Берсенева, Макунина, Рубанова, Суворова

Реферат по статье Building Large-Scale Twitter-Specific Sentiment Lexicon:

A Representation Learning Approach

Ссылка на статью:  <https://www.aclweb.org/anthology/C14-1018.pdf>

В представленной статье авторы ставят перед собой цель построения large-scale тонального лексикона для Твиттера. Основными задачами в работе были разработка эффективного признакового представления тональных фраз и получения обучающего датасета, требующего минимальной ручной разметки.

В качестве датасета для создания лексикона авторы используют архив размером 10 миллионов твитов. Они отобрали твиты со смайликами - “:) : ) :-) :D =) as positive and :( : ( :-( as negative”. На его основе они создают базовый список слов (sentiment seeds), который расширяют, используя схожие слова из Urban Dictionary. Качество работы классификатора твитов (LibLinear) исследователи оценивают на материалах корпуса SemEval 2013.

При обучении LibLinear авторы провели два эксперимента. В первом они брали набор признаков, полученный только с помощью TS-Lex (Unique), а во втором (Appended) этот набор признаков добавляли к уже имеющемуся в SemEval 2013.

Авторы используют обучение с учителем, объясняя это тем, что для обучения без учителя необходим тональный лексикон высокого качества. В их же случае целью было создать не максимально точный, а как можно более широкий лексикон. Так, они допускают возможность большего количества ошибок в угоду большего размера лексикона.

Одним из преимуществ подхода, представленного в статье, является охват разговорной речи, которого они достигли с помощью Urban Dictionary. Кроме того, авторы используют новый подход к построению эмбеддингов - SSPE, который учитывает не только синтаксический контекст, но и тональность предложения.

Авторы приводят сравнение своего тонального лексикона (TS-Lex) с другими: тремя традиционными (HL, MPQA and NRC-Emotion) и двумя также основанными на текстах твитов (HashtagLex and Sentiment140Lex). По размеру TS-Lex выигрывает у всех традиционных лексиконов, но уступает другим твиттерским. Однако классификатор твитов, использующий TS-Lex, справляется лучше, чем классификаторы, использующие все вышеупомянутые лексиконы. Даже без расширения базового списка тональных слов с помощью Urban Dictionary авторы достигают точности 74.64%, что превышает результаты остальных лексиконов. На TS-Lex (с расширенными sentiment seeds) классификатор справляется ещё лучше - с точностью 78.07%.

Также авторы сравнивают свой подход к созданию эмбеддингов (SSPE) с C&W, W2V, MVSA и ReEmbed(C&W) и приходят к выводу, что SSPE даёт лучшие результаты на двух традиционных лексиконах.

Несмотря на то, что авторы довольно подробно объясняют свою идею и процесс работы, некоторые моменты показались нам не совсем понятными. Во-первых, в пункте 3.1, где объясняется, как работает SSPE, неясно, как вычисляется полярность предложения (в формуле 2). Во-вторых, в пункте 3.2, где говорится о расширении sentiment seeds, используются схожие слова из Urban Dictionary. Их схожесть для повышения качества и количества оценивается с помощью KNN. Непонятно, каким представлением слов авторы пользовались для применения KNN: у слов из SSPE имеется своё представление, а как были обработаны слова из Urban Dictionary не описано.

В заключение хотелось бы сказать, что исследование кажется очень проработанным - авторы постарались улучшить все аспекты создания тонального словаря, начиная с подбора материала датасетов и заканчивая созданием собственных эмбеддингов.