Интеграция тематического представления в классические методы автоматического аннотирования новостных кластеров

Алексеев Алексей Александрович

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: a.a.alekseevv@gmail.com

Рассматривается задача автоматического аннотирования новостных кластеров – задача подготовки краткого изложения наиболее значимой информации, содержащейся в наборе новостных сообщений, посвященных одной теме. Большинство существующих методов для решения данной задачи основаны на пословной модели представления текстов (bag–of–words model) — рассмотрение текста как массива отдельных независимых слов. Классическими и широко используемыми методами, основанными на данной модели, являются методы Maximal Marginal Relevance (MMR, [1]) и SumBasic [2]. Необходимо отметить, что пословные методы автоматического аннотирования не учитывают структурные особенности реальных текстов.

В то же время тексты на естественном языке подчиняются определенным законам построения, в частности обладают древовидной тематической структурой [3]. В работе [4] предложен метод построения тематического представления новостных кластеров, комбинирующий набор разнородных характеристик для определения схожести языковых выражений и базирующийся на особенностях внутреннего устройства реальных текстов. Тематическое представление в работе [4] является разбиением всех языковых выражений (слов и многословных выражений) на тематические цепочки – группы выражений, относящихся к одному участнику ситуации, описываемой в исходном новостном кластере.

В докладе рассматривается алгоритм интеграции тематического представления в классические методы аннотирования, основанные на пословной модели, на примере методов MMR и SumBasic. Предлагаемый алгоритм строился в предположении, что добавление информации о внутреннем устройстве исходного новостного кластера позволит улучшить общее качество алгоритмов автоматического аннотирования.

В рамках проведенного исследования разработан программный комплекс, реализующий методы аннотирования MMR и SumBasic с интеграцией и без интеграции тематического представления, а так-

же автоматическую оценку аннотаций пакетом ROUGE [5]. Метод ROUGE является основным методом оценки качества автоматических аннотаций, суть которого заключается в сравнении порожденных аннотаций с эталонными, составленными экспертами. Получены результаты, подтверждающие улучшение качества методов MMR и SumBasic с интегрированным тематическим представлением на 6% (усредненная оценка по различным метрикам ROUGE).

Литература

1. Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries // Proceedings of ACM conference, Melbourne, Australia, 1998, P. 335–336.

2. Vanderwende L., Suzuki H., Brockett C., Nenkova A. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion // In Information Processing and Management Journal, Volume 43, Issue 6, November, 2007, P. 1606–1618.

3. Лукашевич Н. В., Добров Б. В. Исследования тематической структуры текста на основе большого лингвистического ресурса // Труды международного семинара Диалог, Москва, Россия, 2000, стр. 252–258.

4. Alekseev A., Loukachevitch N. Use of Multiple Features for Extracting Topics from News Clusters // Proceedings of SYRCODIS conference, Moscow, Russia, 2012, P. 3–11.

5. Lin C. Y. ROUGE: a Package for Automatic Evaluation of Summaries // Proceedings of the ACL conference, Barcelona, Spain, 2004, P. 74–81.