УДК 004.912

ИСПОЛЬЗОВАНИЕ РАЗЛИЧНЫХ ТИПОВ ЭМБЕДДИНГОВ В КАТЕГОРИАЛЬНОЙ КЛАССИФИКАЦИИ КОРОТКИХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Марченко А.М. (Университет ИТМО), **Овчинников И.Д.** (Университет ИТМО; Dasha.AI)
Научный руководитель – к.т.н, доцент Сметанников И.Б.
(Университет ИТМО)

Аннотация. В данной работе рассматривается возможность использования различных типов предобученных эмбеддингов в категориальной классификации коротких текстов на русском языке. Для изучения данной возможности исследуются существующие типы эмбеддингов, составленных на основе русскоязычных текстов, и сравниваются результаты категориальной классификации с использованием эмбеддингов разных типов.

Введение. Использование предобученных эмбеддингов играет важную роль в ходе категориальной классификации текстов. На данный момент существует множество различных типов предобученных эмбеддингов, по-разному представляющих поданный на вход текст.
В данной работе будут рассмотрены несколько различных типов предобученных русскоязычных эмбеддингов. Также будет построена модель категориальной классификации коротких текстов на русском языке. Полученная модель будет протестирована с различными предобученными эмбеддингами с целью определить их влияние на итоговый результат классификации.

Основная часть. Эмбеддинги или векторные представления слов повсеместно используются в задачах обработки естественного языка. В зависимости от той или иной задачи, могут быть использованы либо эмбеддинги, созданные на основе предоставленных данных, либо какие-либо предобученные эмбеддинги. На текущий момент существует множество различных предобученных векторных представлений, построенных с помощью различных алгоритмов на основе больших наборов данных. Однако такие векторные представления занимают больше места в памяти и могут плохо передавать специфику исследуемых в рамках решаемой задачи данных.

Для анализа в рамках данной работы были выбраны два следующих типа эмбеддингов:

- Эмбеддинги, обученные на основе предоставленных для решаемой задачи данных
- Предобученные эмбеддинги, полученные с помощью алгоритма GloVe и byte-pair encoding (BPEmb)

Первый подход к использованию эмбеддингов является наиболее простым способом получить векторные представления для своей задачи. Для его реализации были выполнены следующие шаги:

1. Подсчитаны уникальные слова в имеющихся данных
2. Вычислена максимальная длина предложения
3. Каждому слову присвоен свой индекс среди уникальных слов
4. Предложения преобразованы в числовые последовательности согласно индексам слов в них и дополнены нулями до вычисленной максимальной длины
5. Полученные числовые последовательности подавались непосредственно в эмбеддинг-слой модели категориальной классификации, инициализированный случайными векторами и обучавшийся в процессе тренировки модели

Второй подход представлен предобученными эмбеддингами BPEmb, которые были созданы на основе текстов из Википедии следующим образом:

1. Тексты были представлены как последовательности символов

2. Два наиболее часто встречающихся символа были заменены на один новый (операция объединения)

После нескольких таких объединений тексты были переданы алгоритму GloVe, с помощью которого были получены итоговые векторные представления. Данный подход интересен тем, что в отличие от так называемых word-level эмбеддингов (где векторное представление строится на основе целого слова, как, например в Word2Vec и обычном GloVe), эмбеддинги, полученные с помощью данного метода, строятся на основе частей слов (subword-level эмбеддинги), что может позволить построить векторные представления для слов, полностью не встречавшихся в исходном тренировочном наборе данных (что актуально для текущей задачи, так как в рассматриваемом наборе данных присутствуют ошибки и жаргонные слова).

С помощью описанных выше двух типов эмбеддингов были построены две модели категориальной классификации — сверточная нейронная сеть и двунаправленная LSTM сеть с использованием механизма внимания. По результатам экспериментов модели показали примерно одинаковый средний F1-Score по результатам 10-fold кросс-валидации, при этом средняя точность предсказаний у модели с BPEmb получилась больше.

Выводы. Данное исследование показывает, что эмбеддинги, полученные на основе byte-pair encoding могут быть использованы в задаче категориальной классификации коротких текстов на русском языке, несмотря на наличие ошибок и жаргонных слов в исходных данных. При этом BPEmb эмбеддинги даже превосходят по точности эмбеддинги, построенные на основе тренировочных данных.

Марченко А.М. (автор)

Сметанников И.Б. (научный руководитель)