УДК 004.912

РАЗРАБОТКА СИСТЕМЫ ДЛЯ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ НАУЧНЫХ ТЕКСТОВ

Гайнутдинов М.Ф. (Национальный исследовательский Томский государственный университет),
Научный руководитель – канд. техн. наук, доц. Шкуркин А.С.
(Национальный исследовательский Томский государственный университет)

В современном мире важность получения новых знаний с наименьшими затратами времени велика как никогда, особенно в случае с научными текстами, так как в них содержится наиболее важная и актуальная информация о той или иной научной сфере или предметной области. Рост важности и спроса на быстрое освоение новой научной информации обеспечивают актуальность систем автоматического реферирования, то есть предоставления самых важных данных в наиболее сокращенном виде. Целью данной работы является создание системы автоматического реферирования научных текстов на русском языке с помощью использования методов машинного обучения.

Введение. На данный момент автоматическое реферирование научных текстов на русском языке не распространено, но подходы, общие для большинства языков, хорошо изучены и предлагают разные варианты: с помощью методов лингвистики, теории автоматов, машинного обучения, но, например, для абстрактивного подхода к реферированию, которое является крайне перспективным, наиболее подходит использование машинного обучения, но в области научных текстов на русском языке этот вопрос не так глубоко изучен. Также одной из ключевых особенностей именно научных текстов является использование формул и специальных символов, чьё реферирование также должно быть обработано, но на текущий момент мало изучено.
Также важной частью работы является предоставление возможности использования системой реферирования конечным пользователям, для чего было решено разработать веб-приложение.
Бесплатных и доступных отечественных решений не было найдено, а зарубежные сервисы не самым лучшим образом обрабатывают тексты на русском языке, что также является проблемой, решить которую берется данная работа.

Основная часть. Разрабатываемая система будет представлять из себя веб-приложение, предоставляющее доступ к реферированию научного текста из доступных научных сфер, с которыми сервис будет работать. Система будет предоставлять на выбор два подхода к реферированию, которые по своей сути и являются основными способами реферирования, а именно:
Экстракция - извлечение из текста самых важных блоков информации (абзацев, предложений и т.п.). В алгоритмах используются модели машинного обучения, например, Word2Vec.
Абстракция - генерация реферата, состоящего из нового текста, который не содержался в исходном тексте. В алгоритмах используются модели машинного обучения, например, Encoder-Decoder LSTM.
Для обучения моделей машинного обучения необходимы дата-сеты — большие массивы данных, содержащие, в данном случае, тексты научных статей по определенной научной области и их аннотации. Для получения и формирования данных для обучения моделей планируется использовать *веб-скрапинг* - получение данных путем извлечения их со страниц веб-ресурсов и дальнейший парсинг для формирования пригодных для обучения данных.

Выводы. Спроектирована система автоматического реферирования научных текстов на русском языке с помощью использования методов машинного обучения.