УДК 81'322

ИНТЕРАКТИВНОЕ РЕФЕРИРОВАНИЕ, ОРИЕНТИРОВАННОЕ НА МАШИННЫЙ ПЕРЕВОД

С.О. Шереметьева

ON INTERACTIVE SUMMARIZATION ORIENTED TO MACHINE TRANSLATION

S.O. Sheremetyeva

В статье описана методика разработки автоматизированной системы создания рефератов[1] в форме, соответствующей требованиям ГОСТа и облегчающей его последующий перевод на иностранный язык. Методика сочетает эмпирические и рациональные приемы обработки естественного языка и иллюстрируется на примере интерактивной системы для генерации рефератов научных статей по математическому моделированию. Реферат/аннотация генерируются на русском языке и сопровождаются выдачей английских эквивалентов использованной лексики.

Ключевые слова: автоматизированное реферирование, многоязычный лексикон, автоматический перевод.

The article describes a method of developing an automated system for creating summaries that meet state standard specification requirements and facilitate their subsequent translation into a foreign language. The approach is a combination of empirical and rational NLP techniques. It is illustrated by an interactive system for generating summaries of research papers on mathematical modeling. The output of the system is summaries in Russian and a list of Russian-English equivalents of the lexical units used in the summary.

Keywords: automated summarization, multilingual lexicon, machine translation.

Введение

Реферирование научно-технической литературы представляет собой важный вид профессиональной коммуникации, целью которого является оперативный обмен информацией между специалистами как в рамках одной страны, так и в международном масштабе [1].

Постоянно увеличивающаяся потребность в издании научно-технической литературы, подлежащей реферированию на различных языках, и возрастание стоимости ручного оформления документов и их перевода превращают автоматизацию реферирования в социальную и экономическую необходимость. Исследования по автоматизации реферирования как отечественными [2, 3], так и зарубежными лингвистами [4, 5] ведутся в рамках различных направлений и, несмотря на полувековую историю [6], далеки от завершения.

В настоящей работе предлагается новая методика разработки автоматизированной системы реферирования, а также ее конкретная реализация на примере системы для генерации рефератов научных статей по математическому моделированию на русском языке с выдачей английских эквивалентов использованной лексики[1].

Обоснование методики

Естественный язык настолько необъятен и неоднозначен, что создание высококачественных компьютерных систем его обработки требует от разработчиков бесконечно много времени и усилий.

В настоящее время достижение наиболее корректных результатов при автоматической обработке текстов возможно только в жестких рамках подъязыка в силу ограниченности его словаря и грамматики. Результаты анализа характеристик подъязыка позволяют выработать требования к конечному продукту работы системы, помогают выявить оптимальные для каждого подъязыка способы представления знаний, а также позволяют упростить или обойти многие проблемы автоматической обработки текстов, не разрешимые для все-

Шереметьева Светлана Олеговна, д-р филологических наук, доцент, профессор кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (Челябинск). E-mail: linklana@yahoo.com

Svetlana O. Sheremetyeva, PhD (Habilitation), professor of the Linguistics and Intercultural Communication department, South Ural State University (Chelyabinsk). E-mail: linklana@yahoo.com

[1] Все, сказанное в статье о реферировании, в равной степени относится к аннотированию.

го языка в целом. В основе нашей методики лежит ориентация на подъязык.

С целью экономии усилий и времени разработчиков предлагаемая методика предусматривает повторное использование отдельных компонентов программ, ранее разработанных для других языков и приложений, с их последующей адаптацией к выполнению новых задач и включением в разрабатываемую систему на определенных этапах обработки языкового материала.

Многие исследования по компьютерному реферированию концентрируются на разработке полностью автоматических систем. Однако в этом процессе достаточно часто невозможно обойтись без участия специалиста, по крайней мере, по двум причинам. Во-первых, в систему нужно ввести семантические знания (содержание реферата), на основе которых должен синтезироваться текст реферата, во-вторых, именно специалисты обладают знаниями относительно того, какое содержание должно быть отражено в реферате. Введение этих знаний в систему – не тривиальная задача. Поэтому одной их основных характеристик нашей методики является интерактивное взаимодействие пользователя с компьютером.

При разработке лингвистического обеспечения, которое включает в себя лексикографический и алгоритмический компоненты, мы следовали лексикалистскому подходу, при котором основная часть лингвистических знаний включается в лексикон, что повышает надежность системы.

Описанный подход был успешно апробирован при разработке интерактивной системы для генерации рефератов и аннотаций научных статей по математическому моделированию, описание которой приводится в следующем разделе.

Описание и реализация методики на примере системы РЕФЕРАТ

В этом разделе на примере системы РЕФЕРАТ приводится компьютерная реализация описанного выше подхода реферирования научно-технической информации. Система РЕФЕРАТ, архитектура которой дана на рис.1, предназначена для предметной области математического моделирования.

Подъязык рефератов по математическому моделированию отражает требования ГОСТа [7] к структуре реферата как такового и специфику предметной области математического моделирования.

По требованиям ГОСТа в тексте реферата следует четко и ясно излагать основные положения статьи, избегая сложных языковых структур и соблюдая единство терминологии. Это требование объясняется тем, что неправильное или сложное языковое оформление реферата даже при корректно отобранном содержании может привести к неправильному пониманию реферата и ошибкам при его переводе на иностранный язык. Длинные придаточные предложения, вставленные в основное предложение, причастные и деепричастные обороты и т. д. усиливают присущую естественному языку неоднозначность, добавляя к лексической омонимии омонимию синтаксическую. Поэтому система оформляет реферат статьи в форме предложений с простой синтаксической структурой и терминологией, использованной в исходной статье.

Специфика предметной области математического моделирования отражена в содержании лингвистической базы знаний системы, которая построена на основе анализа русского корпуса статей по математическому моделированию, опубликованных в «Вестнике ЮУрГУ» в 2008–2012 гг. и англоязычных статей сходной тематики, найденных в Интернете. Основная часть лингвистических знаний представлена в лексиконе системы.

Лексикографический компонент системы РЕФЕРАТ содержит русско-английский лексикон с информацией, необходимой для а) формальной фиксации знаний, б) алгоритмов анализа и синтеза текстов рефератов на русском языке, в) алгоритмов перевода одно- и многокомпонентной лексики (рис. 1).

Алгоритмический компонент содержит: а) алгоритмы обращения к лексикону, б) алгоритмы анализа научно-технической документации, предусматривающие перевод текстовой информации на формальный язык смыслов, в) алгоритмы синтеза текстов рефератов на русском языке и г) алгоритмы перевода одно- и многокомпонентной лексики на английский язык.

Для взаимодействия с пользователем разработан интерактивный модуль извлечения знаний.

Система РЕФЕРАТ повторно использует в качестве отдельных блоков некоторые модули программного обеспечения, ранее разработанного для английского языка [8, 9]. Эти модули были адаптированы для обработки русского языка в соответствии с задачами описываемого приложения.

В целом в систему РЕФЕРАТ входят следующие компоненты:

- *предметно-ориентированная база знаний,* которая включает лексикографический и алгоритмический компоненты
- *предметно-ориентированный анализатор* русских текстов, состоящий
  - из *автоматических модулей*, *выделяющих в тексте статьи именную (ИГ) и глагольную (ГГ)*[1] [10] *терминологию*. На выходе этого модуля выдается текст статьи в интерактивном формате, размеченный на именную терминологию и предикаты;
  - *интерактивного модуля синтаксического анализа* генерируемого реферата, который представляет отобранное автором содержание реферата в виде формальных структур представления знаний;

[1] ИГ – именная группа; ГГ – глагольная группа.

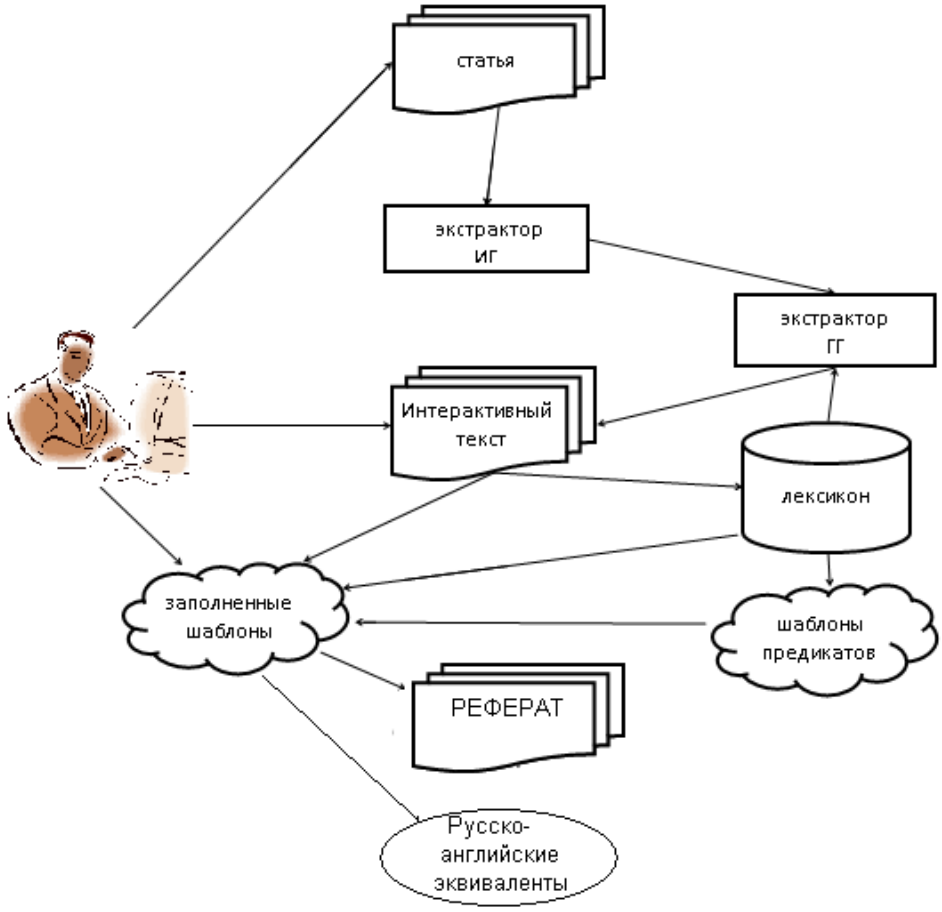

Рис. 1. Архитектура системы РЕФЕРАТ

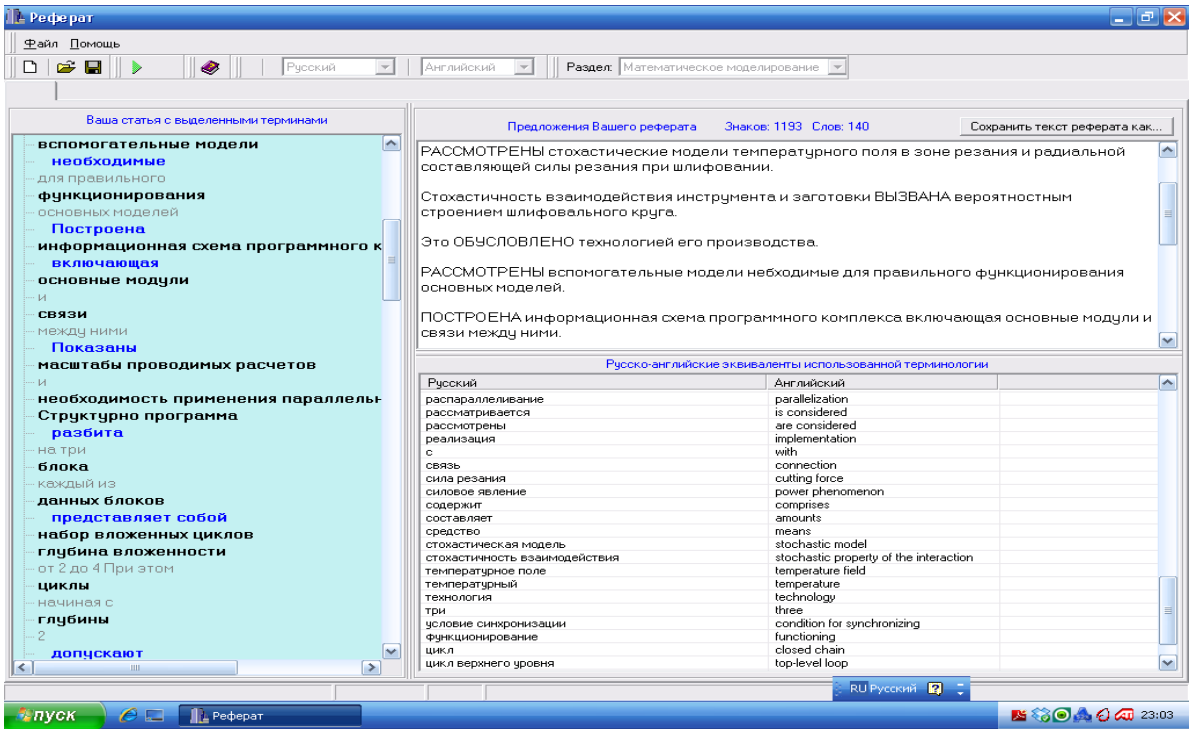

Рис. 2. Общий вид основного окна пользовательского интерфейса

o *автоматического модуля морфологического анализа;*

o *автоматический генератор предложений* реферата на русском языке.

Генерация текста реферата осуществляется на основе концептуальной схемы предложений реферата из базы знаний системы и на основе интерактивно извлеченной из пользователя информации.

Работа с программным инструментом РЕФЕРАТ осуществляется следующим образом. На вход системы подается текст статьи. РЕФЕРАТ автоматически анализирует полученный текст и представляет его в размеченном виде, акцентируя внимание автора на использованных терминах (именных группах и глаголах). При этом система автоматически генерирует первое предложение реферата и выдает интерактивный список использованных в статье глаголов, приведенных к форме, позволяющей использовать эти глаголы в качестве сказуемых остальных предложений реферата. После выбора автором определенного глагола выдается соответствующий шаблон будущего предложения реферата. Релевантные слоты шаблона заполняются автором путем автоматического переноса фраз из проанализированного системой текста в слоты шаблона. На основе заполненного шаблона генерируется грамматически правильное предложение и список русско-английских эквивалентов использованной одно- и многокомпонентной лексики. При этом английские эквиваленты русских сказуемых предложений выдаются в форме, соответствующей текстовой форме (времени, числе и роде) русских сказуемых, а остальные слова и фразы (длиной до шести слов) выдаются в основной форме единственного числа именительного падежа (см. рис. 2).

Заключение

В статье описывается методика разработки компьютерных систем реферирования и аннотирования, основанная на ориентации их на конкретные области науки и техники, а также на заранее стандартизированный текст. Показана возможность повторного использования программных ресурсов для новых языков и приложений и, таким образом, снижения всех видов затрат при экстраполяции системы на новые научно-технические области.

Представлена система РЕФЕРАТ, разработанная по предложенной методике для предметной области математического моделирования в НОЦ ЛиНТ ЮУрГУ.

Литература

1. Каптерев, А.И. Информатизация социокультурного пространства / А.И. Каптерев. – М.: ФАИР-ПРЕСС, 2004. – 512 с.

2. 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.

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

4. Loret, E. A Gradual Combination of Features for Building Automatic Summarisation Systems Text / E. Lloret, M. Palomar // Speech and Dialogue. – Heidelberg, 2009. – P. 16–23.

5. 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.

6. Saggion, H. A classification algorithm for predicting the structure of summaries / H. Saggion // Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLPЬ 2009. – Suntec, 2009. – P. 31–38.

7. ГОСТ 7.9–95. Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования. – Введ. 1997–07–01. – М.: Изд- во стандартов, 1995. – 8 с.

8. Sheremetyeva, S. On Extracting Multiword NP Terminology for MT / S. Sheremetyeva // Proceedings of the Thirteen Conference of European Association of Machine Translation (EAMT-2009). – Barcelona, Spain. May 14–15, 2009.

Поступила в редакцию 26 ноября 2012 г.