УДК 81’322

СОВРЕМЕННЫЕ ПОДХОДЫ К АВТОМАТИЧЕСКОМУ РЕФЕРИРОВАНИЮ И АННОТИРОВАНИЮ

П.Г. Осминин

MODERN APPROACHES TO AUTOMATIC SUMMARIZATION

P.G. Osminin

Представлен краткий обзор современных научных исследований по автоматическому реферированию и аннотированию.

Ключевые слова: автоматическая обработка текста, автоматическое аннотирование, автоматическое реферирование.

The article presents a short review of current research on automatic summarization.

Keywords: natural language processing (NLP), automatic summarization.

В современном мире возрастает актуальность применения методов автоматического реферирования и аннотирования. В настоящее время существует проблема информационной перегрузки. Рефераты и аннотации дают возможность установить основное содержание документа и определить необходимость обращения к первоисточнику. Автоматическое реферирование и аннотирование помогает человеку эффективно обрабатывать большие объемы информации.

По способу построения текста методы автоматического реферирования и аннотирования делятся на две группы: извлекающие и генерирующие[1].

При использовании извлекающих методов из исходного текста выделяются наиболее важные фрагменты (предложения, абзацы). При этом данные фрагменты не обрабатывают, а извлекают в таком порядке и виде в каком они приведены в тексте.

Среди извлекающих методов мы рассмотрим следующие: методы на основе машинного обучения и методы на основе теории графов.

Преимуществом методов на основе машинного обучения является удобство тестирования целого ряда признаков важности.

В работе К.Ф. Вонга[2] рассматривается сочетание различных признаков важности предложения: поверхностные признаки (расположение предложения), содержательные (частота слов). Разработаны два алгоритма: алгоритм обучения с учителем и алгоритм частичного обучения. Оценка результатов показала, что лучшим является сочетание поверхностных и содержательных признаков.

В методах на основе теории графов текст представляется в виде графа, узлы которого представляют фрагменты текста (слова, предложения, абзацы), а ребра обозначают отношения между узлами, например семантические отношения.

В работе Л. Плаза[3] представлен метод реферирования, основанный на представлении текста в концепты с последующим преобразованием документа и предложений в граф. Метод использует дополнительные ресурсы – тезаурус медико-биологической области UMLS и программу MetaMap для преобразования текста в концепты из тезауруса UMLS. Метод состоит из следующих шагов: представление документа в виде графа, кластеризация концептов, выбор предложений.

Генерирующие методы реферирования и аннотирования основаны на лингвистических правилах обработки естественного языка или методах искусственного интеллекта. Генерирующие методы способны создавать новый текст, не представленный явно в тексте исходного документа.

Авторы работы[4] описывают создание аннотаций для числовых данных. Определяются изменения во входных данных (данные сенсоров газовой турбины), происходит их представление в символьном виде, определяются необходимые изменения и происходит генерация текста, описывающего эти изменения. Системе необходимы компонент анализа данных и модуль генерирования текста. Для выполнения этих задач авторы провели процедуру сбора знаний: опрос экспертов по опи-

Осминин Павел Григорьевич, аспирант кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (г. Челябинск); научный руководитель – С.О. Шереметьева, доктор филологических наук, профессор. E-mail: osperevod@gmail.com

Pavel G. Osminin, postgraduate student, linguistics and cross-cultural communication department, South Ural State University (Chelyabinsk); Scientific supervisor – S.O. Sheremetyeva, Doctor of philological Sciences, Professor. E-mail: osperevod@gmail.com

санию числовых данных, разработку онтологии примеров описания данных.

В работе[5] исследуется проблема автоматической генерации структуры аннотаций. Авторы отмечают, что предикаты и предикатные фразы имеют коммуникативную функцию – предупреждение читателя о содержании аннотированного документа путем явного указания («упоминает», «представляет»). Разработанный алгоритм получает на входе набор извлеченных фрагментов предложений и определяет, как соединить фрагменты в аннотацию. На каждом шаге алгоритм выбирает для вставки в начало текущего фрагмента наиболее подходящий предикат (фразу) из заранее определенного словаря. Оценка результатов показала, что разработанный алгоритм может прогнозировать структуру аннотаций более чем в 60 % случаев.

Таким образом, мы можем сделать вывод, что современные подходы к автоматическому реферированию и аннотированию отличаются разнообразием используемых методов. Материалами для реферата и аннотации могут выступать не только тексты, но и числовые данные[6].

[1] Кондратьев М. Аннотирование по запросу: связность или информативность // Труды третьего российского семинара по оценке методов информационного поиска / под ред. И.С. Некрестьянова. СПб.: НИИ Химии СПбГУ, 2005. С. 125–135.

[2] Wong Kam-Fai. Extractive summarization using supervised and semi-supervised learning // Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, 2008. P. 985–992.

[3] Plaza L. Concept-graph based Biomedical Automatic Summarization using Ontologies // Coling 2008: Proceedings of 3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing. Manchester, 2008. P. 53–56

[4] Choosing the content of textual summaries of large time-series data sets / J. Yu, E. Reiter, J. Hunter, Chris Mellish // Natural Language Engineering. 2007. Vol. 13, № 1. P. 25–49.

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

[6] Choosing the content of textual summaries of large time-series data sets…

Поступила в редакцию 10 июня 2012 г.