УДК811.93
DOI 10.18101/978-5-9793-1497-6-97-101

О НЕКОТОРЫХ ПОДХОДАХ В АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТОВ НА ЕСТЕСТВЕННЫХ ЯЗЫКАХ

© Токтохоев Роман Николаевич
студент,
Бурятский государственный университет имени Доржи Банзарова
Россия, 670000, г. Улан-Удэ, ул. Смолина, 24а
E-mail: toktokhoev@inbox.ru

© Токтохоева Татьяна Александровна
старший преподаватель,
Бурятский государственный университет имени Доржи Банзарова
Россия, 670000, г. Улан-Удэ, ул. Смолина, 24а
E-mail: totaal@mail.ru

В статье рассматриваются некоторые вопросы автоматизированной обработки текстовых массивов. Описываются современные подходы к проблеме автоматического реферирования. Представлен краткий обзор наиболее актуальных и перспективных направлений развития исследований в области автоматического реферирования.
Ключевые слова: автоматическая обработка текстовых массивов, автоматическое реферирование, автоматическое аннотирование.

В настоящее время мы столкнулись с тем, что с каждым днем объем информации увеличивается, причем, это происходит в геометрической прогрессии. Инфокоммуникационные технологии внедряются во все сферы человеческой деятельности и отрасли производства: промышленность, транспорт, социальную сферу, науку, образование, медицину, бизнес, финансы, энергетику и т.д. В современном обществе человек сталкивается с тем, что ему необходимо воспринимать, хранить и использовать в своей трудовой деятельности огромный поток информации. Отсюда вытекает необходимость в обработке больших объемов данных: структурированных и неструктурированных (в том числе на естественных языках). Поэтому актуальной задачей для ИТ-специалистов является автоматизация процессов обработки текстовой информации, таких как индексирование, аннотирование, реферирование и др. В данной статье рассмотрим именно автоматическое реферирование и аннотирование, которое позволяет осуществить информационную поддержку лиц, принимающих управленческие решения.

Под аннотацией мы понимаем короткий связный текст, по объему содержащий не более 150-200 знаков, описывающий основную тему или предмет рассматриваемого документа. Рефератом мы называем текст, который в сжатом виде представляет первоисточник и передает его смысловое содержание. Обычно он бывает объемом 1000-1500 знаков и содержит в отличие от аннотации еще и цель, основные методы и результаты описываемого оригинала. Качественный реферат или аннотация позволяют человеку понять основное содержание текста и принять решение о необходимости обращения к первоисточнику. Это

значительно ускорит поиск необходимой информации, так как позволит уже при первом знакомстве отклонить не отвечающий требованиям материал и сэкономить время. Не вызывает сомнения то, что использование вычислительной техники для реферирования и аннотирования позволит более эффективно обрабатывать большие объемы информации.

Попытки создания качественных систем автоматического реферирования и аннотирования ведутся уже давно, начиная с 60-70-х годов прошлого столетия. За это время было предложено множество алгоритмов и методов решения данной задачи. Современные исследования направлены на то, чтобы улучшить качество аннотации или реферата, полученных в результате обработки больших текстовых массивов на естественном языке.

Сложилось два направления в данной области: квазиреферирование и автоматическое реферирование, как краткое изложение содержания. «Квазиреферирование основано на экстракции из первичных документов с помощью определённых формальных признаков «наиболее информативных» фраз (фрагментов), совокупность которых образует некоторый экстракт (квазиреферат). Собственно, автоматическое реферирование же основано на выделении из текстов с помощью специальных информационных языков наиболее существенной информации и порождении новых текстов (рефератов), в большей или меньшей степени изоморфных первичным документам (или их частям)» [2].

Первые работы были связаны в основном с выявлением статистических закономерностей распределения в тексте терминов и их взаимного расположения. Это продемонстрировано на рисунке 1. Созданный в итоге реферат представлял собой совокупность отдельных предложений или фрагментов текста, вырезанных из документа и собранных в порядке их следования. Релевантность выбора того или иного фрагмента текста определялась с учетом частотности слов. Затем разработки в данной области пошли по пути исследования самой структуры текстов, учета синтаксического и морфологического анализаторов, установления семантических связей, определения весов, а также машинного обучения. В данное время алгоритмы обработки текстов широко используют методы искусственного интеллекта.

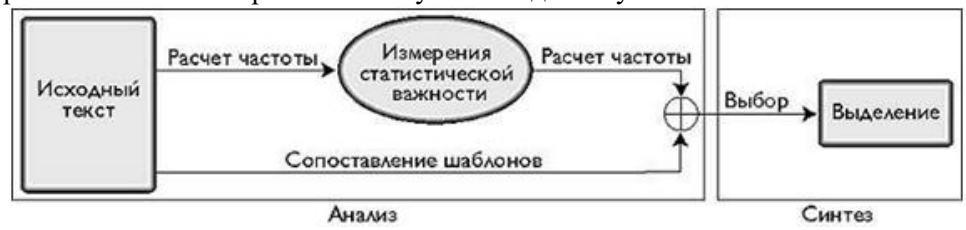

Рис1. Архитектура извлечения фрагментов текста

Используемые методы автоматического реферирования были ориентированы на извлечение предложений (Sentence Extraction) из текста оригинала на основе модели линейных весовых коэффициентов. Наиболее значимые предложения выбирались по позиционным или тематическим критериям. Что предполагает собой позиционный критерий? Здесь учитывается местоположение предложения в документе, определяется тип выделенного элемента, например, заголовок, вступление, заключение. Тематический критерий – это прежде всего наличие ключевых слов в предложении. Дальнейшее развитие методов экстракции

привело к использованию определяющих значимость предложений маркеров. К ним, например, можно отнести фразы типа «таким образом», «в итоге», «наше исследование направлено», «важно», «по результатам анализа» и другие. Кроме того, методы извлечения стали ориентироваться на связность элементов текста. Здесь предполагается учет близости расположения в тексте, частоты повторов и синонимичности и т. п. Методы данного подхода характеризуются наличием оценочной функции (Scoring Function) важности информационного блока или предложения. [2]

«Общий вес всего блока текста U является суммой индивидуальных весов, определяемых в соответствии со специальными параметрами настройки, связанными с каждым весом.

$$Weight(U) = Location(U) + CuePhrase(U) + StatTerm(U) + AddTerm(U).$$

Location(U) – это весовой коэффициент расположения, в данной модели зависящий от того, где встречается данный фрагмент — в начале, в середине или в конце, и используется ли он во введении или в заключении. CuePhrase(U) – весовой коэффициент ключевой фразы, зависящий от того является ли она частью лексической или фразовой резюмирующей конструкции, в сочетании с такими как «в заключение», «в данной статье», «согласно результатам исследования» и так далее или в сочетании с оценочными терминами, принятыми в конкретной предметной области, например, «высочайший», «малоэффективный», «незначительный». StatTerm(U) - весовой коэффициент статистической важности, вычисляющийся на основании данных, полученных в результате анализа автоматической индексации, с использованием целого ряда метрик, определяющих весовые коэффициенты термина. Эти метрики позволяют выделить документ из числа других в определённом наборе документов. Одна группа метрик, например, метрика tf.idf, характеризует баланс между частотой появления термина в документе и частотой его появления в наборе документов. AddTerm(U) - весовой коэффициент дополнительного наличия терминов, появляющихся также в заголовке, в колонтитуле, первом параграфе и в тексте пользовательского запроса. Выделение приоритетных терминов, наиболее точно отражающих интересы пользователя, — это один из путей настроить реферат или аннотацию на конкретного человека или группу. К основному недостатку систем этого класса стоит отнести отсутствие связанности текста получаемого реферата: как правило, выбранные наиболее значимые информационные блоки никак не связаны между собой». [2].

Другой подход, как мы уже говорили выше, предполагает создание автоматического реферата, который является кратким изложением содержания источника. Иными словами, происходит генерация реферата с порождением нового текста, передающего суть оригинала (извлечение содержания, Content Extraction). Данный подход предполагает обязательно три этапа: «анализ исходного текста с генерацией внутреннего представления, семантическое сжатие внутреннего представления и синтез нового текста» [2]. В данном подходе получили развитие два основных направления: абстракция на основе лингвистического сжатия и абстракция с опорой на знания.

При составлении реферата на основе лингвистического сжатия исходный текст анализируется, а затем формируется синтаксическое дерево разбора.

Сжатие происходит путем сокращения ветвей дерева. Оно основано на анализе структуры и исключении незначимых частей. К таким, например, можно отнести подчиненные предложения, скобки и т.д. Данный метод является очень требовательным к ресурсам.

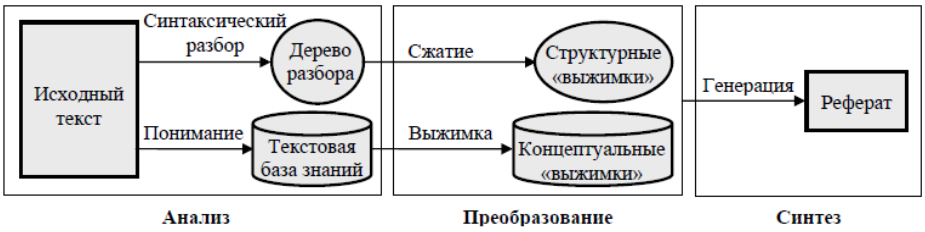

Рис 2. Архитектура формирования краткого изложения

Реферирование по методу абстракции с опорой на знания основывается на предположении о том, что гораздо проще сократить текст, если понимать его значение и вероятнее получить более качественный результат. Данный подход предусматривает «использование базы знаний значительного объема, состоящей из правил, которые извлекаются, поддерживаются и затем адаптируются к новым приложениям и языкам» [4]. Основой здесь являются системы искусственного интеллекта.

В настоящее время подготовка качественного реферата является одной из важных задач компьютерной обработки текстов. Исследования в данной области очень актуальны и направлены на совершенствование методик и подходов.

Литература

1. Леонов В. П. О методах автоматического реферирования / В. П. Леонов // НТИ. Сер. 2. – 1975. – № 6. – С. 16–20.
2. Луканин А. В. Автоматическая обработка естественного языка/ А. В. Луканин; М-во образования и науки Российской Федерации, Южно-Уральский гос. ун-т. – Челябинск: Изд. центр ЮУрГУ, 2011. – 70 с.
3. Тарасов С. Д. Современные методы автоматического реферирования / С. Д. Тарасов // Научно-технические ведомости СПбГПУ 2010. Информатика. Телекоммуникации. Управление, 2010.– С. 68–74.
4. Хан У. Системы автоматического реферирования / У. Хан, И. Мани // Открытые системы. – 2000. – № 12. – [Электронный ресурс]. – URL: http://www.osp.ru/os/2000/12/178370.

ON SOME APPROACHES IN AUTOMATIC PROCESSING OF TEXTS IN NATURAL LANGUAGES

Roman N.Toktokhoev
Student,
Dorzhi Banzarov Buryat State University
24a Smolina St., Ulan-Ude 670000, Russia
E-mail: toktokhoev@inbox.ru

Tatiana A. Toktokhoeva
Senior Lecturer,
Dorzhi Banzarov Buryat State University
24a Smolina St., Ulan-Ude 670000, Russia
E-mail: totaal@mail.ru

The article deals with some issues of automatic processing of texts. Modern approaches to the problem of automatic summarization are described. A brief overview of the most relevant and promising areas of research development in the field of automatic summarization is presented.
Keywords: natural language processing (NLP), automatic summarization.