Применение технологии Word2vec в задаче получения инверторов тональности

Полозов Илья Константинович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: ilya-polozov@mail.ru

Научный руководитель *— Волкова Ирина Анатольевна*

Автоматический анализ тональности является актуальной задачей. Во многих таких системах используются списки тональных слов. При этом тональность может меняться специальными словами — инверторами тональности, например, «избежать», «убрать». Поэтому особенно важной задачей является составление именно таких словарей. Многие системы используют словари инверторов тональности. Например, они используются в работе [3]. Внедрение словарей помогло увеличить точность классификации отзывов. В работе [2] строится классификатор, в котором одним из признаков является частота тональных слов в отзыве.

Большинство работ по составлению словарей инверторов тональности основано на ручном отборе. Например, в работе [1] описана система правил извлечения инверторов для китайского языка. Авторы размечают часть предложений вручную. Затем эта информация используется для пополнения множества инверторов.

В данной работе реализован новой подход к решению такой проблемы. Используются векторные представления слов, полученные с помощью Word2vec. Сначала из текста удаляются стоп-слова, знаки пунктуации, все слова переводятся в начальную форму. Потом для всех нетональных слов подсчитывается частота их встречаемости рядом с тональными словами. Если она больше порога, то такие слова считаются кандидатами в инверторы. Затем каждое появление инвертора в тексте рядом с тональным словом заменяется единым токеном «кандидат_тональное слово».

После этого обучается Word2vec. Определяется тональность каждого токена «кандидат_тональное слово». Считается среднее расстояние по косинусной мере между комбинированным токеном и тональными словами. Если расстояние до положительных слов меньше, чем до отрицательных, то токен имеет положительную тональность. При этом если само тональное слово, входящее в такой токен, является отрицательным, то значит его тональность поменялась инвертором. Нетональное слово из комбинированного добавляется в

список инверторов.

Литература

1. Xu G. Huang C. Mining Chinese Polarity Shifters // Chinese Lexical Semantics: 16th Workshop, Beijing, China, 2015, P. 244-=251.
2. Morsy S. A. Rafea A. Improving Document-Level Sentiment Classification Using Contextual Valence Shifters The American University in Cairo // International Conference on Application of Natural Language to Information Systems, Groningen, The Netherlands, 2012, P. 253–258.
3. Ye Z. Li F. Encoding Sentiment Information into Word Vectors for Sentiment Analysis // Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, USA, 2018, P. 997–1007.