УДК 004.9

СИСТЕМА АННОТИРОВАНИЯ НА ОСНОВЕ ГИБРИДНОЙ ЭКСТРАКТИВНО-АБСТРАКТИВНОЙ АРХИТЕКТУРНОЙ МОДЕЛИ

© Батаев Д.В., Головнин О.К.

Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Российская Федерация

e-mail: denisbataev1@gmail.com

Растущее количество порождаемой текстовой информации и бурное развитие технологий ее обработки открывают новые возможности в области анализа естественного языка [1; 2]. Системы извлечения и суммаризации текстовой информации способны аннотировать большие объемы материала, тем самым выступая полезным инструментом, позволяющим существенно экономить время при проведении такого рода работ [3].

Ведется разработка системы аннотирования на основе гибридной экстрактивно-абстрактивной архитектуры [4], позволяющей объединить достоинства подходов для повышения качества аннотирования (на рис.).

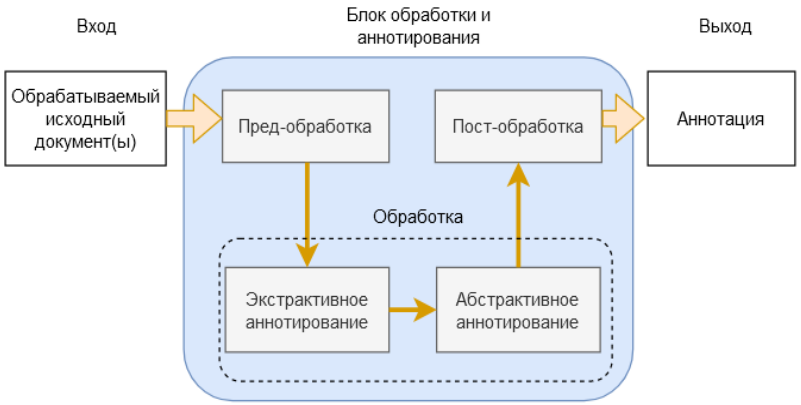

Рис. Архитектура гибридной системы аннотирования [4]

В разрабатываемой системе на этапе предобработки производятся извлечение текста из источника, преобразование его в формат JSON, токенизация, стемминг и удаление стоп-слов.

На этапе экстрактивного аннотирования в системе выполняется оценка предложений путем ранжирования, затем извлекаются предложения с высокими оценками (наиболее важные), осуществляются переупорядочивание предложений и замена в них имен существительных и относительных высказываний на абсолютные.

На этапе абстрактивного аннотирования в системе выполняется построение внутреннего семантического представления и сводное генерирование аннотации посредством модели глубокого обучения Transformer [5], при этом используется параллельная обработка.

На этапе постобработки в системе осуществляется контроль валидности полученного предложения для результирующей аннотации, например с целью контроля его длины и состава.

Система получает на вход исходные текстовые файлы в формате PDF; результирующий файл формируется в формате DOCX для возможности дальнейшего редактирования пользователем. Предложенная система реализует аналитические функции, предоставляя на выходе не только аннотацию, но и список ключевых слов с

учетом частоты встречаемости. В системе для генерации аннотации устанавливаются следующие ограничения: размер аннотации может составлять от 100 до 1000 слов; количество ключевых слов должно быть не более 10.

Разработка системы ведется на языке Python в среде PyCharm. В процессе разработки используются дополнительные программные библиотеки Numpy, Scipy, Matplotlib, Pandas, python-doc, PyPDF2, а также библиотеки машинного обучения TensorFlow и Keras.

Преимущество реализуемого гибридного аннотирования заключается в дополнении экстрактивного подхода абстрактивным, что повышает точность результата. Таким образом, ожидается, что разрабатываемая система позволит получить объективную аннотацию текста, не привязанную к мнению конкретного лица, и сократить время на изучение и анализ исходной текстовой информации.

Библиографический список

1. Review of automatic text summarization techniques & methods / Adhika P.W., Supriadi R., Guruh F.S. [et al.] // Journal of King Saud University. Computer and Information Sciences. 2020.

2. Столбова А.А., Головнин О.К. Теоретические основы и практические аспекты информатики и программирования для решения задач управления и обработки информации на языке C#. Самара: Самарский университет, 2019.

3. Батаев Д.В. Автоматизированная система извлечения и суммаризации текстовой информации из открытых источников // Новые информационные технологии в научных исследованиях: материалы XXV Юбилейной Всерос. научн.-технич. конф. студентов, молодых ученых и специалистов. 2020. С. 196.

4. Wafaa S.E.-K., Cherif R.S., Ahmed A.R., Hoda K.M. Automatic text summarization: A comprehensive survey // Expert Systems with Applications. 2020. P. 113679.

5. Kitaev N. Reformer: The Efficient Transformer. URL: https://ai.googleblog.com/2020/01/reformer-efficient-transformer.html.