УДК 004.912

АВТОМАТИЧЕСКАЯ ТЕМАТИЧЕСКАЯ И СТРУКТУРНАЯ СЕГМЕНТАЦИЯ ТЕКСТОВ МЕДИЦИНСКИХ ЭЛЕКТРОННЫХ КАРТ

Функнер А. А. (Университет ИТМО)
Научный руководитель – к.т.н. Ковальчук С. В.
(Университет ИТМО)

В настоящее время отсутствует автоматический метод для оценки качества медицинских текстов на естественном языке. В данной работе предлагается подход тематического моделирования и сегментации для записей электронных медицинских карт разнообразной структуры и типов.

Введение. В электронной медицинской карте (ЭМК) содержится информация о пациенте и его лечении. Формат ЭМК определяется медицинской информационной системой (МИС) и лечебно-профилактическим учреждением (ЛПУ), в которых данная карта наполняется и хранится. Также в карте информация представлена в структурированном (табличные данные), полуструктурированном (анкеты и другие записи с полями) и неструктурированном виде (тексты без разметки на естественном языке). Для анализа и дальнейшего использования записей ЭМК важно оценивать качество вносимых записей и при возможности рекомендовать ЛПУ и медицинским сотрудникам способы улучшения качества таких записей. Однако, не существует алгоритмов и методов в открытом доступе для оценки качества медицинских текстов на естественном языке. Целью данного исследования является разработать подход для семантической оценки медицинских текстов ЭМК. В данной работе представлены результаты первых экспериментов по тематическому моделированию и сегментации для 80 тыс. текстов на естественном языке, переданных в Медицинский информационно-аналитический центр (МИАЦ) Санкт-Петербурга в 2020 году для больных с артериальной гипертензией и острым коронарным синдромом.

Основная часть. Основная сложность в семантическом анализе представленных данных – это разнообразие форматов и типов записей: тексты были собраны в 13 МИС и 106 ЛПУ. Во-первых, необходимо соотнести типы записей между МИС и ЛПУ, так как одни те же по семантике записи могут называться по-разному (например, «Консультация кардиолога» и «Осмотр врача»). Для решения данной проблемы было предложено классифицировать методом k-ближайших соседей названия записей по частично размеченным данным: было выделено 10 классов.

Во-вторых, записи могут быть состоять из подразделов и включать в себя другие типы записей. Например, запись «Консультация кардиолога» часто содержит раздел «Анамнез заболевания», что так же является самостоятельным типом записи. Поэтому на данном шаге записи двух наиболее представленных МИС были разбиты на подразделы с помощью регулярных выражений и ручной разметки. Данные подразделы стали основой для тематического моделирования и будущей сегментации неразмеченных данных.

В-третьих, были выделены темы с помощью тематического моделирования и была обучена модель тематической сегментации на основе разработанного ранее модуля тематической сегментации. Кроме того, была обучена нейронная сеть RNN для классификации текстов, с которой были сопоставлены результаты тематической сегментации. В отличии от первого подхода нейронная сеть требует больших вычислительных ресурсов и менее интерпретируема.

Выводы. Предложенный выше подход позволяет узнать, какие темы и разделы содержатся в записи ЭМК на естественном языке, а также определить среднюю по семантике запись, с

которой можно сравнивать новые записи и формировать рекомендации для ЛПУ по заполнению таких записей. С помощью данного подхода можно оценивать качество поступающих в МИАЦ записей. Кроме того, модели любого этапа данного подхода могут быть переобучены или дообучены для других нозологий, ЛПУ и МИС.

Функнер А.А. (автор) Подпись

Ковальчук С.В. (научный руководитель) Подпись