УДК 004.91

АНАЛИЗА СЛАБО ФОРМАЛИЗОВАННЫХ КОРПУСОВ ТЕКСТА: ОБЗОР МЕТОДОВ

Абрамовский И.С. (Национальный исследовательский университет ИТМО)
Научный руководитель –к. пол. н. директор ЦТЭП Чугунов А.В.
(Национальный исследовательский университет ИТМО)

В данной работе приводятся возможные методы анализа слабоформализованных текстов: комментарии и сообщения в социальных сетях, применяемые в машинном обучении. Автор рассматривает существующие корпуса текстов, а также методы анализа текстов, включенных в корпус.

Введение.

Вопросы анализа слабоформализованных текстов: новости, публикации и комментарии в социальных сетях рассматриваются в рамках корпусной лингвистики.

Корпусная лингвистика — актуальное направление современных прикладных исследований. Корпусная лингвистика – это раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов с использованием компьютерных технологий. На сегодняшний день, ввиду распространения методов машинного обучения в компьютерной лингвистике. возникает потребность в больших объемах аннотированного в том или ином аспекте текстового материала. Как отмечают Н.А. Власова, И.В. Трофимов, И.П. Сердюк и др., в рамках современных исследований компьютерной лингвистики с использованием методов машинного обучения актуальны два направления. Первое касается создания корпусов большого объема, различающихся лежащими в основе разметки теоретическими подходами, второе направление основано на поиске эффективных методов создания больших корпусов.

В.П. Захаров определяет корпус текстов, как унифицированный, структурированный и размеченный массив языковых (речевых) данных в электронном виде, предназначенный для определенных филологических и, более широко, гуманитарных изысканий. М.А. Костерин и К.В. Лагутина приводят классификацию корпусов текстов. Основанием классификации является тип текста. Авторы разделяют корпуса текстов на следующие категории:

1) корпуса интернет-текстов;
2) корпуса статей и официальных документов;
3) художественные тексты.

На сегодняшний день наиболее актуальным вопросом анализа корпусов текстов является использование методов машинного обучения.

Основная часть.

Машинное обучение – это научное исследование алгоритмов и статистических моделей, которые компьютерные системы используют для эффективного выполнения конкретной задачи без использования явных инструкций, опираясь на шаблоны и выводы. Эксперты выделяют в машинном обучении 3 основных направления в зависимости от характера обучения: обучение с учителем, обучение без учителя, обучение с подкреплением.

Анализе текстовых данных в машинном обучении используется методы регрессии, классификации и кластеризации. Алгоритмы регрессии обычно используются для статистического анализа. Такие методы позволяют анализировать модельные отношения между точками данных, также они могут количественно определять силу корреляции между переменными в наборе данных.

При кластеризации объекты с аналогичными параметрами группируются вместе (в кластер). Все объекты в кластере более похожи друг на друга, чем на объекты других кластеров. При классификации происходит процесс прогнозирования класса заданных точек

данных. Классы иногда называются метками или категориями. Классификационное прогнозирующее моделирование представляет собой задачу аппроксимации функции отображения (f) от входных переменных (x) к дискретным выходным переменным (y).

Т. Батура приводит алгоритм классификации текстов. Он состоит из 5 этапов: предобработка, индексация, выбор признаков, построение и обучение классификатора, оценка качества. Предварительная обработка текста включает в себя токенизацию, удаление функциональных слов (союзы, предлоги, артикли и пр.). Далее осуществляется морфологический анализ. В результате в качестве признаков документа выступают все значимые слова, встречающиеся в документе. Индексация документов – это построение некоторой числовой модели текста, которая переводит текст в удобное для дальнейшей обработки представление. Автор предлагает использовать на данном этапе метод «Мешка слов» (Bag-of-Words). Он позволяет представить документ в виде многомерного вектора слов и их весов в документе.

На этапе выбора признаков возможен выбор ряда методов: функция TF-IDF, LDA, Метод Байеса (NB), метод латентного распределения Дирихле (LDA), 5-балльная шкала типа Лайкерта, алгоритма случайного леса (RF), DEMATEL. Суть метода TF-IDF состоит в том, чтобы больший вес получали слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах. В основе метода DLA предположение о том, что в каждом документе смешаны разные темы, а в каждой теме – присутствует определенное распределение слов. Метод предполагает, что процесс порождения каждого слова состоит в том, чтобы сначала выбрать тему по распределению, соответствующему документу, а затем выбрать слово из распределения, соответствующего этой теме.

Выводы.

Таким образом, мы можем утверждать, что сегодня существует широкий список методов и алгоритмов текстового анализа, применяемых в корпусной лингвистике. Также методы машинного обучения активно используются в данной области, что делает их актуальными для анализа слабоформализованных корпусов текстов.

Работа выполнена в рамках проекта НИР Университета ИТМО № 621304, «Разработка сервиса тематической кластеризации корпуса текстов «Развитие цифрового государственного управления в Российской Федерации» на основе машинного обучения».

Абрамовский И.С. (автор) Подпись

Чугунов А.В. (научный руководитель) Подпись