# ПРОБЛЕМЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

**Ю. В. Кулинкович**
Минск, МГЛУ

## МЕТОДЫ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ЭМОЦИОНАЛЬНО ОКРАШЕННОЙ ЛЕКСИКИ ИЗ ПИСЬМЕННОГО ТЕКСТА

В статье рассматривается проблема определения тональности текста отзыва пользователя Интернета о некотором объекте. Отмечается, что тональность текста формируется с помощью тонально окрашенных слов и выражений, образующих сентиментный лексикон. Анализируются преимущества и недостатки нескольких распространенных методов автоматического выделения эмоционально окрашенных лексических единиц из письменных текстов: метода анализа пар прилагательных; метода, основанного на семантической близости синонимов, и метода, основанного на семантический близости слов с учетом частотных данных об их употреблении в текстах определенной тональности.

В последнее время в компьютерной лингвистике активно развивается направление, связанное с анализом и извлечением мнений из разных типов текстов, определением на их основе общей тональности текстов (автоматическим проведением так называемого сентимент-анализа). С помощью автоматического анализа тональности можно идентифицировать эмоциональную окраску текста на трех уровнях: на уровне всего документа, на уровне предложения или короткого высказывания, на уровне объекта, относительно которого высказывается мнение [1, с. 23–24]. Анализ тональности на уровне документа позволяет определить сформулированную автором текста общую оценку, например, анализ отзыва пользователя о продукте или услуге для определения позитивного или негативного отношения автора к объекту отзыва в целом. Этот уровень классификации подразумевает, что один документ содержит мнение только об одном объекте и только одного субъекта и не подходит для анализа текстов, в которых рассматриваются несколько объектов или они сравниваются между собой. Однако мнения даже об одном объекте часто бывают противоречивыми, поэтому анализ тональности на уровне документа дает только общее представление об отношении автора к объекту. Анализ тональности отдельного предложения или короткого высказывания определят оценочное мнение автора, высказанное в одном предложении. Если автор высказывает несколько противоположных оценок, то принято говорить о конфликтном мнении, например, *Этот телефон очень красивый, но ужасно тяжелый.* Классификация тональности предложения тесно связана с определением субъективности предложения. Однако, как подчеркивают некоторые исследователи, автор текста может выражать свое мнение и с помощью объективных предложений, например, *Я купил телефон на прошлой неделе, и корпус вчера треснул* [1; 2].

Основным индикатором тональности текста являются тонально окрашенные слова, например, лексические единицы с положительной семантикой *хороший*, *вкусный* или единицы с негативной семантикой *плохой*, *ужасный*, *злой*. Помимо отдельных слов, в тексте могут присутствовать тонально окрашенные выражения, например, *не могу жить без…* , *голову бы отдал за…* . В совокупности слова и выражения со значением тональной ориентации формируют сентиментный лексикон. В настоящее время внимание исследователей в сфере компьютерной лингвистики направлено на разработку оптимальных алгоритмов автоматического создания тонального лексикона, поскольку от него зависит качество анализа тональности текста. Рассмотрим некоторые методы автоматического выделения эмоционально окрашенной лексики из письменного текста и формирование сентиментного лексикона подробнее.

Метод анализа пар прилагательных предполагает определение тональности прилагательных, соединенных в пары с помощью союзов *и*, *или*, *но*, *или...или*, *ни...ни* и выделенных из большого массива неразмеченных текстовых документов. Факт объединения прилагательных с помощью перечисленных выше единиц задает лингвистические ограничения тональности для данных прилагательных. Так, союзом *и* обычно объединяют два прилагательных одинаковой тональности, в то время как союз *но* объединяет два прилагательных противоположной тональности. Данный метод позволяет автоматически выделять прилагательные, которые имею разную тональную окраску в зависимости от предметной области [3, p. 176]. Например:

*Кофе* ***горячий*** *и* ***очень вкусный*** (позитивная тональность);

*Ноутбук очень* ***шумный*** *и* ***горячий*** (негативная тональность).

Для реализации метода анализа пар прилагательных необходим небольшой список общеупотребительных слов с явно выраженной тональной окраской, например, *хороший*, *замечательный*, *плохой*, *ужасный*. В ходе анализа из массива текстовых документов извлекаются все прилагательные, которые употребляются в перечисленных выше союзных конструкциях вместе с прилагательными из списка общеупотребительной лексики, и им приписывается соответствующая тональность. Несмотря на то, что анализ пар прилагательных позволяет установить тональность прилагательных в рамках определенной предметной области, некоторые прилагательные могут иметь противоположную тональность даже в одной предметной области [4, p. 142]. Например:

*Батарея работает очень* ***долго*** (позитивная тональность);

*Камера очень* ***долго*** *фокусируется* (негативная тональность).

Основная идея второго метода, основанного на семантической близости синонимов, заключается в том, что синонимы слова с положительной семантикой также имеют позитивную окраску, а антонимы слова – противоположную. Данный метод предполагает использование таких онтологических ресурсов, как, например, *WordNet* для английского языка, которые содержат списки синонимов и антонимов для каждого значения слова. На начальном

этапе необходимо сформировать небольшой список слов с однозначной тональной ориентацией. Далее система рекурсивно просматривает иерархический граф онтологии, выбирая прилагательные, синонимичные или антонимичные для списка слов с положительной или отрицательной семантикой. Найденные слова добавляются в исходный список, поэтому при рассмотрении каждого последующего прилагательного из дерева онтологии список слов увеличивается. Метод учета семантической близости синонимов ограничен прилагательными, которые зафиксированы в онтологиях [5, p. 1116].

Среди методов формирования тонального лексикона наибольшую популярность приобрел метод, основанный на семантический близости слов с учетом частотных данных об их употреблении в текстах определенной тональности. Этот метод имеет следующие основные характеристики. В о - п е р в ы х, по качеству получаемых результатов он значительно превосходит описанные выше методы. В о - в т о р ы х, лексику можно выбирать из текстов разных предметных областей. И, в - т р е т ь и х, полнота словаря ограничена только используемым для тренировки корпусом. Чем больше объем корпуса, тем больше различных слов будет покрыто словарем. Идея данного метода заключается в следующем: чем чаще слово используется в текстах с положительной тональностью, тем большую позитивную ориентацию оно имеет, и наоборот. Равная частота употребления слова свидетельствует о его нейтральности.

Своеобразным толчком к широкому распространению данного метода послужило исследование П. Терни, в котором предложен трехэтапный алгоритм классификации тональности текста отзыва на основе сентиментной ориентации его слов и словосочетаний [2]. На первом этапе классификации определяется список отдельных слов и последовательностей слов для их дальнейшего анализа по заданным маскам на основе тегов частей речи (прилагательное, прилагательное + существительное, наречие + прилагательное и т.д.). Таким образом, можно получить список слов и словосочетаний, например, *long battery life*, *short battery life*, *quiet car*, *quiet audio.* На втором этапе для каждого предложения вычисляется значение его тональной ориентации на основе поточечной взаимосвязанной информация (PMI) от +1 до –1:

$$SO(t) = PMI(t, pos) - PMI(t, neg),$$

где $SO(t)$ – значение тональной ориентации термина $t$; $PMI(t, pos)$ – поточечная взаимосвязанная информация, отражающая вероятность встречаемости термина t в позитивном контексте *pos*; $PMI(t, neg)$ – поточечная взаимосвязанная информация, отражающая вероятность встречаемости термина $t$ в негативном контексте *neg*. Положительный знак тональной ориентации предложения указывает на его положительную тональность, а большее абсолютное значение – на более явно выраженную тональность.

*PMI* отражает степень синтаксической зависимости двух словарных единиц и показывает, насколько заданное слово или предложение соотносится с положительным контекстом.

$$PMI(t, pos) = \log \frac{P(t, pos)}{P(t)P(pos)} \qquad PMI(t, neg) = \log \frac{P(t, neg)}{P(t)P(neg)}$$

где $P(t, pos)$ – вероятность употребления термина $t$ в положительном контексте pos; $P(t, neg)$ – вероятность употребления термина $t$ в отрицательном контексте neg; $P(t)$ – вероятность появления термина $t$ в обучающей выборке (отношение количества употреблений к общему количеству слов в корпусе); $P(pos)$ – вероятность появления положительного контекста *pos* в обучающей выборке; $P(neg)$ – вероятность появления отрицательного контекста *neg* в обучающей выборке.

Для вычисления вероятности употребления термина в позитивном и негативном контексте предложено два подхода. Первый подход использует поисковые запросы пользователя в виде конкатенации анализируемого слова и слова с явной положительной (например, *love*) или явной отрицательной (например, *hate*) тональностью. Так, поисковая система Google дает следующие результаты, на основе которых можно определить тональную ориентацию выражения:

| Запрос | Количество найденных |
|---|---|
| *long battery life* | 178,000,000 |
| *hate . * long battery life* | 6 |
| *love . * long battery life* | 36,700,000 |
| *hate* | 579,000,000 |
| *love* | 3,290,000,000 |

| | |
|---|---|
| *awesomeness* | 0.91325 |
| *love* | 0.86665 |
| *fabulous* | 0.86415 |
| *hilarious* | 0.84615 |
| *disgust* | –0.758 |
| *horrid* | –0.875 |
| *hate* | –0.92365 |

Второй подход связан с созданием корпусов текстов с положительной и отрицательной тональностью и вычислением частоты встречаемости в них определенных лексических единиц. С этой целью можно использовать отзывы, в которых пользователь указывает рейтинг продукта. Отзывы с низким рейтингом буду содержать в основном предложения с негативной тональностью, а отзывы с высоким рейтингом – с позитивной тональностью. В результате применения второго подхода был, например, получен следующий список слов и выражений со значением тональной ориентации [2, p. 897]:

Необходимо отметить, что качество и полнота сентиментного лексикона имеет большое значение не только для определения тональности текста, но и для других задач анализа мнений.

## ЛИТЕРАТУРА

1. *Сибиряков, А.* Извлечение мнений о товарах из форумов и блогов с учетом тональности / А. Сибиряков. – М. : Прогресс, 2012. – 117 с.
2. *Turney, P. D.* Thumbs up or thumbs down: semantic orientation applied to unsupervised classification of reviews / P. D. Nurney // Proc. of Annual Meeting of the Assoc. for Computational Linguistics (ACL–2002). – 2002. – P. 891–899.
3. *Hatzivassiloglou, V.* Predicting the semantic orientation of adjectives / V. Hatzivassiloglou, K. R. McKeown // Proc. of ACL–97, 35th Annual Meeting of the Assoc. for Computational Linguistics. – Madrid, 1997. – P. 174–181.
4. *Ding, X.* A holistic lexicon-based approach to opinion mining / X. Ding, B. Liu, P. S. Yu // Proc. of the Conf. on Web Search and Web data Mining (WSDM–2008). – 2008. – P. 141–149.
5. *Kamps, J.* Using WordNet to measure semantic orientation of adjectives / J. Kamps, M. Marx, R. J. Mokken, M. D. Rijke // Proc. of LREC–04, 4th Intern. Conf. on Lang. Resources and Evaluation. – Libon, PT. – 2004. – Vol. IV. – P. 1115–1118.

The article deals with the problem of emotionally colored lexical units automatic extraction from a written text. Some progressive methods of sentiment lexicon formation are under consideration.

**М. В. Масловская**
Минск, МГЛУ

## СТРУКТУРА МАШИННОГО ПЕРЕВОДА НА БАЗЕ НЕЙРОННЫХ СЕТЕЙ

В статье рассматривается структура машинного перевода на базе нейронных сетей и проводится сравнение с ранее существующими системами машинного перевода.

Искусственные нейронные сети (ИНС) разработаны по аналогии с процессами обработки информации человеческим мозгом. Нейрон живого организма – нервная клетка, единица нервной системы, обрабатывающая, хранящая и передающая некую информацию с помощью импульсов по сети другим нейронам. Искусственный нейрон выполняет схожую функцию в искусственной нейронной сети.

ИНС получили широкое распространение при решении различного вида прикладных задач, в том числе и задач прикладной лингвистики в рамках