УДК 004.912

Применение языковой модели BERT в задаче разрешения омонимии для русскоязычных текстов

Слапогузов А.П. (Университет ИТМО), **Малюга К.** (Университет ИТМО), **Цопа Е. А.** (Университет ИТМО)
Научный руководитель – кандидат технических наук, доцент Перминов И.В. (Университет ИТМО)

В данной работе рассматривается применение языковой модели BERT в задаче разрешения омонимии для русскоязычных текстов с использованием машинного обучения без учителя. Языковая модель BERT использовалась для получения векторных представлений слов, которые в дальнейшем были использованы для кластеризации с использованием алгоритма Affinity Propagation. Разработанный подход позволил достигнуть результатов, которые сопоставимы с другим подходами для русского языка.

Введение. С увеличением количества генерируемой информации задача обработки текстов на естественном языке становится всё актуальнее. Одна из отличительных особенностей таких текстов — это наличие в них неоднозначностей. Несмотря на то, что грамматическая и синтаксическая неоднозначность успешно решаются с помощью частеречной разметки и анализа зависимостей, морфологической и синтаксической информации недостаточно, чтобы однозначно определить значение слова в употребляемом контексте. Поэтому возникло такое направление в обработке естественных текстов как разрешение лексической многозначности. Из этой области отдельно выделяют задачу Word Sense Induction(WSI), которая отличается тем, что определяется не конкретное значение слова, а контексты, в котором употребляется слова, разбиваются на кластеры, в каждом из которых заданное слово употребляется в одном и том же смысл. В ходе активного развития данной области было организовано несколько соревновательных задач для английского языка (Senseval 2, 3, SemEval 2007, 2013, 2015) и одна для русского RUSSE'18. Для английского языка долгое время лучшие результаты были основаны на сложных графических моделях, но недавняя работа с использованием языковой модели BERT и динамическим числом кластеров показала лучший результат. Наилучший подход для русского языка в ходе RUSSE'18 использовал предобученную модель CBOW (Continuous Bag of Words) для получения векторных представлений слов и кластеризацию ближайших соседей. До текущего момента, языковая модель BERT не применялась в рамках задачи WSI для русскоязычный текстов, поэтому было решено оценить ее применимость для русскоязычных текстов.

Основная часть. Для сравнения был выбран подход А. Кутузова, который был разработан в рамках RUSSE'18. Данный подход для извлечения векторного представления слов использовал модель ruscorpora_upos_skipgram_300_5_2018 из RusVectōrēs, которая была обучена на Национальном Корпусе Русского Языка с использованием Continuous Skip-gram алгоритма. Далее, После преобразования слов в векторы, составлялся "семантический отпечаток"(как средний вектор между всеми словами в предложении), полученные "семантические отпечатки" кластеризовались с использованием алгоритма Affinity Propagation. Таким образом, чтобы оценить применимость языковой модели BERT для русского языка, в описанном подходе модель из RusVectōrēs была заменена моделью BERT, при этом был использован тот же алгоритм кластеризации. В качестве вектора слова использовались 4 последних слоя модели BERT. Каждый слой содержить по 768 значений, поэтому итоговый вектор имеет 3072-размерность. Такая высокая размерность данных может плохо на результаты кластеризации, поэтому размерность данных была уменьшена до

2 с использованием метода главных компонент. Таким образом, разработанный подход состоит из следующих шагов:

1. Получите векторное представление слов из модели BERT для каждого предложения.
2. Удалите знаки препинания, предлоги и союзы.
3. Создать «семантический отпечаток» как усредненный вектор уникальных слов.
4. Уменьшите размер данных с помощью метода главных компонент.
5. Примените алгоритм Affinity Propagation для кластеризации векторных представлений слов в группы, которые представляют смыслы слов.

Для оценки точности кластеризации использовалась метрика ARI, которая показывает сходство между двумя кластерами.

Выводы. В ходе данной работы был описан простой и эффективный метод для решения задачи Word Sense Induction. Данный метод основан на кластеризации векторных представлений слов, которые были получены с использованием языковой модели BERT. Подход был протестирован на трех наборах данных из RUSSE'18, для тестового набора wiki-wiki результат был выше на 15% по сравнения с базовым подходом от Кутузова А. Для двух других наборов данных результаты оказались незначительно ниже, что обусловлено различной структурой этих наборов. Продемонстрированные результаты позволяют положительно оценить применимость модели BERT для русского языка.

Слапогузов А.П. (автор) Подпись

Перминов И.В. (научный руководитель) Подпись