УДК 811.111
ББК Ш143.21-1

БАЗА ЗНАНИЙ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ СОДЕРЖАНИЯ РЕФЕРАТА

С.О. Шереметьева, П.Г. Осминин

Описывается методика построения базы знаний для анализирующего модуля системы автоматизированного реферирования научно-технических текстов. База знаний строится на основе анализа предметной области и ориентирована на автоматизацию определения содержания реферата по тексту статьи.

Ключевые слова: автоматизированное реферирование, определение содержания, база знаний, научно-технический текст.

1. Введение

Автоматическое реферирование позволяет оперативно обрабатывать большие объемы информации, что особенно важно при постоянно возрастающих объемах информации в современных условиях. Исследования в этой области продолжаются уже более 50 лет. Большой вклад в разработку систем автоматического реферирования внесли такие отечественные ученые, как В.П. Леонов [8], Э.Ф. Скороходько [9], В.А. Яцко [11], С.А. Тревгода [10], а также зарубежные специалисты Г.П. Лун [2], Г. Эдмундсон [1], Д. Марку [3], Д. Радев [4].

По способу построения текста реферата все методы автоматического реферирования делятся на три группы: извлекающие, генерирующие и смешанные.

Извлекающие методы реферирования формируют реферат, извлекая без изменения наиболее значимые фрагменты документа (предложения, абзацы) в порядке их появления в тексте. Мерами значимости фрагментов считаются наличие в них наиболее частотных слов, расположение фрагмента в документе (заголовки, первые и последние предложения абзацев), присутствие сигнальных слов и выражений, таких как «важно», «определенно», «в частности», присутствие слов из заголовка или подзаголовка [1, 2].

Генерирующие методы предполагают автоматическое определение содержания реферата с последующей генерацией нового текста, не представленного явно в тексте исходного документа [4]. При использовании генерирующих методов текст реферата строится, основываясь на правилах, предполагающих наличие лингвистической базы знаний. Генерирующие методы позволяют получать более качественные, чем извлекающие методы, результаты, но из-за сложности построения лингвистических баз знаний недостаточно разработаны.

В связи с этим большое распространение получили смешанные подходы, использующие для построения текста реферата комбинации извлекающих и генерирующих приемов реферирования.

Несмотря на большое количество исследований, проблема автоматического реферирования до сих пор не решена, так как естественный язык характеризуется неоднозначностью, неограниченностью и чрезвычайно сложно поддается формализации.

Наиболее многообещающие подходы по автоматической обработке естественного языка в общем и автоматическому реферированию в частности ориентированы на подъязык, который имеет ограниченный словарь и грамматику, характерные для текстов конкретной предметной области.

В статье делается попытка внести определенный вклад в решение проблемы автоматизированного реферирования и предлагается методика разработки базы знаний анализирующего модуля системы, ориентированной на автоматизацию определения содержания реферата по тексту статьи ограниченной предметной области. Методика иллюстрируется на материале научных статей по математическому моделированию.

2. Формальные критерии включения информации в реферат

Реферат научной статьи представляет собой краткое точное изложение ее содержания, включающее основные фактические сведения и выводы, без дополнительной интерпретации или замечаний автора реферата [7].

База знаний разрабатываемой нами модели ориентирована на решение следующих проблем:

1) отобрать информацию из текста статьи для включения в реферат (анализ);

2) представить найденную информацию в виде грамматически правильного и логически связного текста реферата (синтез).

В центре внимания настоящей статьи находится первая из названных проблем, которая решается на основе анализа подъязыка на примере предметной области математического моделирования. Цель анализа состоит в выявлении формальных критериев определения в тексте статьи фрагментов релевантного для реферата содержания и принятия решения относительно методики реферирования: извлечения или генерации, в зави-

симости от того, какие из этих приемов используются человеком при создании реферата.

Анализ корпуса текстов по математическому моделированию объемом 105 статей показал, что статьи либо представляют собой сплошной текст, либо структурированы и включают такие разделы, как «Введение», «Заключение», «Выводы», а также разделы, озаглавленные в соответствии с рассматриваемой в них проблемой. Исследование пересечений множеств предложений авторских рефератов со множествами предложений соответствующих статей проводилось с учетом следующих параметров:

- текстовая презентация одного и того же содержания в реферате и статье;
- локализация в статье фрагмента, релевантного для реферата;
- лексические и другие маркеры релевантной для реферата информации.

Было выявлено, что отдельные предложения реферата могут: а) совпадать с предложениями из статьи полностью, б) совпадать частично и в) не совпадать вообще. В случаях б) и в) при построении реферата частично или полностью используются лексико-грамматические средства, отличные от тех, что использованы в статье.

Тексты рефератов могут полностью состоять из предложений, извлеченных из статьи (полное извлечение), представлять собой комбинацию фрагментов статьи и нового текста, возможно даже в рамках одного предложения (сочетание извлечения и генерации), а также не содержать предложений статьи вообще (полная генерация).

В таблице приведены количественные данные о предпочтительных способах построения рефератов человеком и данные о локализации предложений, передающих фрагменты содержания рефератов. Для неструктурированных статей мы ввели условные разделы «Начало» (первые два абзаца статьи), «Конец» (последние два абзаца) и «Середина» (текст статьи, который находится между «Введением/Началом» и «Заключение/Концом»).

Большая часть рефератов – 54,3 % составлена авторами полным извлечением предложений из текста. В 36,2 % случаях авторы обрабатывали извлеченные предложения из текста – перефразировали, опускали второстепенную информацию, добавляли новый текст, т. е. сочетали приемы извлечения и генерации. В 9,5 % случаев авторы составляли рефераты только из нового текста полной генерацией. Так как большинство рефератов составлено авторами из предложений статьи и/или отредактированных фрагментов статьи, то наш метод отбора информации для включения в реферат ориентирован на извлечение текстовых фрагментов статьи релевантного для реферата содержания с последующей их обработкой, т. е. на сочетание методик извлечения и генерации.

В соответствии с требованиями ГОСТа [7] в тексте реферата достаточно четко выделяются три информационные части: «Проблема» – информация о предмете, теме и цели работы, «Метод» – информация о методе или методологии проведения работы, «Результат» – информация о результатах работы, области применения результатов. В статье информация о проблеме, методе и результатах исследования может повторяться в различных разделах и в различной языковой репрезентации, в то время как в реферате каждый тип информации (проблема, метод, результат) представляется один раз. Порядок изложения в реферате может не соответствовать порядку представления соответствующих типов информации в статье. Одно и то же содержание в статье и реферате часто выражается языковыми сегментами разной длины и структуры. Например, о полученных результатах, как правило, говорится во Введении/Начале статьи, а метод получения этих результатов указывается в Заключении/Конце. В авторском реферате предложения, содержащие информацию о результатах и методах их получения, часто сливаются в одно, как правило, опускаются начальные слова и убираются ссылки на литературу.

В статье каждый из перечисленных выше трех типов информации часто сопровождается языковыми маркерами, такими как «в работе», «показано», «впервые», «изучено» и др. для описания проблемы; «реализация», «способ» и др. для описания метода; «доказать», «установить» для описания результатов. Отметим, что в нашей работе термин «маркер» обозначает лексемы широкого значения, наиболее часто повторяющиеся в информационных частях реферата, что отличается от общепринятого понимания маркеров только как устойчивых выражений типа «во-первых», «во-вторых» и т. д. Анализ показал, что множество всех маркеров в статье не только делится на 3 категории, соответствующие определенному типу информации, включенной в реферат, но в каждой из категорий маркеры имеют различную семантико-информационную природу и выражаются различными морфо-синтаксическими средствами.

Способ построения и локализация содержания рефератов

Способ построения реферата	Кол-во, %	Раздел статьи с информацией для реферата	Кол-во предложений, %
Полное извлечение	54,3	Введение/Начало	57,4
Извлечение/Генерация	36,2	Середина	30,8
Полная генерация	9,5	Заключение/Конец	11,8
Рефератов всего	100	Предложений всего	100

При этом лексические, семантико-информационные и морфо-синтаксические свойства маркеров находятся в характерной для каждой категории корреляции. Например, маркеры, обозначающие объекты, выражаются существительными и местоимениями; маркеры, обозначающие отношения между объектами – глаголами, а маркеры, обозначающие атрибуты объектов и отношений, выражаются прилагательными, местоимениями, наречиями. В предложениях статьи маркеры могут функционировать либо в качестве самостоятельных лексем, либо быть частью более длинных групп. В последнем случае маркеры в зависимости от категории являются частью именных групп, глагольных групп или групп прилагательных, содержащих термины предметной области математического моделирования.

Важным результатом анализа подъязыка является информация о совместном появлении, локализации и линейной последовательности маркеров различного типа во фрагментах статьи, включающих содержание, релевантное для определенных информационных частей реферата. Нами выдвинута гипотеза о том, что в качестве формальных критериев определения фрагментов статьи, содержащих релевантную для реферата информацию, можно использовать наличие в предложениях статьи маркеров в определенной семантико-информационной и структурной корреляции, эмпирически определенной в результате анализа подъязыка.

3. Структура базы знаний анализирующего модуля

По результатам анализа подъязыка построены важнейшие компоненты базы знаний анализирующего модуля модели, используемой для автоматизации определения содержания реферата: информационно-концептуальная сеть в виде корневого дерева и множество шаблонов, кодирующих знания о допустимой совместной встречаемости и линейной последовательности маркеров и других слов в релевантном для реферата фрагменте статьи.

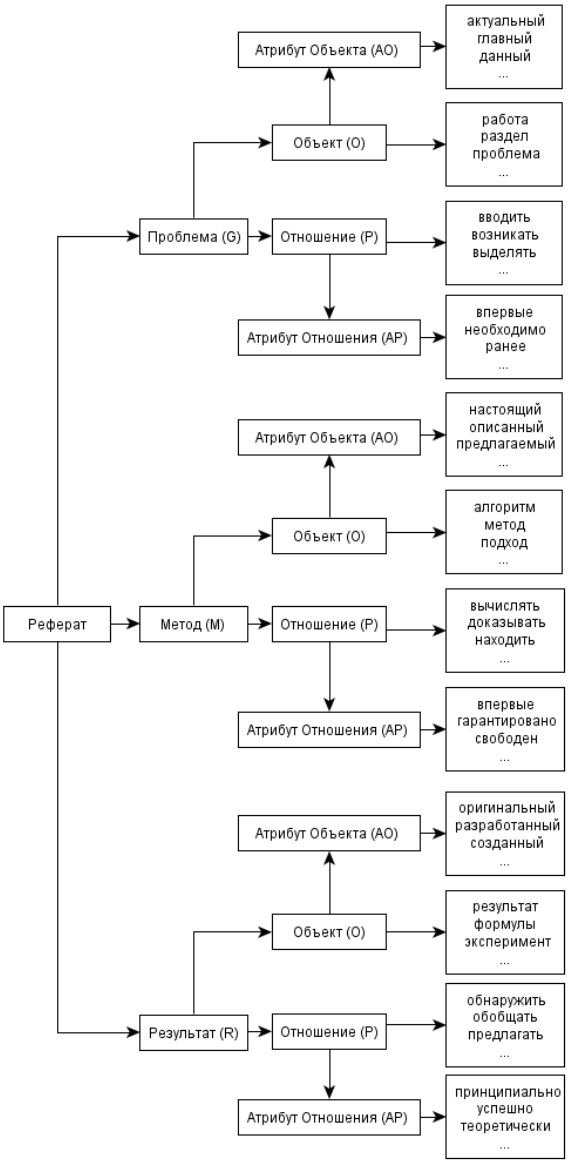
Информационно-концептуальная сеть (см. рисунок) состоит из терминальных и нетерминальных узлов и дуг, реализующих отношение включения. Корнем дерева является концепт Реферат, в нетерминальных узлах находятся концепты, соответствующие информационным частям реферата «Проблема (G)», «Метод (M)», «Результат (R)» и семантическим типам маркеров «Объект (O)», «Отношение (P)», «Атрибут Объекта (AO)» и «Атрибут Отношения (AP)».

Термин «концепт» мы понимаем как «общее понятие» [6]. Объекты описывают то, о чем говорится в предложении, отношения описывают связи между объектами, атрибуты дают дополнительную характеристику объектам или отношениям. Терминальные узлы содержат лексемы, реализующие определенный тип маркера в подъязыке.

Концепты не зависят от конкретного языка, а их реализации – маркеры, естественным образом определяются лексическими средствами конкретного, в нашем случае русского языка.

К моменту подготовки настоящей статьи информационно-концептуальная сеть реферата содержит 275 терминальных узлов (маркеров), из них наиболее частотными являются маркеры отношений (112 лексем), следующими по частоте идут маркеры объектов (98), затем – маркеры атрибутов объектов (55) и маркеры атрибутов отношений (10).

Сетевая структура (см. рисунок) позволяет формализовать корреляцию лексических, семантико-информационных и морфо-синтаксических свойств маркеров, специфических для каждой информационной части реферата. Эта корреляция



Информационно-концептуальная сеть реферата

соответствует пути от терминального узла-маркера к вершине сети. Например, коррелирующие свойства маркера «актуальный» описываются сетевым кодом AOOG (Атрибут Объекта-Объект-Проблема), а сетевой код маркера «находить» – PM (Отношение-Метод).

Вторым компонентом базы знаний анализирующего модуля является множество шаблонов, кодирующих знания о допустимой совместной встречаемости и линейной последовательности маркеров и других слов в релевантном для реферата фрагменте статьи. Шаблоны имеют вид:

шаблон	::= (ИЧ (структура))
ИЧ	::= {проблема, метод, результат}
структура	::= (X Группа X … Группа …X)
X	::= (слово слово ... слово)
группа	::= {NP(МАРКЕР T), VP(МАРКЕР T), AP(МАРКЕР T)}
МАРКЕР	::= (маркер(сетевой код))

где ИЧ – информационная часть реферата; структура – структура фрагмента текста статьи; X – цепочка из последовательных слов фрагмента (может быть пустой);
маркер – терминальный узел сети; код – сетевой код; T – термин (может быть пустым); NP – именная группа, VP – глагольная группа, AP – группа прилагательного.

Ниже в качестве примера приведен один из шаблонов базы знаний:

*(Результат (X NP(**Маркер(OR)** T) X (AP(**Маркер(APR)** T) VP(**Маркер(PR)** **T**) X))*

Такому шаблону соответствует, например, такой фрагмент статьи:

*На основе полученного **результата** в среде Maple 6 **впервые создан программный продукт**, позволяющий численно находить приближенное решение поставленной обратной задачи.*

На момент публикации статьи база знаний содержит 12 шаблонов. Идентификация фрагментов статьи с возможно релевантным для реферата содержанием осуществляется путем а) идентификации маркеров с помощью сопоставления лексического состава статьи со множеством терминальных узлов информационно-семантической сети, б) определением фрагментов статьи, соответствующих шаблонам базы знаний и в) оценки веса фрагмента на основе экспериментально разработанной метрики. Вес фрагмента статьи зависит от наличия, количества корреляционных свойств, совместной встречаемости и линейной последовательности маркеров во фрагменте. Кроме этого на вес фрагмента влияет наличие в нем ключевых слов статьи.

Описание метрики взвешивания и алгоритмической процедуры извлечения фрагментов статьи, а также их последующей обработки при генерации окончательного текста реферата выходит за рамки настоящей статьи. Отметим только, что ключевые слова не входят в базу знаний анализирующего модуля системы реферирования, а определяются динамически для каждой статьи с помощью программы Lana-Key [5].

4. Заключение

В статье предложена методика разработки базы знаний анализирующего модуля системы реферирования, ориентированной на автоматизацию определения содержания реферата по тексту статьи, основанная на закономерностях подъязыка определенной предметной области. Определены основные параметры и процедура анализа подъязыка для построения базы знаний системы автоматизированного реферирования. Разработана структура формализованного представления результатов анализа в базе знаний системы, которая состоит из двух компонентов: а) информационно-концептуальной сети, кодирующей знания о корреляции лексических, семантико-информационных и морфо-синтаксических свойств маркеров и б) шаблонов, представляющих знания о допустимой совместной встречаемости, локализации и линейной последовательности маркеров в релевантном фрагменте статьи. Описана конкретная реализация предложенной методики на примере подъязыка научных статей по математическому моделированию.

Литература

1. Edmundson, H.P. New methods in automatic extracting / H.P. Edmundson // Newspaper of ACM tea, 16-2 (1969). – P. 264–285.

2. Luhn, H.P. The Automatic Creation of Literature Abstracts / H.P. Luhn // IBM Journal of Research and Development. – 1958. – V. 2, № 2. – P. 159–165.

3. Marcu, D. Improving summarization through rhetorical parsing tuning / D. Marcu // Proceedings of the Sixth Workshop on Very Large Corpora. – Montreal, Canada. – 1998. – P. 206–215.

4. Radev, Dragomir R. Generating natural language summaries from multiple on–line / Dragomir R. Radev, Kathleen R. McKeown // Computational Linguistics – Special issue on natural language generation. – 1998. – V. 24, № 3. – P. 470–500.

5. Sheremetyeva, S. Automatic Extraction of Linguistic Resources in Multiple Languages. Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012. – Wroclaw, Poland, 2012. – P. 44–52.

6. Большой толковый словарь русского языка / сост.: В.В. Виноградов и др.; под ред. Д.Н. Ушакова. – М.: АСТ: Астрель, 2008. – 1268 с.

7. ГОСТ 7.9-95. Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования. – М.: Изд-во стандартов, 1995. – 8 с.

8. Леонов, В.П. Реферирование и аннотиро-

вание научно-технической литературы / В.П. Леонов. – Новосибирск: Наука, 1986. – 176 с.

9. Приходько, С.М. Автоматическое реферирование на основе анализа межфразовых связей / Э.Ф. Скороходько, С.М. Приходько // НТИ. Сер. 2. – 1982. – № 1. – С. 27–32.

10. Тревгода, С.А. Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений: автореф. дис. ... канд. техн. наук / С.А. Тревгода. – СПб., 2009. – 18 с.

11. Яцко, В.А. Симметричное реферирование: теоретические основы и методика / В.А. Яцко // НТИ. Сер. 2. – 2002. – № 5. – С. 18–28.

Шереметьева Светлана Олеговна, доктор филологических наук, доцент, профессор кафедры «Лингвистика и межкультурная коммуникация», Южно-Уральский государственный университет (г. Челябинск). E-mail: linklana@yahoo.com

Осминин Павел Григорьевич, аспирант кафедры «Лингвистика и межкультурная коммуникация» Южно-Уральский государственный университет (г. Челябинск). Научный руководитель – доктор филологических наук, профессор С.О. Шереметьева. E-mail: osperevod@gmail.com

Bulletin of the South Ural State University
Series "Linguistics"
2013, vol. 10, no. 2, pp. 77–81

KNOWLEDGE BASE FOR AUTOMATED EXTRACTION OF SUMMARY CONTENT

***S.O. Sheremetyeva**, South Ural State University, Chelyabinsk, Russian Federation, linklana@yahoo.com*
***P.G. Osminin,** South Ural State University, Chelyabinsk, Russian Federation, osperevod@gmail.com*

The present paper focuses on a methodology of constructing a knowledge base for the analysis module of an automated summarization system for scientific and technical texts. The knowledge base is constructed based on sublanguage analysis and is oriented to automated extraction of summary content from the document text.

Keywords: automated summarization, content extraction, knowledge base, scientific and technical texts.

Поступила в редакцию 21 марта 2013 г.