АЛГОРИТМ АВТОМАТИЗИРОВАННОГО РЕФЕРИРОВАНИЯ НОВОСТНЫХ АНГЛОЯЗЫЧНЫХ ТЕКСТОВ

Дашкевич Е.С.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 0978314534,
e-mail: esdashkevich@gmail.com

Целью данного исследования является разработка математической модели автоматизированного реферирования текстов на английском языке. В процессе выполнения работы были поставлена задача разработать математическую модель построения автоматического реферата и на основе разработанной модели построить алгоритм программной реализации автоматического реферирования.

Процесс разработки методов автоматического формирования краткого представления или реферата (англ. summary) текстовых документов длится с конца 1950-х годов. Автоматическое реферирование – это создание коротких выкладок материалов, аннотаций или дайджестов, т.е. получение важнейших сведений из одного или нескольких документов, и генерация на их основе лаконичных и информационно-насыщенных отчетов [1]. Существуют два направления автоматического реферирования - квазиреферирование и краткое изложение содержания. Современные методы реферирования дают возможность автоматизировать данный процесс таким образом, что вмешательство человека и конечная корректировка продукта будут минимизированы. На сегодняшний день пользователю доступны системы, осуществляющие автоматическое реферирование разными методами и на разных языках, и работающие с текстами разных тематик.

Информация, которую человек получает ежедневно – это новости. Наибольшим спросом, по мнению новостного агентства БиБиСи, пользуются сжатые новостные сообщения аналитического характера. Основным методом, который используется для реферирования новостных сообщений, является статистический. Это обусловлено тем, что новостные сообщения представляют собой четко структурированные тексты, не требующие специальной семантической обработки.

Учитывая, что язык – это конечный, дискретный, детерминированный объект, для своего моделирования он требует методов и средств дискретной математики. Дискретная математика позволяет моделировать большинство языковых явлений и анализировать языковые процессы.

Для расчёта веса термина используется формула TF. Как правило, наибольший вес в документе имеют общеупотребительные слова и термины, которые не дают представления о содержании текста. Термины со слишком малым весом также могут быть ключевыми за редким исключением, поэтому должны быть исключены из текста [2]. С целью получения ключевых слов по данному принципу, будет использовано равенство веса ключевых терминов

$TF_{min} < TF_k < TF_{max}$, где TF_{min} – нижняя граница веса терминов в документе, TF_{max} – верхняя граница веса терминов в документе, TF_k – диапазон веса ключевых слов. Таким образом, задача автоматического реферирования состоит в том, чтобы создать реферат, максимально приближенный по качеству к получаемому в результате человеческой когнитивной деятельности.

На основе этого утверждения был создан алгоритм реферирования англоязычных новостных сообщений, который опирается на метод веса ключевых слов. Суть алгоритма заключается в том, чтобы выделить из текста лексемы со средним весом и считать их ключевыми. Для удобства работы с текстом, он будет приведен к массиву строк. На стадии предварительной обработки будет определен объем текста. Так мы получим обработанный текст, благоприятный для дальнейшего анализа.

В полученном «скелете» текста будет проведен анализ количества вхождений конкретного термина в текст. Для этого все лексемы, начиная с первого слова в заголовке, будут сравниваться со следующими лексемами в массиве предложений. Найденная лексема сразу удаляется из исходного предложения и заносится в структуру, где ей присваивается значение инкременты 1. Так, количество вхождений термина в текст инкрементируется начиная со значения 1++. Посредством удаления уже обработанных лексем осуществляется оптимизация алгоритма. Вследствие этого этапа получим необходимые параметры для определения веса терминов. После вычета веса всех терминов, анализатор определит нижнюю и верхнюю границу веса ключевых терминов. Все лишние лексемы будут удалены из структуры. Последним этапом автоматизированного создания реферата является запись заголовка текста и всех предложений, содержащих ключевые слова, в текстовый файл. Таким образом, получим реферат новостного сообщения.

Данный алгоритм также может работать с корпусом текстов. Основой создания реферата корпуса текстов является нахождение ключевых слов, общих для всех текстов данной коллекции. Благодаря выделению цельных ключевых предложений в реферат, минимизируется возможность нарушения синтаксических связей и дробления смысла и идеи текста. Цель заключается в том, чтобы сохранить оригинальную идею текста и использованные в нем семантические единицы.

Таким образом, в результате работы представлена математическая модель автоматизированного реферирования новостных текстов. Был предложен алгоритм для программной реализации системы автоматизированного реферирования, основанный на методе веса ключевых слов. Следующим шагом в решении задачи автоматизированного реферирования является создание соответствующего программного обеспечения.

Список литературы

1. *Nenkova A., McKeown K.* Automatic Summarization./ A. Nenkova, K. McKeown. – NY. : Springer US, – 2011. – pp. 216.

2. *Sparck Jones K.* A statistical interpretation of term specificity and its application in retrieval./ K. Sparck Jones. – L.: Journal of Documentation, –1972. – pp. 12.