Автоматическое извлечение оценочных слов для конкретной предметной области

Automatic extraction of domain-specific opinion words

Четвёркин И. И. (ilia2010@yandex.ru)

Факультет вычислительной математики и кибернетики МГУ

Лукашевич Н. В. (louk_nat@mail.ru)

НИВЦ МГУ

Для эффективного извлечения мнений из текстов необходимо знание оценочных слов и выражений из рассматриваемой предметной области. Мы предлагаем новый подход к автоматическому извлечению оценочных слов, основанный на работе с несколькими корпусами текстов и вычислении с их помощью характеристик слов.

1. Введение

В настоящее время на страницах сети Интернет можно найти множество отзывов о тех или иных товарах, фильмах, книгах и т. п. Эти отзывы содержат много полезной информации, поэтому ее важно автоматически извлекать и предъявлять пользователям [6, 11].

Мнения пользователей о продукте часто выражаются посредством оценочных слов и выражений, которые несут в себе некоторую положительную или отрицательную оценку. Поэтому важным фактором качественного извлечения мнений о той или иной сущности является знание оценочных слов и выражений, которые используются в данной области. Проблема состоит в том, что невозможно заранее собрать список оценочных слов и выражений, которые будут применимы для всех предметных областей, поскольку некоторые оценочные выражения употребляются только в конкретных предметных областях, другие являются оценочными в одной области и не являются оценочными в другой.

Так, например, выражение «хочу еще сходить посмотреть» является характерным оценочным выражением для фильмов и небольшого количества других областей. Значимую часть оценочных слов не найти ни в каких словарях, например, «никакущий», Иногда трудно догадаться, что то или иное слово в контексте может употребляться как оценочное, как, например, слово «скомканный» о фильмах или книгах.

Таким образом, актуальной является задача автоматического формирования списка наиболее употребительных оценочных слов для данной предметной области.

В данной статье мы рассмотрим методы автоматического получения оценочных слов на основе нескольких корпусов текстов, которые можно автоматически построить для многих предметных областей, а именно, корпуса отзывов о сущности с вручную проставленными потребителями оценками, корпуса нейтральных описаний сущностей и нейтрального контрастного корпуса, например, составленного из потока общезначимых новостей.

Из указанных корпусов мы извлекаем списки слов, упорядоченные по значению различных признаков, оцениваем качество этих списков относительно содержания в них оценочных слов, и исследуем способы комбинирования этих списков для получения лучшего качественного состава по оценочным словам. Определение ориентации оценочных слов не производится.

Рассматривается предметная область отзывов о фильмах.

2. Методы автоматического извлечения оценочных слов

Существует два основных подхода к автоматическому выделению оценочных слов из текстов. Первый

подход базируется на информации из словарей или тезаурусов. В данном подходе обычно выбирается небольшое начальное множество слов, которое формируется вручную, и затем дополняется с помощью словарей и тезаурусов. Основной принцип заключается в том, что если слово оценочное, то и его синонимы будут оценочными и антонимы (возможна только смена ориентации), Поэтому, имея слова из начального множества, можно с помощью этих связей составить новое множество, которое будет более полным [5], В [3] на основе толкований слов в словаре выясняется их ориентация (положительная или отрицательная), Основная идея заключается в том, что слова с одинаковой ориентацией имеют «похожие» толкования. Таким образом, основываясь на этой идее, был построен классификатор слов на положительно ориентированные слова и отрицательно ориентированные.

Корпусный подход основан на поиске правил и закономерностей в текстах. В работе [8] оценочная характеристика слова вычисляется путем сопоставления совместной встречаемости данного слова со словами *отличный (excellent)* и *плохой (poor)* в данной предметной области. Полученная оценочная направленность слов используется для классификации отзывов на положительные и отрицательные.

В работе [4] выделение оценочных слов и определение их семантической направленности основано на синтаксических шаблонах и союзах между словами. Основное внимание уделяется союзам И, ИЛИ и НО. Предполагается, что, если два прилагательных связаны союзами И или ИЛИ, то они оба являются или не являются оценочными, а так же одинаково семантически направлены. В случае союза НО, семантическое направление различается. Основываясь на этом принципе, был построен классификатор, определяющий семантическую направленность множеств прилагательных, работающий с точностью 92 %.

В работе [7] представлена система OPINE. Система служит для извлечения из отзывов разных атрибутов представленных продуктов, а также оценок по ним. OPINE выделяет следующие атрибуты продукта: свойства продукта, части продукта, атрибуты частей продукта, связанные сущности, свойства и части связанных сущностей. Предполагается, что оценочные фразы появляются в непосредственной близости от атрибутов объекта. Для извлечения оценочных слов используется 10 правил, основанных на синтаксической структуре предложения. Определение семантической ориентации слов базируется на ряде факторов, включая употребление с союзами, учет словообразования, информации о синонимах и антонимах из WordNet.

В работе [2] производился поиск мнений, выраженных придаточными предложениями. Как один из факторов извлечения таких мнений использовалась характеристика относительной частотности слов в документах с предполагаемым большим количеством мнений (колонка редактора и письма читателей) и в документах с меньшим количеством мнений (новости и бизнес-публикации).

Особенностью предлагаемого нами подхода является то, что значительное количество потенциальных характеристик оценочных слов вычисляется на основе сочетания разных текстовых корпусов в рамках заданной предметной области.

3. Получение корпусов и характеристик слов

3.1. Подготовка входных данных

Для подготовки данных с сайта *www.imhonet.ru* были собраны тридцать тысяч отзывов пользователей по различным фильмам. Кроме того, для каждого отзыва была извлечена численная оценка (от одного до десяти) фильма пользователем. Этот корпус является основным для работы, назовем его *корпус мнений*.

Пример отзыва (1): *Неплохой фильм, главное не выключить его в начале, где он напоминает просто ужасную пародию на Адреналин. Ну а в целом в фильме есть, как и положительные (адреналиновые, захватывающие и интересные сцены) так и отрицательные (неоднозначный финал, не везде удачная режиссура) качества.*

Для формирования нейтральной коллекции, где концентрация мнений значительно меньше, с того же сайта были собраны двадцать тысяч описаний фильмов. Назовем этот корпус *корпусом описаний*.

Собранные тексты были обработаны программой морфологического анализа и получен список лемм с информацией о части речи.

Для работы также использовался список лемм, с информацией об их встречаемости в новостном корпусе размером в один миллион документов. Условно этот список назовем *новостным корпусом*.

3.2. Составление корпуса с более высокой концентрацией оценочных слов

Было высказано предположение, что можно выделить некоторые части корпуса мнений, в которых концентрация оценочных слов больше, а именно:

- Предложения, заканчивающиеся на «!»;
- Предложения, заканчивающиеся на «…»;
- Короткие предложения не более чем из 7 слов;
- Короткие отзывы, состоящие из одного предложения;
- Предложения, содержащие слово «фильм» без других существительных.

Условно назовем это корпус — *малый корпус*. Его размер примерно в 2,5 раза меньше чем у *корпуса мнений*.

3.3. Предлагаемые характеристики

Для выделения качественного списка оценочных слов был предложен набор различных характеристик. Для подсчёта были выбраны следующие характеристики:

- Частотность.
- Количество документов, в которых встречается слово.
- Странность.
- TFIDF.
- Отклонение от средней оценки.
- Частотность слов, употребляемых с большой буквы (в корпусе мнений).

Остановимся более подробно на каждой из них.

3.3.1. Частотность

Частотность вычисляется как число появления слова во всем корпусе. Далее все слова упорядочиваются по убыванию частоты встречаемости.

3.3.2. Странность

Для подсчета характеристики странности необходимо два корпуса, один содержащий мнения, другой контрастный. Идея в том, что слова, которые несут оценки, будут «странными» в контексте контрастного корпуса [1], Сама характеристика вычисляется так:

$$\text{Странность} = (FRL/FRC)/(FRLC/FRCC)$$

FRL — частотность леммы в исследуемой коллекции.
FRC — число словоупотреблений во всей исследуемой коллекции.
FRLC — частотность леммы в контрастной коллекции.
FRCC — число словоупотреблений в контрастной коллекции.

Вместо частотности можно использовать количество документов, в котором встретилось слово.

3.3.3. TFIDF

Характеристика TFIDF хорошо известна в информационном поиске. Обычно она вычисляется на основе частотности некоторого слова в отдельном документе и в коллекции в целом. Мы подсчитываем эту характеристику на основе целых корпусов, тем самым также выявляем слова, которые «вдруг» повышают свою относительную частотность относительно другого корпуса.

Существует довольно большое количество способов подсчёта характеристики TFIDF, мы используем формулу из работы [9].

$$\text{TFIDF}(l) = \beta + (1 - \beta) * tf(l) * idf(l)$$

tf(l) — частота леммы l в корпусе с мнениями.
Idf(l) = log((|c|+0,5)/df(l))/log(|c|+1) — фактическая форма штрафования часто используемых в колекции слов.
df(l) — количество документов в контрастной коллекции, где встречалась лемма l.
β = 0,4.
|c| — количество документов в контрастной коллекции.

3.3.4. Отклонение от средней оценки

Как уже упоминалось, для каждого собранного текста мнения, сохранялась еще и числовая оценка (от одного до десяти), поставленная пользователем. Суть данной характеристики состоит в том, чтобы для каждого слова посчитать его среднюю оценку (т, е, взять оценки тех мнений, где оно встретилось, и разделить на количество словоупотреблений) и вычислить модуль разности со средней оценкой всего корпуса. Таким образом, мы получаем суммарную оценочную ориентацию этого слова.

$$dev(l) = \left|\frac{\sum_{i=1}^{n} m_i k_i}{k} - \frac{\sum_{i=1}^{n} m_i}{n}\right|$$

$$\sum_{i=1}^{n} k_i = k$$

l — рассматриваемая лемма.
n — общее количество отзывов.
m_i — оценка i-го отзыва.
k_i — число словоупотреблений леммы в i-ом отзыве (если не употребляется, тогда 0).

3.3.5. Частотность слов употребляемых с большой буквы

Суть этой характеристики в том, что имена собственные обычно не являются оценочными словами. Поэтому мы подсчитываем, сколько раз каждое слово употреблялось с большой буквы и при этом не находилось в начале текста или в начале предложения.

3.4. Комбинации характеристик и корпусов

Для экспериментов были взяты первые десять тысяч слов по частотности, и вся дальнейшая работа проводилась с ними. Слова были разделены на прилагательные и неприлагательные. Смысл такого разделения состоит в том, что многие исследователи указывали, что большинство оценочных слов являются прилагательными, и оценка качества нашего подхода на них, представляет отдельный интерес. В неприлагательные входят: существительные, глаголы и наречия. Все характеристики считались отдельно по этим двум категориям.

Таким образом, получаются такие комбинации характеристик и корпусов:

- TFIDF по парам корпусов: *малый-новости, малый-описания, мнения-новости, мнения-описания*;
- Странность по парам корпусов: *мнения-новости* и *мнения-описания* по количеству документов, *малый-описания* и *мнения-описания* по частотности;
- Отклонение от средней оценки;
- Частота по корпусу мнений и *малому корпусу*;
- Количество документов, в которых встречается слово в *корпусе мнений*;
- Частотность слов употребляемых с большой буквы в *корпусе мнений*;

Кроме этого, отдельно для *корпуса описаний* были посчитаны характеристики: частотность, количество документов, странность *описания-новости* по количеству документов и TFIDF по *корпусам описания-новости*.

Таким образом, для каждой леммы получается 17 признаков.

4. Оценка качества и комбинирование полученных списков

4.1. Метрика оценки качества

Для оценки качества получаемых списков слов мы использовали метрики оценки качества, применяемые для информационного поиска [10], В данной работе будут использоваться три метрики: точность, полнота и F-мера.

4.2. Точность (precision)

Точность вычисляется как отношение количества оценочных слов к общему количеству слов в списке:

$$P=a/(a+b)$$

Здесь a — количество слов, которые являются оценочными; b — количество слов, которые не являются оценочными. Например, если точность равна 50 %, то это значит, что в рассматриваемом списке половина слов — оценочные и половина — неоценочные.

4.3. Полнота (recall)

Полнота вычисляется как отношение найденных оценочных слов к общему количеству оценочных слов:

$$R=a/(a+c)$$

Здесь a — количество слов, которые являются оценочными; c — количество слов, которые являются оценочными, но не найдены. Например, если полнота равна 50 %, то это значит, что половина оценочных слов не найдена.

4.4. F-мера

F-мера часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера вычисляется по формуле:

$$F=2PR/(P+R)$$

4.5. Разметка

Для оценки качества работы алгоритмов необходимо эталонное множество оценочных слов. Изначально была опробована идея построения эталонного множества по общезначимому списку оценочных слов. Были взяты около пятисот слов, но оценки качества, сделанные с помощью этого множества, не соответствовали действительности. Характеристики получались несоответствующие реальному положению дел, поскольку с помощью общезначимого списка нельзя получить слова, которые зависят от предметной области, жаргонные слова и некоторые другие. Поэтому было решено взять исходный десятитысячный список слов и вручную разметить в нем оценочные слова.

При разметке оценочных слов выяснилось, что во многих случаях нельзя сделать однозначный вывод о том, является ли слово оценочным, поскольку иногда слово является оценочным в другой области, но не является оценочным в рабочей области. Многие слова могли употребляться как в оценочном, так и неоценочном смысле при обсуждении фильмов. Поэтому было принято правило, что слово размечалось как оценочное, если можно представить какое-либо оценочное суждение по отношениям к фильмам и их атрибутам. Кроме того, разметка делалась обоими авторами работы.

В результате разметки получился список оценочных слов размером три тысячи двести слов (1262 прилагательных, 296 наречия, 857 существительных, 785 глаголов).

4.6. Результаты по отдельным характеристикам

Приведем результаты, полученные по каждой характеристике при подсчете среди первой тысячи слов:

Таблица 1. Прилагательные

Характеристика	Коллекция	Точность %
TFIDF	малый-новости	60,7
TFIDF	малый-описания	59,4
TFIDF	мнения-новости	60,0
TFIDF	мнения-описания	58,7
Странность	мнения-новости (количество документов)	64,0
Странность	мнения-описания (количество документов)	61,8
Странность	малый-описания (частотность)	60,7
Странность	мнения-описания (частотность)	61,4
Отклонение от оценки		56,3
Частотность	мнения	57,4
Частотность	малый	58,2
Количество документов	мнения	58,7

Таблица 2. Неприлагательные

Характеристика	Коллекция	Точность %
TFIDF	малый-новости	25,9
TFIDF	малый-описания	25,3
TFIDF	мнения-новости	23,2
TFIDF	мнения-описания	21,0
Странность	мнения-новости (Количество документов)	41,7
Странность	мнения-описания (Количество документов)	39,2
Странность	малый-описания (Частотность)	40,5
Странность	мнения-описания (Частотность)	38,2
Отклонение от оценки		30,6
Частотность	мнения	18,4
Частотность	малый	21,4
Количество документов	мнения	19,2

4.7. Машинное обучение

Имея для каждого слова набор характеристик, можно построить классификатор для автоматического разделения слов на оценочные и неоценочные. Для классификации использовалась свободно распространяемая система Rapid Miner [12], В данной работе использовались следующие алгоритмы:

- Метод k ближайших соседей (kNN)
- «Наивный» байесовский классификатор (Naïve Bayes)
- Персептрон (Perceptron)
- Нейронная сеть (2х и 3х-слойная)
- Логистическая регрессия (Logistic Regression)
- Метод опорных векторов (SVM стандартный и с радиальной ядровой функцией)

Оценки качества и подбор параметров алгоритмов производился с помощью кросс-валидации. Кроме того, воспользовавшись байесовским подходом к теории вероятностей, можно получить «вероятность» принадлежности объекта к классу оценочных слов. Если отсортировать слова по значению этой «вероятности», то можно узнать количество оценочных слов в первой тысяче слов списка.

4.8. Результаты классификации

Приведем результаты классифицирования для прилагательных и неприлагательных.

Таблица 3. Прилагательные

Алгоритм	Precision	Recall	F
kNN	63,98	70,75	67,17
Naïve Bayes	73,90	20,69	32,29
Perceptron	59,30	94,34	72,76
Neural Net(2 layers)	65,51	78,95	71,08
Neural Net(3 layers)	66,01	75,29	69,39
Logistic Regression	67,77	68,63	68,09
SVM	63,32	74,72	67,54

Таблица 4. Неприлагательные

Алгоритм	Precision	Recall	F
kNN	34,67	34,50	34,59
Naïve Bayes	28,44	88,39	42,56
Perceptron	53,48	5,88	10,39
Neural Net(2 layers)	38,22	28,32	32,52
Neural Net(3 layers)	55,90	14,90	23,19
Logistic Regression	58,70	9,18	15,84
SVM	37,15	27,00	31,27

В результате классификации: для прилагательных наилучшее значение F-меры получилось равным 72,76 % с использованием персептрона, для неприлагательных соответственно 42,56 % при классификации «наивным» байесовским алгоритмом.

Отдельный интерес представляет вычисление точности для первой тысячи слов, взятых от начала

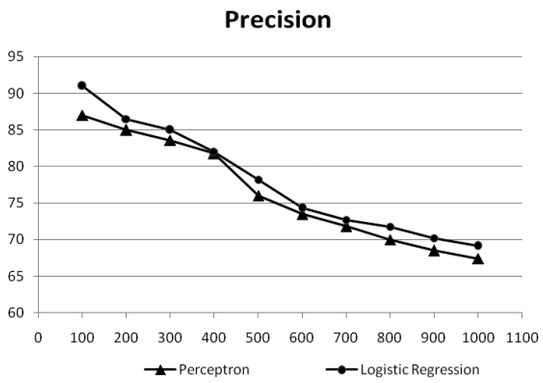

Рис. 1. Зависимость Точности от числа слов взятых сначала списка (прилагательные)

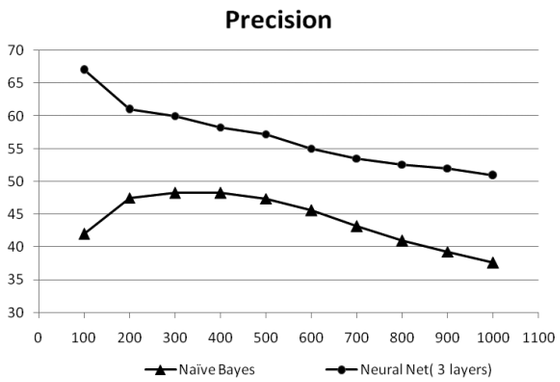

Рис. 2. Зависимость Точности от числа слов взятых сначала списка (неприлагательные)

списка, отсортированного по значению «вероятности» принадлежности объекта к классу оценочных слов. Эти показатели могут быть полезны для дальнейшего автоматического использования классифицированных слов. Наилучшее значение точности получились: для прилагательных при использовании логистической регрессии 69,1 %, для неприлагательных при использовании 3-х слойной нейронной сети 50,9 %, Интересным также показалось изобразить зависимость точности от количества слов взятых от начала. Для этого были взяты лучшие алгоритмы по точности на первой тысяче и лучшие по F-мере. Полученные результаты представлены в виде графиков.

Таким образом, удалось получить рост качества полученных списков на первой тысяче слов (по сравнению со списками по характеристикам) по точности для прилагательных — на 8,28 %, для неприлагательных — на 20,6 %.

В качестве примера приведем первые десять слов из лучших двух списков: по прилагательным и неприлагательным:

позитивный	*пересматривать*
отличный	*простой*
интересный	*тягомотина*
замечательный	*высосанный*
затянутый	*хавать*
смешной	*плоско*
добрый	*наигранно*
обалденный	*фигня*
предсказуемый	*блин*
потрясающий	*отвратительно*

5. Заключение

В работе мы показали, что путем извлечения нескольких корпусов заданной предметной области и вычисления с их помощью нескольких характеристик слов можно автоматически получить достаточно качественные списки оценочных слов. В качестве дальнейшей работы мы планируем добавить число характеристик слов, на основе которых можно улучшить качество выделения оценочных слов, а также предполагается оценить устойчивость предложенной технологии для другой предметной области.

Литература

1. *Ahmad K., Gillam L., Tostevin L.* University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval // In the Proceedings of Eigth Text Retrieval Conference (Trec-8), 1999.
2. *Bethard S., Yu H., Thornton A., Hatzivassiloglou V., Jurafsky D.* Automatic Extraction of Opinion Propositions and their Holders // AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004.
3. *Esuli A., Sebastiani F.* Determining the Semantic Orientation of Terms through Gloss Classification // Conference of Information and Knowledge Management, 2005.
4. *Hatzivassiloglou V., McKeown K.* Predicting the Semantic Orientation of Adjectives // ACL, 1997. P. 174–181
5. *Hu M., Liu B.* Mining and Summarizing Customer Reviews // KDD, 2004.
6. *Pang B., Lee L.* Opinion mining and sentiment analysis // Foundations and Trends® in Information Retrieval, Now Publishers, 2008
7. *Popescu A., Etzioni O.* Extracting Product Features and Opinions from Reviews // EMNLP, 2005.
8. *Turney P. D.* Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // ACL, 2002, 417-424
9. *Агеев М. С., Добров Б. В., Лукашевич Н. В., Сидоров А. В.* Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // РОМИП , 2004.
10. *Агеев М., Кураленок И., Некрестьянов И.* // Петрозаводск: Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2009), 2009.
11. *Ермаков А. Е.* Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии, 2009.
12. *The RapidMiner toolset* // http://rapid-i.com