ИСПОЛЬЗОВАНИЕ ФОНОСЕМАНТИЧЕСКОЙ ОЦЕНКИ СЛОВ-МОДИФИКАТОРОВ В СЛОВАРЯХ ОЦЕНОЧНОЙ ЛЕКСИКИ

Игнатьев А.М.
Национальный университет гражданской защиты Украины,
г. Харьков, ул. Чернышевская, 94, тел. 050-733-78-33,
e-mail: Ignatiew@yandex.ru

За последние десять лет интерес к области анализа эмоциональной тональности текстов сильно возрос. Стоит отметить, что на текущем этапе развития в данной области существует много нерешенных проблем. Анализ эмоциональной окраски текста затруднителен не только в связи с проблемой выделения единиц оценки тональности, но и ввиду неоднозначности эмоциональной составляющей лексических компонент. Например, в рамках одной и той же предметной области "высокая стоимость" – отрицательный аспект товара, в то время как "высокое качество" – положительный. Таким образом, используемые методы тонального анализа предметно зависимы, т.е. для различных предметных областей необходимо составлять различные словари.

Основные подходы к определению тональности можно разделить на следующие категории [1]:

1. Подход, основанный на правилах (rule-based approach), заключается в применении набора правил, выявленного экспертами на основе анализа предметной области.

2. Подход, основанный на использовании словарей оценочной лексики (affective lexicons). Для каждого слова, встречаемого в документе, из словаря получают значение тональности. Чтобы получить итоговую тональность необходимо взять среднее арифметическое или вычислить сумму значений тональности всех слов из документа.

3. Подходы, основанные на обучении с учителем (supervised learning). Алгоритм классификации тренируется на основе обучающей выборки (корпуса), состоящей из документов, классы которых заранее известны.

4. Подходы, основанные на обучении без учителя (unsupervised learning). Отличие состоит в том, что в этом случае для тренировки алгоритма используется обучающая выборка, состоящая из документов, классы которых заранее неизвестны (или известны, но эта информация не используется алгоритмом).

В ходе экспериментов методы, основанные на словаре эмоциональной лексики, при решении задачи автоматической классификации текстов по тональности показали результаты, несколько превосходящие результаты метода опорных векторов (Support Vector Machine, SVM) и простейшего способа классификации (baseline) [2]. Исследования показали, что методы на основе словаря показывают достаточно неплохие результаты при классификации текстов.

Однако, кроме оценочных слов для выбранной предметной области, в текстах встречается множество слов-модификаторов, в зависимости от которых можно увеличивать или уменьшать вес следующего за ним оценочного слова. Все слова-модификаторы можно разделить на две группы в зависимости от их

направленности. К первой группе относятся слова-модификаторы, которые увеличивают эмоциональный вес соседнего слова (например, «особенно»), ко второй – те, которые уменьшают ее (например, «незначительно»). Для изменения веса следующего за модификатором слова можно использовать метод простого сложения и вычитания. Если модификатор увеличивает эмоциональный вес слова, то к его оценке можно добавлять фиксированное число, иначе – вычитать это же число.

Однако, недостатком данного подхода является то, что он не учитывает широкий диапазон модификаторов в пределах группы. Например, модификатор «абсолютно» очевидно сильнее изменяет эмоциональный вес слова, чем модификатор «значительный». Также при усилении слова с уже большим весом увеличение его эмоционального веса должно быть больше по сравнению со словом, обладающим меньшим весом. Например, «действительно восхитительный» и «действительно хороший». Возможно использовать подход, который в зависимости от слова-модификатора изменяет вес соседнего слова на некоторый процент. Например, если слово «хорошо» имеет вес 5, а модификатор «действительно» имеет относительную оценку 20%, то «действительно хорошо» будет иметь вес 5*(100% + 20%) = 5*1,2 = 6. В качестве модификаторов используются наречия и прилагательные. Такой подход был рассмотрен в работе [3], в которой процентные значения для слов-модификаторов фиксировались на основе экспертных оценок.

Нами предлагается поход, в котором процентные значения для слов-модификаторов будут вычисляться на основе их фоносемантических оценок по различным шкалам. Таких шкал может быть несколько («хороше-плохое», «быстрое-медленное» «сильное-слабое» и т.д.). Оценка каждого слова-модификатора вычисляется как среднее арифметическое всех его фоносемантических оценок по всем предложенным шкалам. Такой поход позволит учесть эмоциональную окраску самих слов-модификаторов и, как следствие, получить более точные веса анализируемых слов.

Предлагаемый способ оценки весов слов позволит избежать коллизий, возникающих при применении слов, выражающих отрицание, к словам-модификаторам. Простое инвертирование эмоционального веса оценочного слова хорошо работает лишь в некоторых случаях, но часто может привести к нежелательному результату. Например, «не очень хорошо» может оказаться более отрицательно, чем «плохо». В работе [3] вместо смены знака, значение эмоционального веса сдвигается к противоположной полярности на фиксированную величину. Нами предлагается учитывать отрицание изменением веса соседнего слова на процент его модификатора, взятого со знаком «минус».

Список литературы

1. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008. - pp. 1-135.

2. Российский семинар по оценке методов информационного поиска (РОМИП). URL: http://romip.ru.

3. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* Lexicon-based methods for sentiment analysis // Computational Linguistics, 37(2): 2011. - pp. 267–307.