Фамилия	Имя	Отчество	Факультет	Курс
Буканова	Ольга	Владимировна	НИУ ВШЭ НН, Факультет гуманитарных наук	4
Телефон			**email**	
89047818792			ovbukanova@mail.ru	

Научный руководитель	Тема доклада
Малафеев Алексей Юрьевич, к. филол. н., доцент департамента прикладной лингвистики и иностранных языков, факультет гуманитарных наук, НИУ ВШЭ	«Автоматическое реферирование нескольких статей одной тематики на русском языке»

Автоматическое реферирование нескольких статей одной тематики на русском языке

С развитием информационных технологий увеличиваются необходимые на поиск интересующих сведений временные затраты, что снижает личную эффективность пользователя. В таких условиях область автоматического реферирования становится всё более актуальной, поскольку эта сфера обработки естественного языка позволяет представить необходимую информацию, содержащуюся в источниках, в сокращенном виде, что значительно экономит временные ресурсы.

Когда приходится иметь дело с несколькими документами примерно одинакового содержания (например, новостными статьями, размещенных на разных Интернет-ресурсах), проблема информационной избыточности встает наиболее остро, так как поиск релевантной информации затрудняется огромными объемами схожей информации, повторяющейся в каждом отдельно взятом документе.

Таким образом, программа, способная перерабатывать массивы текстов со схожей информацией и тем самым экономить временные ресурсы пользователя имеет высокую практическую значимость.

В данном исследовании была создана программа, позволяющая

генерировать реферат из массивов текстов со схожей информацией. Итогом автоматического реферирования этой программы является документ, который содержит в себе релевантную информацию, неоднократно появляющуюся в коллекции, без повторов, который при этом также включает в себя дополнительную информацию, специфичную для какого-то из текстов в массиве. Здесь избыточность помогает точнее идентифицировать важную информацию, которая впоследствии должна присутствовать в готовом реферате.

Для анализа работы программы были отобраны коллекции статей с новостных порталов «РИА Новости», «Интерфакс», «Корреспондент» и др.

Алгоритм программы берет за основу коэффициенты подобия между парами предложений и коэффициенты информативности каждого предложения в исходных документах, которые высчитываются определенным образом с помощью статистических и лингвистических критериев:

- частота словоупотреблений в массиве текстов;
- частота словоупотреблений в заранее отобранной коллекции текстов заданной тематики;
- частеречная принадлежность слов;
- имена собственные;
- информативность предложений;
- метаданные (время публикации статьи и др.)

Следующим этапом программы является отбор предложений из повторяющихся блоков текста. Здесь делается акцент на полноту информации, а также на её актуальность.

Полученные результаты и алгоритм программы будут рассмотрены в ходе доклада.

Библиографический список

1. Barzilay, R.: Information Fusion for Multidocument Summarization: Paraphrasing and Generation// Doctoral Dissertation. - Columbia University New

York, NY, USA. – 203p. – 2003.

2. Barzilay, R., McKeown, K.: Sentence Fusion for Multidocument News Summarization. – Computational Linguistics, vol. 31 (3). – pp. 297-328. – 2005.

3. Barzilay, R., Elhadad, N., McKeown, K.R.: Inferring Strategies for Sentence Ordering in Multidocument News Summarization, Journal of Artificial Intelligence Research, vol. 17 – pp. 35–55 – 2002.

4. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitzt, M.: Multi-Document Summarization By Sentence Extraction. – In NAACL-ANLP-AutoSum '00 Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization – Vol. 4– pp. 40-48 – 2000.

5. Hägerstrand, A.: Multi Document Summarization. A search based approach// Master's Thesis. – KTH, School of Computer Science and Communication (CSC) – 39 p. – 2011.

6. McKeown, K.R., Klavans, J., Hatzivassiloglou, V., Barzilay, R., and Eskin, E.: Towards multidocument summarization by reformulation: Progress and prospects. In Proceedings of the National Conference on Artificial Intelligence. – pp. 453-460. – 1999.