МЕТОДЫ АВТОМАТИЧЕСКОЙ ФУНКЦИОНАЛЬНОЙ КОМПРЕССИИ ТЕКСТА

Керножицкая В.И.

Минский государственный лингвистический университет

Аннотация: В статье рассматривается ряд методов автоматического функционального реферирования текста первичного документа, позволяющих создавать вторичные документы путем экстрагирования важных смысловых единиц исходного текста. Основными характеристиками данных методов является учет статистических параметров текстовых единиц, число связей между предложениями текста, объективность и полная автоматизация.

Ключевые слова: вторичный документ, первичный документ, смысловое сжатие, текст, функциональная компрессия

Одной из самых больших проблем современного общества является его информационное переполнение. Постоянно растущий объем информации не позволяет находить необходимые материалы в массивах первичных документов, поскольку такой поиск требует слишком много времени. Поэтому он осуществляется в массивах вторичных документов: библиографических описаниях, аннотациях и рефератах. Вторичные документы представляют собой результат свертывания, т.е. смыслового сжатия или компрессии текста первичного документа. Один из подходов к смысловой компрессии текста ориентирован на автоматическое извлечение фрагментов исходного текста, из которых составляется текст вторичного документа. При этом наиболее информативные единицы извлекаются из текста первоисточника на основе лексико-грамматических признаков, зафиксированных в словарях. Таким образом, данный подход предполагает осуществление трех основных этапов: составление словаря; просмотр текста первичного документа и выявление предложений, содержащих лексико-грамматические единицы, указанные в

словаре; экстрагирование из исходного текста выявленных предложений и составление текста реферата как вида вторичного документа [4, с. 5–6].

В зависимости от структуры словаря можно выделить два метода смыслового сжатия текста: на основе словарей нетематической лексики и на основе словарей тематической лексики. Примером автоматической смысловой компрессии исходного текста на основе словарей тематической лексики может служить функциональное сжатие текста, основные идеи которого были разработаны Х.П. Лунном и Э.Ф. Скороходько [4, с. 8–9]. Это направление основано на предположении, что наиболее значимая информация содержится в тех предложениях исходного текста, которые имеют наибольший функциональный вес. В свою очередь, функциональный вес предложения рассматривается как количество смысловых связей между данным предложением и остальными предложениями текста первичного документа. Каждая связь реализуется посредством повторения в предложениях одного и того же имени существительного, глагола, причастия, прилагательного, или наречия, принадлежащего к данной предметной области и зафиксированного в словаре тематической лексики. На основе такого словаря и подсчитывается функциональный вес предложений в исходном тексте. Предложения с большим функциональным весом включаются в текст реферата. Функциональная компрессия может применяться только по отношению к текстам определенной, достаточно узкой предметной области. Рассмотрим подробнее некоторые методы автоматической функциональной компрессии исходного текста.

Наиболее распространенный статистико-дистрибутивный метод представляет собой единство дистрибуции, как совокупности всех окружений рассматриваемой языковой единицы, и статистики, как количественных данных об этих окружениях [2, с. 423–429]. В общем статистико-дистрибутивном методе выделяется несколько частных типов. К их числу относится, в первую очередь, метод Х.П. Луна, согласно которому наиболее информативными предложениями текста исходного документа считаются те,

в которых содержатся скопления значимых (ключевых) для данного текста слов, расположенных достаточно близко друг к другу. При этом позиционные ограничения на предложения не накладываются: они могут выбираться из любого фрагмента текста первичного документа. Скоплением считается любая цепочка слов предложения, крайние элементы которой являются значимыми для данного текста, причем между ближайшими значимыми словами цепочки находится не более пяти не ключевых слов. Хотя метод Х.П. Луна не учитывает смысловых связей между словами, его преимущество заключается в возможности осуществления полной автоматизации. Статистико-дистрибутивный метод позволяет простейшим образом выделить из исходного текста те предложения, которые можно считать наиболее информативными и которые в силу этого можно включить в текст вторичного документа. Однако его недостаток заключается в том, что между информативностью фрагмента и частотой входящих в него ключевых слов нет прямой и однозначной зависимости. Кроме того, текст реферата не всегда отличается особой связностью. Поэтому статистический критерий должен, по необходимости, дополняться другими критериями. Существует ряд модификаций метода Х.П. Луна [3, с. 11–13]. Например, с целью его оптимизации В.А. Освальд предложил оценивать информативность предложений не только наличием скоплений значимых слов, но и количеством таких скоплений. По мнению Л. Эрла, информативными следует считать предложения с числом скоплений ключевых слов не менее трех. Л. Дойл и М. Квиллиан предложили метод учета совместной встречаемости в предложениях терминов или ключевых слов и вычисления их коэффициента подобия с помощью особой матрицы взаимосвязанных элементов.

Еще одной разновидностью функциональной компрессии исходного текста является метод симметричного сжатия, основанный на принципах симметричности, отождествления, последовательности и контактной связи [4, с. 10–15]. В процессе смысловой компрессии из текста первичного документа выбираются повторяющиеся слова, которые не относятся к предметной

области текста и, следовательно, не принимаются во внимание в процессе построения текста реферата, а также повторяющиеся слова, принадлежащие предметной области, которые обязательно необходимо учитывать. Важно отметить, что при формировании текста вторичного документа учитываются как правосторонние, так и левосторонние связи предложений. Основным принципом данного метода является принцип симметричности отношения, суть которого заключается в следующем: если предложение *X* имеет *n* связей с предложением *Y*, то предложение *Y* имеет *n* связей с предложением *X* (в обоих предложениях есть повторяющиеся, важные в смысловом плане лексические единицы). Другой принцип заключается в том, что повторяющиеся слова, имеющие одну основу, но разные словообразовательные и формообразующие суффиксы, отождествляются и рассматриваются как одно слово. Метод симметричного сжатия исходного текста имеет следующие преимущества: его достаточно легко автоматизировать при наличии словаря терминов, принадлежащих к данной области знания; можно достаточно просто изменить размер текста вторичного документа, установив определенный пороговый уровень функционального веса отбираемых в него предложений; симметричное сжатие может применяться как к большим, так и к небольшим научным и газетным текстам; метод симметричной компрессии текста позволяет изменять параметры информационного поиска [4, с. 20–21; 3, с. 14].

Как отмечают ученые, описанные выше и целый ряд других формальных методов предполагают только преобразование одной внешней формы в другую, минуя внутренний этап свертывания содержания текста, который всегда реализуется человеком. Именно на этом этапе совершается смысловое преобразование, которое является необходимым условием семантической адекватности первичных и вторичных текстов [1, с. 119].

ЛИТЕРАТУРА

1. Карпилович, Т. П. Когнитивно-коммуникативная модель смысловой компрессии научного текста: дис. … док. филол. наук: 10.02.19, 10.02.21 / Т. П. Карпилович. – Минск : МГЛУ, 2005. – 223 с.

2. Прикладное языкознание: учебник / Л. В. Бондарко [и др.]; под общ. ред. А. С. Герда. – СПб., 1996. – 528 с.

3. Хан, У. Системы автоматического реферирования / У. Хан // Открытые системы. – М., 2000. – № 12. – С. 10–15.

4. Яцко, В. А. Симметричное реферирование: теоретические основы и методика / В. А. Яцко // НТИ. Сер.2. Информационные системы. – М., 2002. – № 2. – С. 5–25.