всего на *объект* и *результат* исследования указывают следующие выражения:

1. В (данной/настоящей) статье/работе...
2. Данная статья/работа/обзор ...
3. Целью данной/настоящей статьи/работы...

Далее идут глаголы:

1.... рассматривается / предлагается / описывается /исследуется / анализируется...

2.... посвящена / представляет собой ...

3....является ...

Примеры из текстов статей:

В данной статье рассматривается вопрос ...

В настоящей статье математически описываются

В статье рассматривается возможный подход...

В статье предлагается математическая модель...

В данной работе предлагаются и рассматриваются некоторые процедуры...

Данная статья посвящена проблемам...

Следует обратить внимание на слова-указатели *Целью данной/настоящей статьи/работы*.... В таких предложениях речь идет об *объекте* и *результате*, а не о *цели* исследования. На *цель* иссле-

Таблица 1

Pp			**Результат**		**Объект**		**Цель**	**Место**	**Метод**
Выражение с предлогом			Прилагательное	Существительное	Прилагательное	Существительное	Предлог ***для…***	Предлог ***в…***	Существительное
–	–	–	**И.п.**	**И.п.**	**Р.п.**	**Р.п.**	**Р.п.**	**Пр.п.**	**Р.п.**
К	вопросу ...Еще раз.. Учителя..	*О(б)*, *о(б)*	Пр.п.	Пр.п.	Р.п.	Р.п.			
О(б)	*одном подходе...*	*К, к*	Д.п.	Д.п.	Р.п.	Р.п.			
О(б)	*одном методе*	_	Р.п.	Р.п.					
	Один из подходов (*вопросов, аспектов*)	_	Р.п.	Р.п.					

дования указывает, как было отмечено при анализе заголовка, предлог *для*.

Отталкиваясь от слов-указателей и глаголов, указывающих на смысловые аспекты в тексте, необходимые для построения реферата, выбираются предложения для текстовой базы.

Для этого анализируется текст статьи.

1. Ищутся слова-указатели и глаголы, указывающие на нужные для реферата смысловые аспекты.

2. Выделяется существительное в соответствующем словам-указателям падеже и стоящие слева от него слова в таком же падеже, а также все слова, стоящие справа от выделенного существительного.

3. Выделенные в п.2 слова сравниваются со словами из заголовка, обозначающими результат и объект.

4. Если имеется совпадение (хотя бы частичное), предложение записывается в текстовую базу.

Таким образом, анализируются предложения из первых и последних абзацев текста, в которых обнаруживаются слова-указатели и соответствующие глаголы, а также именные группы, совпадающие со словами из заголовка.

Такая процедура во многом пересекается с концепцией понимания текста, предложенной в работах голландского лингвиста ван Дейка. «Быстрый анализ поверхностных структур и выстраивание относительно простой и жесткой семантической конфигурации» представляют собой обобщенное описание основного содержания дискурса, которое читатель строит в процессе понимания, и являются фактически рефератом или резюме. Разработка макро правил обобщенного описания содержания текста позволяет построить формальный переход от исходного текста к реферату [4].

Такие правила были разработаны нами на этапе изучения семантико-синтаксической структуры реферата и легли в основу процедуры построения реферативных предложений.

Разработанная на первом этапе исследований модель реферата была положена в основу первой версии компьютерной программы АвтоРеферат, в которой осуществлялось порождение реферативного предложения, но пока без глубинного анализа смысла первичного текста.

Разработка процедуры смыслового анализа исходного текста позволила перейти к выбору предложений, содержащих необходимые для реферата смысловые аспекты.

Если в первой версии системы АвтоРеферат мы опирались на заголовок и частоту встречаемости слов в тексте для порождения реферативных предложений в соответствии с разработанной моделью реферата, то на нынешнем этапе за основу берутся предложения из текстовой базы и заголовка, на основе которых строятся реферативные предложения.

Алгоритм заполнения реферативного предложения соответствующими актантами и предикатам в главном остался неизменным, однако ушли издержки частотного подхода.

Так, на первом этапе при частотном подходе осуществлялся поиск наиболее часто встречающегося слова путем сравнения каждого слова из заголовка с каждым словом текста, и, в случае совпадения, подсчитывалась абсолютная частота встречаемости этого слова в тексте. На основании всего этого выбиралось слово с максимальной частотой в качестве основного для построения актантной цепочки в реферативном предложении.

При использовании текстовой базы необходимое слово выбирается из ограниченного набора предложений, входящих в текстовую базу.

На втором этапе выбранное слово помещается в конец реферативного предложения, заполняя актант *А*1 для последующего построения актантной цепочки, являющейся семантической основой реферативной конструкции.

Алгоритм синтеза реферата работает в двух блоках: условно блок 1 можно назвать – «поиск влево», блок 2 – «поиск вправо», то есть в одном блоке анализируются слова, стоящие слева от выбранного слова, а в другом блоке – слова, стоящие справа. По результатам этого анализа формируется актант. Далее все последующие актанты добавляются справа от предыдущих. В реферативных текстах предложения могут включать в себя различное количество актантов (от одного до четырех). На этом заканчивается построение актантной цепочки для первого предложения реферата.

На третьем этапе завершается построение всего реферативного предложения. Поскольку на входе есть уже готовая актантная цепочка, отражающая основной смысл предложения для реферата, главной задачей этого этапа является выбор предикативного ядра. В первой версии системы выбор предиката осуществлялся автоматически из словаря, в который вошли глаголы, не несущие важной семантической информации и являющиеся взаимозаменяемыми, без соотнесения с реальными значениями по тексту оригинала. Для проверки правильности синтеза реферативных предложений, что являлось главной задачей первой версии системы автоматического реферирования АвтоРеферат, это было допустимо, поскольку семантический блок не был еще разработан. Но для полноценной семантической обработки необходим точный выбор предиката. При оформлении окончательного варианта реферативных конструкций с помощью предикатов из других словарных групп, действует совсем другой механизм, который предполагает поиск необходимого слова в исходном тексте и анализ его окружения для определения уровня компрессии смысла, и только после этого можно использовать его в реферативной конструкции. При использовании текстовой базы эта задача была решена.

Однако не стоит думать, что нам удалось разрешить все вопросы автоматизации процесса реферирования.

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.

Более того, использование текстовых баз, задающих контекстную семантику текстов, и формируемых на их основе ситуационных моделей, а также заголовков всех текстов, описывающих схожие ситуации, позволяют выделить наиболее важные характеристики определенной ситуации. Эти характеристики, представляющие 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, образуют инвариантную репрезентацию ситуации. В результате последних исследований мы вплотную подошли к моделированию процесса реферирования на уровне глубинной семантики текста.

Выводы

В статье рассмотрена процедура смыслового анализа текста и заголовка, позволяющая обеспечить более качественный результат автоматического реферирования за счет использования информации, необходимой для наполнения смысловой структуры реферата.

Описан алгоритм синтеза реферата на смысловом уровне.

Проведено сравнение процедуры реферирования первой версии системы АвтоРеферат со второй версией системы, работающей на основе текстовых баз.

Отмечается необходимость дальнейших исследований для автоматизации процедуры построения ситуационных моделей и инвариантной репрезентации ситуации, обеспечивающих анализ глубинной семантики текста.

Список литературы: 1. *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. **2.** *Лазаренко О.В.* Семантіко-контекстна модель реферування / О.В. Лазаренко, Д.І. Панченко // Бионика интеллекта: науч.-техн. журнал. 2014. № 1(80). С. 19-24. **3.** *Лазаренко О.В.* Разработка интеллектуальной системы автоматического реферирования с использованием текстовых баз и ситуационных моделей / О.В. Лазаренко // MegaLing'2013. Горизонти прикладної лінгвістики та лінгвистичних технологій : доп. міжнар. наук. конф., Україна, Київ, 20-23 листопаду 2013 г. **4.** *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.

Поступила в редколлегию 29.05.2015

УДК 81'322.2'33

Розробка алгоритму змістовного аналізу тексту для синтезу реферату в системі автоматичного реферування / О.Ю. Буряк, О. В. Лазаренко, Д.І. Панченко // Біоніка інтелекту: наук.-техн. журнал. – 2015. – № 2 (85). – С. 127–130.

У статті розглянуто питання розробки алгоритму аналізу заголовка та тексту в системі автоматичного реферування. Запропонована процедура створення текстової бази для синтезу реферативних конструкцій.

Табл. 1. Бібліогр.: 4 найм.

УДК 81'322.2'33

The development of the text semantic analysis algorithm for summary synthesis in the automatic summarization system / O. Y. Buriak, O. V. Lazarenko, D. I. Panchenko // Bionics of Intelligense: Sci. Mag. – 2015. – № 2 (85). – P. 127–130.

The development issues of the heading and text analysis algorithm in the automatic summarization system are considered. The procedure of a text database creation to synthesize abstract formations is offered.

Tab. 1. Ref.: 4 items.