ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 122
45th International Philological Conference (IPC 2016)

Автоматическое составление тонального словаря для процедур сентиментного анализа

Протопопова Екатерина Владимировна[1] Букия Григорий Теймуразович[2] Митрофанова Ольга Александровна[3]

[1]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия
[2]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия
[3]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия

Sentiment analysis of reviews based on automatically developed lexicon

Protopopova Ekaterina[1] Bookia Grigoriy[2] Mitrofanova Olga[3]

[1]Saint Petersburg State University (SPSU), St. Petersburg, Russia
[2]Saint Petersburg State University (SPSU), St. Petersburg, Russia
[3]Saint Petersburg State University (SPSU), St. Petersburg, Russia

Аннотация

В докладе описаны эксперименты по автоматической обработке тональной лексики. Особенностью нашего подхода является применение авторской метрики близости для ранжирования тональных лексем и конструкций. Исследование выполнена на основе русскоязычных корпусов текстов отзывов по данным сервиса «Яндекс.Маркет» и ряда интернет-магазинов.

Abstract

The talk deals with experiment on automatic processing of sentiment words and constructions. Peculiarity of our approach is determined by the similarity measure developed by the authors and applied to ranging of sentiment units. Our research is based on the Russian corpora of camera reviews from Yandex.Market and various Internet-shops.

Ключевые слова: сентиментный анализ, тональный словарь, русскоязычные корпусы.

Keywords: Sentiment analysis, sentiment lexicon, Russian corpora.

Введение

Для автоматической оценки тональности текстов в системах сентиментного анализа используются

тональные словари, содержащие как общие тональные слова (*хороший*, *удобно*, *нравиться* и т.д.), так и относящиеся к специальной области (*мыльный*, *нерезко*, *мажет* и т.д.). Каждому слову или словосочетанию приписывается оценка, характеризующая его положительную или отрицательную окраску. Часто такие словари составляются вручную, и их слабые стороны проявляются при подключении к системам автоматической обработки текстов.

Мы предлагаем статистический метод извлечения оценочной лексики из корпуса отзывов, имеющих жесткую структуру: «Достоинства» и «Недостатки». Лексемы и двухсловные конструкции, выражающие оценку, ранжируются на основании связи с одним из указанных двух полюсов. Эта связь оценивается с помощью критерия (Bukia et al., 2015), характеризующего вероятность того, что отзыв с данным словом относится к достоинствам / недостаткам. Для извлечения конструкций используются шаблоны и определяется обобщение меры взаимной информации.

Опыт создания русскоязычных тональных словарей

Традиционные методы составления тонального словаря имеют слабые стороны: невозможно охватить все особенности интернет-лексики; необходимо пополнять тональный словарь специальной лексикой для разных типов задач; пропускаются конструкции, характерные для отрицательных отзывов, но не воспринимаемые человеком как имеющие тональность; встречаются сложные лексические конструкции, не попавшие в словарь, и пр.

Многие системы сентиментного анализа используют словари, составленные вручную и/или полуавтоматическими методами. Так, в проекте SentiWordNet (http://sentiwordnet.isti.cnr.it) синсетам из WordNet приписывается положительная, нейтральная или отрицательная тональность. Для русского языка существующие ресурсы обсуждаются в соревновании SentiRuEval (http://www.dialog-21.ru/evaluation/2016/sentiment/); все свободно распространяемые словари оценочных слов (Kotelnikov et al., 2016) собирались полуавтоматически. Отбор кандидатов в оценочные выражения производится с помощью машинного обучения или иначе (в частности, путем перевода с английского и последующей коррекции (Ulanov, Sapozhnikov, 2013)), используются начальные словари, включающие узкий круг тонально окрашенных лексических единиц. В (Chetviorkin, Loukachevitch, 2012) описывается метод построения словаря на основе текстов отзывов с использованием классификатора, для обучения которого применяются грамматические, частотные и другие характеристики слов.

В (Blinov, Kotelnikov, 2014) с целью пополнения тонального словаря используются векторные представления слов. В (Dubatovka et al., 2016) описан метод, опирающийся на минимальное количество размеченных обучающих данных и позволяющий извлекать оценочные слова при помощи графовой модели, построенной с применением синтаксических шаблонов.

Простейший способ извлечения слов для последующей тональной разметки предложен в (Ivanov et al., 2015): авторы используют специфический корпус отзывов, в которых выделяются разделы «Достоинства» и «Недостатки». Из каждого раздела

выделяются наиболее частотные прилагательные, наречия и глаголы, таким образом, в словарь после ручной оценки попадают только однословные единицы.

Процедура построения тонального словаря

Корпусы интернет-отзывов зачастую содержат пользовательские оценки продукта, тем не менее, эти оценки слабо коррелируют с текстом отзыва. Однако если структура отзыва предполагает разделение на два раздела: «Достоинства» и «Недостатки», пользователи их не смешивают. Это позволяет отнести к тональным словам лексику, характерную для раздела достоинств и не характерную для раздела недостатков (положительная тональность), и наоборот (отрицательная тональность).

Мы использовали корпус отзывов на фотоаппараты сервиса «Яндекс.Словари», который организован на основе бинарной структуры. Таким образом удалось сформировать два подкорпуса одинакового объёма (20758 отзывов), состоящих из текстов – описаний а) достоинств и б) недостатков.

Основа нашего критерия – оценка корреляции двух случайных событий: «отзыв содержит слово » и «отзыв описывает достоинства». Ясно, что если одно событие с большой вероятностью влечет другое, то, скорее всего, лексема имеет положительную тональность. Для оценки корреляции мы построили таблицу сопряженности и воспользовались статистикой .

Таблица сопряженности представляет собой набор статистик (a, где a – частота данной лексемы в корпусе достоинств, b – в корпусе недостатков, c – частота прочих лексем в корпусе достоинств и d – частота прочих лексем в корпусе недостатков. Статистика критерия согласия Пирсона для таблиц сопряженности определяется следующим образом:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)},$$

где $n = a + b$.

Тем самым, формируется ядро тонального словаря – тональные лексемы.

Выделение тональных конструкций опирается на простейшие биграммные шаблоны, например, ADJ + N, V + ADV, V + N и т.п. Поскольку частоты биграмм намного меньше частот составляющих, применить стандартный метод не получается из-за высокой доли шума. Поэтому при анализе учитывается тональность как конструкции, так и каждой ее составляющей.

Результаты экспериментов

Для выделения тональных лексем и конструкций был эмпирически подобран порог значения критерия , который позволял отделить нейтральные единицы словаря от окрашенных. Таким образом был получен тональный словарь, включающий около 1,5 тыс. положительно окрашенных и около 2 тыс. отрицательно окрашенных единиц.

Наибольший вес получили тональные слова, выражающие общие

свойства объектов: *отличный*, *хороший*, *удобный*, *качество*, *компактный* (достоинства) и *недостаток*, *иногда*, *слабый*, *отсутствие* (недостатки) и т.д. Оказалось, что наречие *иногда*, на первый взгляд, нейтральное, встречается почти исключительно в корпусе недостатков.

Удалось выделить конструкции, характеризующие как продукт в целом, так и отдельные его характеристики: *стильный фотоаппарат*, *симпатичный дизайн*, *добротный корпус* и *отвратительный видеоискатель*, *большой шум*, *нечеткие фото* и т.д.

Заключение

Представленный метод составления тонального словаря легко адаптируется для решения широкого спектра задач. Алгоритм предполагает лишь наличие тематического корпуса отзывов с выделенными полями достоинств и недостатков. Обучение производится на основе простого в реализации критерия согласия Пирсона ( ) для таблиц сопряженности. С помощью шаблонов извлекаются не только оценочные конструкции, но и характеристики отдельных составляющих продукта (качество фотографии, объективов, вес и пр.), что дает определенные преимущества представленному методу в прикладных задачах.

Благодарности

Исследование поддержано грантом РФФИ № 16-06-00529.

Литература

Blinov P., Kotelnikov E. Using Distributed Representations for Aspect-Based Sentiment Analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2014». №. 13 (20). Vol. 2. 2014. P. 68–79.

Bukia G., Protopopova E., Mitrofanova O. A Corpus-Driven Estimation of Association Strength in Lexical Constructions // Proceedings of the AINL-ISMW FRUCT. 2015. P. 147–152.

Chetviorkin I.I., Loukachevitch N.V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // Proceedings of COLING 2012: Technical Papers. 2012. P. 593–610.

Dubatovka A., Kurochkin Yu., Mikhailova E. Automatic Generation of the Domain-Specific Sentiment Russian Dictionaries // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016». №. 15 (22). 2014. P. 146–154.

Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2015». №. 14 (21). Vol. 2. 2015. P. 22–33.

Kotelnikov E.V., Bushmeleva N.A., Razova E.V., Peskisheva T.A., Pletneva M.V. Manually Created Sentiment Lexicons: Research and Development // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016». №. 15 (22). 2016. P. 300–316.

Ulanov A., Sapozhnikov G. Context-Dependent Opinion Lexicon

Translation with the Use of a Parallel Corpus // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2013». №. 12 (19). 2013. P. 165–174.

References