УДК: 004.896

Название: Разработка методов автоматической разметки полуструктурированных медицинских данных

Автор: Сластён Евгения Сергеевна, «Федеральное государственное автономное образовательное учреждение высшего образования „Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики“», г. Санкт-Петербург;

Контакты: slasten_ev@mail.ru, +79817961703;

Научный руководитель: Копаница Георгий Дмитриевич, «Федеральное государственное автономное образовательное учреждение высшего образования „Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики“», г. Санкт-Петербург;

Тезис доклада:

В настоящее время в медицинской сфере наибольшей ценностью обладают тексты (анамнезы, выписки, листы процедур и т.д.), которые содержат в себе колоссальный объем полезной информации для прогнозирования, моделирования и оценки медицинских процессов и течений заболеваний. Однако процесс получения необходимых сведений (извлечения сущностей) значительно затрудняется ошибками любых видов. Поэтому возникает потребность в разработке алгоритма для исправления ошибок, который бы качественно обрабатывал узкоспециализированные тексты (такие как медицинские). В работе рассматриваются основные существующие методы и алгоритмы обработки ошибок: их достоинства и недостатки. Решается проблема обработки ошибок в узкоспециализированных терминах. Проводится анализ публикаций на заданную тему. С целью выделения наиболее популярных методов и алгоритмов и выводов по практике применения.

Проблема ошибок в текста не является новой областью исследований, однако, как показывает практика, для каждого типа текста необходим свой подход: методы, которые подходят для обработки ошибок в художественный и социальных текстах без специальной лекции мало подойдут для обработки ошибок в текстах с “узкой” лексикой или значительным количеством аббревиатур, - необходимо решение.

В ходе работы было опробовано несколько популярных подходов:

- редакционное расстояние;
- расстояние Дамерау-Левенштейна;
- методы N-грамм слов и N-грамм букв;
- динамическое программирование;
- BK-деревья;
- Noisy channel;
- Soundex и т.д.

По каждому из них приведен анализ применения к, основные плюсы и минусы. Итоговый алгоритм состоит из комбинаций описанных выше методов, которые показали наилучший результат (время выполнения и качество исправления ошибок). Проводилась обработка реальных общих и аллергологических анамнезов (в приведением примеров) полученных из ФГБУ «НМИЦ им. В. А. Алмазова».

Результатом работы является пул методов, которые показывают желаемые результаты по обработке ошибок на реальных данных. Также вводится общая подготовка

данных (нормализация, стемминг/лемматизация, векторизация) для упрощения дальнейшей работы с текстовыми данными для выделения сущностей.