УДК 81'322
ББК Ш111

МОДЕЛЬ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ НА ОСНОВЕ БАЗЫ ЗНАНИЙ, ОРИЕНТИРОВАННАЯ НА АВТОМАТИЧЕСКИЙ ПЕРЕВОД

П.Г. Осминин

Представлена модель автоматического реферирования для научно-технических текстов, ориентированная на автоматический перевод. Модель состоит из трех основных компонентов: экстрактора ключевых слов, базы знаний и алгоритма автоматического реферирования. Текст реферата генерируется в форме, исключающей лингвистические явления, которые могут вызвать проблемы при автоматическом переводе (контролируется синтаксическая сложность предложения и ограничивается его длина, не допускаются сложные придаточные предложения и эллипсис). Правила генерации текста реферата описывают используемую грамматическую структуру предложений. Алгоритм автоматического реферирования состоит из четырех основных процедур – предварительной обработки и анализа текста статьи, отбора содержания реферата, генерации текста реферата.

Ключевые слова: автоматическое реферирование, автоматический перевод, извлечение информации, база знаний.

1. Введение

Рефераты научных статей служат приоритетным средством обмена информацией – ученые и специалисты различных отраслей знаний используют рефераты для оперативного знакомства с последними публикациями и в этом заключается коммуникативная функция рефератов. Кроме этого, рефераты в профессиональной коммуникации выполняют следующие важные функции: отражают основное содержание документа (информативная функция), используются в информационно-поисковых системах и содержат информацию об авторах, заглавии документа (поисковая функция и справочная функция) и сигнализируют о наличии документа (сигнальная функция) [5].

Количество публикаций постоянно растет, и в связи с этим возрастает необходимость разработки методов автоматического реферирования. При этом в последнее время необходимым требованием для публикации научных статей в ведущих рецензируемых журналах является перевод реферата статьи на английский язык. Следовательно, возрастает необходимость повышения оперативности перевода рефератов, для чего все чаще применяются системы автоматического перевода, качество которых оставляет желать лучшего. Тем не менее, многих проблем при автоматическом переводе можно избежать, если заранее ориентировать текст на машинную обработку. На качество автоматического перевода отрицательное влияние оказывают большая длина и синтаксическая сложность предложений исходного текста, большое количество вводных слов, эллипсис, однородные члены, неоднозначные лексические единицы и т. п. [2]. Указанные признаки часто встречаются в научно-технических статьях и их авторских рефератах. Реферат, который предполагается переводить автоматически, должен иметь структуру, позволяющую избежать перечисленных проблем.

В данной статье мы представляем модель автоматического реферирования научных статей, которая, с одной стороны, позволит автоматизировать процесс реферирования на русском языке, а с другой – позволит получить текст, пригодный для дальнейшей автоматической обработки, в частности, автоматического перевода. Модель основана на сочетании извлекающих и генерирующих методов автоматического реферирования.

2. Модель автоматического реферирования

При построении модели мы уделяли особое внимание двум основным проблемам: 1) технике извлечения корректного содержания реферата и 2) представлению извлеченного содержания в виде текста, позволяющего избегать проблем при автоматическом переводе.

Модель построена на основе сравнительного анализа полных текстов научных статей и соответствующих авторских рефератов. Материалом для исследования послужили корпусы 107 полных научных статей (объем 203729 словоформ, без учета библиографических списков) и соответствующих им авторских рефератов (объем 4924 словоформ), средняя компрессия полных текстов статей составила 41,4.

В соответствии с требованиями ГОСТа [4] в тексте реферата достаточно четко выделяются четыре информационные части: «Тема» – информация о предмете и теме статьи, «Цель» – информация о цели работы, «Метод» – информация о методе или методологии проведения работы, «Результат» – информация о результатах работы, области применения результатов.

В тексте статьи каждый из перечисленных выше четырех типов информации, как правило, сопровождается лексическими маркерами. Для описания темы статьи могут использоваться, например, маркеры «содержать», «глава», «раздел» и т.д., для описания метода исследования могут применяться маркеры «прием», «инструмент», «находить» и т.д. Мы разбили все лексические маркеры на группы, соответствующие указанным информационным частям – «Тема», «Цель», «Метод» и «Результат». Внутри каждой группы маркеры делятся на следующие семантические типы: объекты, атрибуты объектов, отношения, атрибуты отношений. Объекты описывают предмет, отношения описывают связи между объектами, атрибуты описывают признаки объектов или отношений. Методика проведения анализа и более подробные результаты изложены в работе [7].

2.1. База знаний

База знаний модели автоматического реферирования состоит из следующих основных компонентов: 1) правил анализа полного документа, 2) стоп-лексикона, 3) информационно-концептуальной сети в виде корневого дерева, 4) множества шаблонов для извлечения информации и правил генерации текста реферата.

Правила анализа полного документа описывают порядок выполнения обработки документа.

Стоп-лексикон используется для удаления из полного текста статьи нерелевантной для реферата информации с целью облегчения дальнейшего анализа.

Информационно-концептуальная сеть используется для семантического анализа – выделения лексических маркеров в тексте документа. Более подробное описание информационно-концептуальной сети дано в работе [7].

Шаблоны для извлечения информации представляют собой фреймовые структуры:

Шаблон	::= (ИЧ (структура))
ИЧ	::= {тема, цель, метод, результат}
Структура	::= (X Группа X … Группа …X)
X	::= (слово слово ... слово)
Группа	::= {NP(МАРКЕР Т), VP(МАРКЕР Т), AP(МАРКЕР Т)}
Маркер	::= (маркер(сетевой код))
Номер шаблона	::= (натуральное число)
Номер предложения	::= (натуральное число)
Вес	::= (натуральное число)

Слоты фреймов при анализе текста по разработанной алгоритмической процедуре заполняются извлеченными текстовыми фрагментами. Заполненный шаблон выглядит следующим образом:

ИЧ	::= Метод
NP(Маркер(ОМ) Т)	::= Методом последовательных приближений
VP(Маркер(РМ) Т)	::= найдено
X	::= решение уравнения
Номер Шаблона	::= 7
Номер Предложения	::= 18
Вес	::= 3

Правила генерации текста разработаны так, чтобы облегчить последующий автоматический перевод реферата. При автоматическом переводе могут возникать ошибки как вследствие синтаксической сложности и многозначности, так и вследствие неполной покрываемости лексики и многозначности лексических единиц. Проблему неполной покрываемости лексики возможно решить с помощью специализированных словарей предметных областей, а проблемы многозначности и грамматической сложности можно снять на этапе композиции исходного текста для перевода. Поэтому в предлагаемой нами модели несмотря на то, что основу текста реферата составляют фрагменты, извлеченные из текста статьи, для получения окончательного текста проводится дополнительная обработка: удаление текстовых фрагментов, части оригинального текста могут быть перефразированы с целью исключить такие проблематичные явления как длинные предложения, эллипсис, однородные члены, длинные придаточные предложения [8].

На основе каждого шаблона, как правило, генерируется одно предложение. Если на одно предложение текста статьи можно наложить несколько шаблонов или шаблоны начинаются с одинаковых слов («рассматривается», «в работе»), то выполняется слияние шаблонов по определенным правилам и генерируется одно предложение.

Порядок расположения сгенерированных предложений в тексте реферата описывается следующим образом: предложения следуют в порядке их принадлежности к следующим информационным частям: Тема, Цель, Метод, Результат, причем в каждой из этих частей шаблоны упорядочиваются по весу, который складывается из веса ключевых слов и веса маркеров.

2.2. Алгоритм автоматического реферирования

Разработанный алгоритм автоматического реферирования выглядит следующим образом:

1. Процедура предварительной обработки текста статьи:

a) сегментация текста на предложения,

b) сжатие текста,

c) извлечение ключевых слов.

2. Процедура анализа текста статьи:

a) частичный морфосинтаксический анализ,

b) семантический анализ.

3. Процедура отбора содержания для реферата:

a) взвешивание предложений,

b) заполнение шаблонов.

4. Процедура генерации текста реферата.

Процедура предварительной обработки состоит из трех последовательных подпроцедур – сегментации текста на предложении, сжатия текста, извлечения ключевых слов.

Первой выполняется подпроцедура сегментации текста полной статьи на предложения. Предложением считается текстовый сегмент от точки до точки.

Затем подпроцедура сжатия текста выполняет удаление из текста неинформативных для реферата частей. На этом шаге применяется стоплексикон из базы знаний, после этого удаляются предложения, содержащие менее пяти слов (словом считается текстовый фрагмент от пробела до пробела).

После сжатия выполняется извлечение из текста ключевых слов с помощью экстрактора LanAKey_Ru [1], который способен извлекать именную лексику длиной до четырех слов без предварительной разметки текста. Ключевые слова в нашей модели – это наиболее релевантные именные группы (ИГ) статьи. Из текста полной статьи извлекаются 10 наиболее релевантных ключевых слов.

Процедура анализа состоит их двух подпроцедур – частичного морфосинтаксического анализа и семантического анализа. На этапе морфосинтаксического анализа происходит выделение лексических групп и определение их принадлежности к частям речи. Фразам, совпадающим с ключевыми словами, приписывается метка именной группы и вес (релевантность), автоматически определенный LanAKey_Ru. Затем морфосинтаксический анализ завершается с помощью программного обеспечения проекта Aot.ru [3].

В размеченном таким образом тексте статьи выполняется семантический анализ с помощью информационно-концептуальной сети. Лексикон, сопоставленный терминальным узлам сети, сравнивается с размеченным текстом, при совпадении маркеру присваивается сетевой код – путь от терминального узла-маркера к вершине сети.

Третья процедура – процедура отбора содержания, состоит из подпроцедур оценки веса (релевантности) предложений и заполнения шаблонов. Вес предложения (релевантность) определяется по следующей формуле:

$$W = 10N + M_i + K_i,$$

где W – вес предложения (релевантность);

N – количество ключевых слов в предложении, множитель 10 был получен эмпирически;

M_i – вес всех маркеров в предложении (вес одного маркера равен 10);

K_i – вес всех ключевых слов в предложении (вес из Lana-key_Ru).

Мы определили, что для малых текстов (до 9000 знаков с пробелами) порог отбора составляет 10 наиболее значимых предложений, для больших текстов – 10 % наиболее значимых предложений.

Подпроцедура заполнения шаблонов начинается с просмотра отобранных предложений по очереди слева направо. В шаблонах для извлечения информации задан порядок последовательности маркеров и других слов в предложении. Если предложение (или его часть) подходит под шаблон, то слоты шаблона заполняются частями предложения. Наложение шаблонов происходит на фрагменты текста между знаками препинания. Шаблоны налагаются в следующем порядке: Тема, Цель, Метод, Результат.

Пример формально составленного реферата статьи [6] приведен ниже.

Обратные спектральные задачи в различных постановках играют фундаментальную роль в различных разделах математики. В настоящей работе исследуется обратная задача для степени оператора Лапласа, порожденного краевой задачей Дирихле в случае непростого спектра.

Основным методом исследования является резольвентный метод.

Получены теоремы существования решения обратной задачи. Разработан вычислительный алгоритм восстановления потенциала по спектру.

В сгенерированном реферате дается описание темы работы, метода исследования и результатов работы. Информация о цели отсутствует, так как в исходном документе эта информация не была представлена. В реферате используются простые синтаксические конструкции, что облегчит последующий автоматический перевод.

Ниже представлен автоматический перевод сгенерированного реферата. Перевод выполнен системой AutoMath, которая разрабатывается в НОЦ «Лингво-инновационные технологии ЮУрГУ» для перевода текстов по математическому моделированию с русского языка на английский. Полученный автоматически перевод полностью корректен.

Inverse spectral problems in different formulations play a fundamental role in different branches of mathematics. The inverse problem for a degree of the Laplace operator generated by the Dirichlet boundary value problem for the case of a non-simple spectrum is investigated in the current work.

The resolvent method is the main method of investigation.

Existence theorems of the inverse problem are obtained. An algorithm for computing the recovery of the potential by spectrum is developed.

3. Заключение

В данной статье представлена модель автоматического реферирования, ориентированная на автоматический перевод. Описаны составные части модели: 1) экстрактор ключевых слов, 2) база

знаний и 3) алгоритм автоматического реферирования. Разработаны шаги формализации алгоритма автоматического реферирования. Правила генерации русского текста разработаны так, чтобы избегать использования проблематичных для автоматического перевода явлений.

Литература

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

2. Underwood, N.L. Translatability Checker: A Tool to Help Decide Whether to Use MT / N.L. Underwood, B. Jongejan // Proceedings of MT Summit VIII 18th–22nd September 2001, Santiago de Compostela. – P. 363–368.

3. Автоматическая обработка текста. – http://aot.ru/ (дата обращения: 20.02.2014)

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

5. Дубинина, Е.Ю. Компрессия научного текста: методы и модели: автореф. дис. ... канд. филол. наук / Е.Ю. Дубинина. – СПб., 2013. – 24 с.

6. Закирова, Г.А. Восстановление потенциала в обратной спектральной задаче для оператора Лапласа с кратным спектром / Г.А. Закирова // Вестник ЮУрГУ. Серия «Математическое моделирование и программирование». – 2010. – № 35 (211). – С. 25–28.

7. Шереметьева, С.О. База знаний для автоматического определения содержания реферата / С.О. Шереметьева, П.Г. Осминин // Вестник ЮУрГУ. Серия «Лингвистика». – 2013. – Т. 10, № 2. – С. 77–81.

8. Шереметьева, С.О. Интерактивное реферирование, ориентированное на машинный перевод / С.О. Шереметьева // Вестник Южно-Уральского государственного университета. Серия «Лингвистика». – 2013. – Т. 10, № 1. – С. 89–92.

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

Поступила в редакцию 18 марта 2014 г.

Bulletin of the South Ural State University
Series "Linguistics"
2014, vol. 11, no. 2, pp. 65–69

AUTOMATIC SUMMARIZATION MODEL ORIENTED TOWARD AUTOMATIC TRANSLATION (BASED ON THE KNOWLEDGE BASE)

Pavel G. Osminin, *South Ural State University, Chelyabinsk, Russian Federation; osperevod@gmail.com*

The present paper is concerned with a model of automatic summarization for scientific and technical texts, oriented toward automatic translation. The model consists of three main components: a keyword extractor, a knowledge base and a summarization algorithm. The summary text is generated in the form excluding linguistic phenomena that can cause problems during automatic translation (the syntactic complexity of the sentence is controlled and its length is limited, ellipsis and long subordinate clauses are not allowed). Rules for summary generation define the grammar of summary sentences. The summarization algorithm consists of four top level procedures – preprocessing and analysis of the article text, summary content selection and summary text generation.

Keywords: automatic summarization, automatic translation, information extraction, knowledge base.

References

1. 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* (2012, Wroclaw), pp. 44–52.

2. Underwood N.L. Translatability Checker: A Tool to Help Decide Whether to Use MT. *Proceedings of MT Summit VIII 18th–22nd September 2001, Santiago de Compostela,* pp. 363–368.

3. *Avtomaticheskaia obrabotka teksta* [Automatic Text Processing]. Available at: http://aot.ru/ (date of check: 20.02.2014).

4. *GOST 7.9-95. Sistema standartov po informatcii, bibliotechnomu i izdatelskomu delu. Referat i annotatciia. Obshchie trebovaniia.* [GOST 7.9-95 1995, System of Standards on Information, Librarianship and Publishing. Informative Abstract and Indicative Abstract. General Requirements], Moscow, Izd-vo standartov Publ., 1995, 8 p.

5. Dubinina E.Iu. *Kompressiia nauchnogo teksta: metody i model. Avtoref. dis. kand. filol. nauk* [Scientific Text Compression: Methods and Models. Abstract of Cand. Diss.], St. Petersburg, 2013, 24 p.

6. Zakirova G.A. [Potential's Restore in the Inverse Spectral Problem for Laplace Operator with Multiple Spectrum], *Bulletin of the South Ural State University. Mathematical Modelling, Programming & Computer Software Ser.*, 2010, № 35 (211), pp. 25–28. (in Russ.)

7. Sheremetyeva S.O., Osminin P.G. [Knowledge Base for Automated Extraction of Summary Content], *Bulletin of the South Ural State University. Linguistics Ser.*, 2013, vol. 10, №. 2, pp. 77–81. (in Russ.)

8. Sheremetyeva S.O. [On Interactive Summarization Oriented to Machine Translation], *Bulletin of the South Ural State University. Linguistics Ser.*, 2013, vol. 10, № 1, pp. 89–92. (in Russ.)

Pavel G. Osminin, postgraduate student of the Linguistics and Intercultural Communication department, South Ural State University (Chelyabinsk), osperevod@gmail.com. Scientific adviser – PhD (Habilitation), professor S.O. Sheremetyeva.

Received 18 March 2014