Модел. и анализ информ. систем. Т. **20**, № **2** (2013) 70–79

УДК 004.853

Построение модели для извлечения оценочной лексики в различных предметных областях

Лукашевич Н.В., Четвёркин И.И.[1]

Московский государственный университет
119991, Москва, ГСП-1, Ленинские горы, д. 1, стр. 4, НИВЦ
119991, Москва, ГСП-1, Ленинские горы, д. 1, стр. 52, ВМК

e-mail: louk@mail.cir.ru, ilia2010@yandex.ru

получена 26 ноября 2012

Ключевые слова: анализ тональности, оценочные слова, адаптация к предметной области

В данной работе предлагается новый подход к извлечению оценочных слов для различных предметных областей. В рамках этого подхода была разработана модель, включающая набор характеристик и комбинацию алгоритмов, которые позволяют извлекать оценочные слова в конкретной предметной области. Данная модель была обучена в предметной области о фильмах и затем применена в четырёх других областях. Качество работы метода оценивалось на основании разметки экспертов и оставалось на высоком уровне при переносе модели на различные предметные области. Кроме того, созданная модель была использована в предметной области о фильмах на английском языке и продемонстрировала высокое качество извлечения оценочных слов.

1. Введение

В связи с бурным развитием Веб 2.0 актуальной является задача анализа тональности отзывов и мнений людей в Интернете. Одной из серьезных проблем при решении данной задачи становится проблема настройки на предметную область. Каждая предметная область может иметь свойственную только ей оценочную лексику, либо значения оценочных слов могут меняться в разных областях [15]. Например, «нужно увидеть» является сильным оценочным выражением в предметной области о фильмах, но нейтральным в предметной области о политике [3].

Для решения проблемы настройки или переноса модели на другие предметные области в литературе предлагаются различные подходы по адаптации алгоритмов, такие как ансамбль классификаторов [2], или подходы на основе графов [18]. Тем не менее, такие подходы плохо работают для предметных областей с существенно

[1]Работа частично поддержана грантом РФФИ N11-07-00588-а.

различающейся лексикой. Поэтому недавние исследования пытаются найти взаимосвязи между специфичными словами в разных предметных областях и построить отображение между ними [13].

В данной работе рассматривается проблема автоматического извлечения словарей оценочных слов для *различных предметных областей*. Такие словари могут быть полезными при адаптации алгоритмов анализа отзывов. Ранее было показано, что словари оценочной лексики, адаптированные под конкретную предметную область, улучшают качество работы в различных задачах, например в поиске оценочной информации [9], или в классификации выражений по тональности [5]. Кроме того, извлечение оценочных слов непосредственно из текстовых коллекций позволяет найти сленг и другие несловарные слова, которые могут быть важными факторами при обработке отзывов [16].

В данном исследовании разработана новая модель для извлечения оценочных слов в заданной предметной области. Предложенная модель обучается на данных из предметной области о фильмах и затем применяется в нескольких других областях.

Дальнейшее изложение статьи будет организовано следующим образом: в разделе 2 будут рассмотрены современные методы формирования словаря оценочной лексики, в разделе 3 описываются текстовые коллекции и характеристики, которые используются в модели. В разделе 4 описывается применение разработанной модели к четырем различным предметным областям. В разделе 5 представлен эксперимент по извлечению оценочных слов из текстов на английском языке.

2. Методы автоматического извлечения оценочных слов

Словари оценочных слов играют важную роль в большинстве задач анализа мнений [6], включающих в себя поиск и извлечение мнений, автоматические ответы на субъективные вопросы, оценочное аннотирование и т.д. Хотя методы машинного обучения с учителем показали свою эффективность в задаче классификации отзывов по тональности [14], авторы в [5] демонстрируют, что включение характеристик на основе оценочной лексики существенно улучшает качество классификации.

Существует два глобальных подхода к автоматическому извлечению оценочной лексики: на основе словарей и на основе текстовых коллекций.

Первый подход основан на информации из различных словарей и тезаурусов. Одним из наиболее распространенных методов является итеративное формирование словаря оценочных выражений на основе тезауруса WordNet или других семантических ресурсов. Основной принцип в данном методе заключается в том, что синонимы и антонимы оценочного слова также являются оценочными. Таким образом, из начального множества слов может быть получено новое, более полное множество оценочных слов [8, 12]. В [7] для конструирования оценочного словаря используются толкования слов. Основная идея заключается в том, что слова с одинаковой оценочной ориентацией имеют схожие толкования.

Второй подход основан на поиске закономерностей, правил и шаблонов в текстовых коллекциях [10, 11]. В [17] авторы тестируют качество оценочного словаря, полученного по огромному массиву текстовой информации, собранному с четырех

миллиардов Веб-страниц. Авторы формируют граф совместной встречаемости слов и используют алгоритм распространения меток на нем. Полученный словарь позволил получить более высокое качество работы в нескольких задачах обработки мнений по сравнению с уже существующими словарями на английском языке. Кроме того, данный ресурс содержит большое количество сленговых, вульгарных и других несловарных слов.

В литературе также встречаются некоторые работы, которые комбинируют вышеописанные два подхода [6].

3. Построение модели для извлечения оценочных слов

В данном разделе будет описано построение и обучение модели извлечения оценочных слов на данных из предметной области о фильмах.

3.1. Коллекции данных

Модель извлечения оценочных слов базируется на нескольких текстовых коллекциях.

Во-первых, для построения модели было собрано 28 773 отзыва о фильмах различного жанра с рекомендательного портала *www.imhonet.ru*. Каждому отзыву соответствовала оценка автора по десятибалльной шкале. Эта коллекция является основной для работы, назовем её *коллекцией мнений*.

Пример отзыва: *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.*

Во-вторых, для формирования нейтральной коллекции, где концентрация мнений значительно меньше, с того же сайта были собраны 17 680 описаний фильмов. Назовем эту коллекцию — *коллекцией описаний*.

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

Кроме того, было высказано предположение, что можно выделить некоторые части корпуса мнений, в которых концентрация оценочных слов больше, а именно: предложения, заканчивающиеся на «!» или «...»; короткие предложения не более чем из 7 слов; короткие отзывы, состоящие из одного предложения; предложения, содержащие слово «фильм» без других существительных.

Условно назовем этот корпус – *малый корпус*.

3.2. Статистические признаки

Чтобы построить модель, которая качественно отличает оценочные слова от неоценочных, для каждой леммы из корпуса мнений вычисляется набор статистических признаков:

- **Частотные характеристики**: частота слова во всей коллекции и подокументная частота; частота слов с большой буквы; признак «странности»; признак TFIDF.

- **Характеристики на основе оценки пользователя**: отклонение от средней оценки; дисперсия оценки слова; вероятность встретить заданное слово с каждой из оценок.

Рассмотрим некоторые характеристики более подробно.

Частота слов с большой буквы. Суть этой характеристики в том, что имена собственные обычно не являются оценочными словами. Поэтому для каждого слова вычисляется, сколько раз оно употреблялось с большой буквы и при этом не находилось в начале текста или в начале предложения.

Странность. Для подсчета странности необходимо два корпуса, один — с высокой концентрацией оценочных слов, другой — контрастный. Идея в том, что слова, которые несут оценки, будут «странными» в контексте контрастного корпуса. Сама характеристика вычисляется так [1]:

$$Weirdness = \frac{P_s(w)}{P_g(w)},$$

где $P_s(w)$ — вероятность слова в исследуемой коллекции; $P_g(w)$ — вероятность слова в контрастной коллекции. Для вычисления оценок вероятностей вместо частотности можно использовать количество документов, в котором встретилось слово.

Характеристика странности вычислялась на базе следующих пар коллекций: *мнения-новости, мнения-описания, описания-новости* с документной частотой и *малый-описания, мнения-описания* с частотой по всей коллекции.

TFIDF. Вес TFIDF является известным в информационном поиске методом взвешивания слов. В данной работе использовался вариант подсчета TFIDF, который предложен в [4] (на основе BM25).

$$TFIDF = \beta + (1 - \beta) \cdot tf \cdot idf$$

$$tf_D(l) = \frac{f(w)}{f(w) + 2} \quad idf(l) = \frac{log\left(\frac{|c| + 0.5}{df(w)}\right)}{log(|c| + 1)},$$

где $f(w)$ — частота леммы w в коллекции, $df(w)$ — количество документов в коллекции (описаний или новостей), где встречалась лемма w, $\beta = 0.4$, $|c|$ — количество документов в коллекции.

Признак TFIDF вычислялся на базе следующих пар коллекций: *малый-новости, малый-описания, мнения-новости, мнения-описания* и *описания-новости.*

Как уже было описано, каждому отзыву соответствует численная оценка по 10-балльной шкале. Пусть $C = 1 \ldots 10$ будет множество возможных оценок в коллекции отзывов.

Определение 1. Условная вероятность каждой категории в зависимости от слова:

$$P(c|w) = \frac{f(w, c)}{\sum_{c_i \in C} f(w, c_i)}.$$

Определение 2. Условная вероятность для каждого слова в зависимости от категории:

$$P(w|c) = \frac{f(w,c)}{\sum_{w_i \in c} f(w_i, c)}.$$

Определение 3. Условное математическое ожидание для каждой категории в зависимости от слова:

$$E(c|w) = \sum_{c_i \in C} c_i \cdot P(c_i|w).$$

Определение 4. Математическое ожидание каждой категории в коллекции отзывов:

$$E(c) = \sum_{c_i \in C} c_i \cdot P(c_i).$$

Используя данные определения, введем несколько характеристик на их основе.

Отклонение от средней оценки.

$$Dev(w) = |E(c|w) - E(c)|.$$

Данный признак позволяет выделять слова, которые употребляются в широком спектре оценочных категорий. Как следствие, вероятность принадлежности таких слов к оценочным ниже, чем у слов с более детерминированным поведением.

Дисперсия оценки слова. Еще одной важной характеристикой является дисперсия оценки слова. Если у оценки слова маленькая дисперсия, это значит, что данное слово употребляется в отзывах с близкими оценками. Такие слова более вероятно являются оценочными.

$$Var(w) = E(c^2|w) - E(c|w)^2.$$

Вероятность встретить заданное слово с каждой из категорий. Чтобы формализовать информацию о встречаемости слов в различных категориях, вводится логарифм нормированной условной вероятности для каждого слова, в зависимости от категории

$$Lhc = \log \frac{P(w|c)}{P(w)}.$$

Нормировка необходима для сравнения значений данной функции у различных слов. Также были добавлены несколько характеристик на основе Lhc, такие как её максимум или среднее.

3.3. Лингвистические характеристики

В модель также был добавлен набор лингвистических признаков, так как они могут играть важную роль в улучшении качества извлечения оценочных слов.

– Четыре бинарных признака частей речи (существительное, глагол, прилагательное и наречие);

– Два бинарных признака, отражающих неоднозначность употребления леммы в разных частях речи (т.е. лемма может быть разными частями речи, в зависимости от контекста) и нахождение данной леммы в словаре морфологического анализатора;

– Нахождение в слове заранее заданного списка приставок. Эта характеристика является важным индикатором слов, начинающихся с отрицания (например, *несмешной*);

Для русского и английского языка использовался один и тот же морфологический анализатор *Cirmorph*[2].

4. Алгоритмы и оценка качества

Для обучения алгоритмов нам необходимо было размеченное множество слов. Чтобы его получить, все слова из коллекции отзывов о фильмах с частотой выше трёх (18362 слова) были размечены вручную. Слово считалось оценочным, если можно представить какой-либо оценочный контекст с его участием в предметной области о фильмах. Каждое слово было размечено двумя экспертами. В результате данной процедуры было получено множество из 4079 оценочных слов.

С помощью размеченных данных были обучены классификаторы, разделяющие все слова на два класса: оценочные и неоценочные. Для обучения алгоритмов использовался программный пакет Weka[3]. Были использованы три различных алгоритма для классификации: Logistic Regression, LogitBoost и Random Forest. Для всех экспериментов применялась кросс-валидация на 10 частей.

Эти алгоритмы применялись для формирования списков слов, упорядоченных по вероятности принадлежности каждого слова к классу оценочных. Для оценки качества извлеченных оценочных слов использовалась метрика *Precision@n*. Эта метрика хорошо подходит для оценки качества комбинаций списков, а также может быть использована с различными порогами. Для сравнения качества работы алгоритмов в различных предметных областях был выбран порог n = 1000. Этот порог не слишком велик для ручной разметки и достаточен для демонстрации качества работы модели. Результаты классификации в предметной области о фильмах можно найти в Таблице 1.

Таблица 1. Precision@1000 в предметной области о фильмах

Logistic Regression	**LogitBoost**	**Random Forest**	**Average**
75.7%	75.3%	72.4%	**81.5%**

В процессе работы было замечено, что списки оценочных слов, извлеченные разными алгоритмами, существенно отличаются. Поэтому было произведено усреднение значений вероятностей для каждого слова из трёх списков. Результат усреднения можно найти в последней колонке в Таблице 1. Данный мета-алгоритм будет использоваться и оцениваться в других предметных областях. Таким образом, качество извлечения оценочных слов для предметной области о фильмах составляет *Precision@1000* = 81.5%.

Для сравнения были вычислены метрики *Precision@1000* по отдельным характеристикам: частоте и отклонению от средней оценки. Так, *Precision@1000* по признаку частоты составила 26.9%, а по признаку отклонения от средней оценки — 35.5%.

[2]http://ru-eval.ru/participants.html#Cirmorph
[3]http://www.cs.waikato.ac.nz/ml/weka/

Таблица 2. Precision@1000 для различных признаков

Характеристика	**Коллекции**	**Precision@1000**
TFIDF	малый–новости	38.5%
TFIDF	малый–описаний	36.4%
TFIDF	мнений–новости	30.5%
TFIDF	мнений–описаний	39.8%
Странность	мнений–новостей (док. частота)	31.7%
Странность	мнений–описаний (док. частота)	48.1%
Странность	малый–описаний (частота)	49.1%
Странность	мнений–описаний (частота)	46.6%
Dev	мнений	35.5%
Var	мнений	21.5%
Lhc	мнений	33.0%
Частота	мнений	26.9%
Частота	малый	31.9%
Док. Частота	мнений	27.8%

Таким образом, алгоритм извлечения оценочных слов дает существенный прирост качества по сравнению с простыми характеристиками. Результаты классификации на основе других признаков представлены в Таблице 2. Приведем несколько примеров оценочных слов с высокими значениями вероятности в результирующем списке: *трогательный, отстой, фигня, отвратительно, посредственный, предсказуемый, любимый* и др.

5. Применение модели к другим областям

В предыдущем разделе был описан подход к построению модели для извлечения оценочных слов в предметной области о фильмах. Следующим этапом данного исследования является применение обученной модели к четырем другим предметным областям и оценка качества извлечения оценочных слов в этих предметных областях.

Для проведения экспериментов по переносу модели были собраны[4] данные в четырёх дополнительных предметных областях. Структура коллекций данных такая же, как и в предметной области о фильмах. Характеристики коллекций можно найти в Таблице 3. При переносе модели использовался тот же новостной корпус, что и раньше.

Для всех слов из заданной предметной области (за исключением низкочастотных) вычисляются векторы признаков и из них строится матрица «слова-характеристики». К данной матрице применяется обученная модель классификации и вручную оценивается первая тысяча наиболее вероятных оценочных слов. Результаты работы алгоритма для разных областей можно найти в Таблице 4.

Таким образом, качество работы алгоритма в каждой из областей существенно превышает значения по индивидуальным признакам (см. Таблицу 2). Некоторое

[4]Отзывы о книгах и цифровых камерах были получены от семинара РОМИП (www.romip.ru)

Таблица 3. Характеристики коллекций данных

	Коллекция отзывов	**Коллекция описаний**	**Источник**
Книги	23 883	22 321	Имхонет
Игры	7 928	1 853	Имхонет
Цифровые камеры	10 208	920	Яндекс Маркет
Мобильные телефоны	30 620	890	Яндекс Маркет

Таблица 4. Результаты переноса модели

Книги	**Игры**	**Цифровые камеры**	**Мобильные телефоны**
86.0%	72.2 %	62.0%	73.2%

снижение качества классификации связано с недостаточным размером коллекций описаний и обилием специфичной лексики (особенно в отзывах о фотокамерах, например «апертура»), которой алгоритм склонен давать высокие веса. Для решения данной проблемы необходимо формирование более полных корпусов описаний сущностей, которые помогут снизить влияние различных специфических терминов в каждой из областей.

6. Применение модели к другим языкам

После успешного переноса модели на другие предметные области была высказана гипотеза, что после стадии морфологической обработки предложенный метод является независимым от конкретного языка. Для подтверждения этой гипотезы был проведен эксперимент по переносу модели с русского языка на отзывы о фильмах на английском языке. В данном эксперименте использовалась коллекция отзывов о фильмах из [3]. Коллекция описания фильмов была получена с популярного портала IMDb[5], а в качестве новостной коллекции использовалась коллекция Reuters-21578[6]. Таким образом, всего было 34 180 отзывов о фильмах, 40 000 описаний фильмов и документная частотность, вычисленная по коллекции Reuters-21578.

На основе данных коллекций была сформирована матрица «слова-характеристики», следуя стандартной процедуре, описанной ранее. Далее, к этой матрице применялась модель извлечения оценочных слов, обученная на русскоязычных отзывах о фильмах. Качество извлеченных английских слов составило **70.5%** в соответствии с Precision@1000.

Наиболее вероятными оценочными словами были следующие: *remarkable, overdo, understated, respected, overlook, lame* и др. Некоторые из этих слов (например *overlook*) являются оценочными только в предметной области о фильмах.

[5]Information courtesy of The Internet Movie Database (http://www.imdb.com). Used with permission.

[6]http://www.daviddlewis.com/resources/testcollections/reuters21578

7. Заключение

В данной работе был предложен новый метод извлечения оценочных слов для конкретной предметной области на основе нескольких текстовых коллекций. Разработанный алгоритм был применен к различным предметным областям и продемонстрировал хорошую обобщающую способность. Кроме того, было показано, что предложенный подход может быть использован для извлечения оценочных слов на других языках.

Список литературы

1. Ahmad K., Gillam L., Tostevin L. University of Surrey participation in Trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder) // The Eighth Text REtrieval Conference (TREC-8), 1999.

2. Aue A. and Gamon M. Customizing sentiment classifiers to new domains: A case study // Proceedings of recent advances in natural language processing (RANLP). 2005.

3. Blitzer J., Dredze M., Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification // Annual Meeting-Association For Computational Linguistics. 2007. V. 45. P. 440–447.

4. Callan J.P. Croft W.B., Harding S.M. The INQUERY retrieval system // Proceedings of the third international conference on database and expert systems applications. 1992.

5. Choi Y., Cardie C. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009. V. 2. P. 590–598.

6. Ding X., Liu B., Yu P.S. A holistic lexicon-based approach to opinion mining // Proceedings of the international conference on Web search and web data mining. 2008. P. 231–240.

7. Esuli A., Sebastiani F. Determining the semantic orientation of terms through gloss classification // Proceedings of the 14th ACM international conference on Information and knowledge management. 2005. P. 617–624.

8. Hu M., Liu B. Mining and summarizing customer reviews // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004. P. 168–177.

9. Jijkoun V., de Rijke M., Weerkamp W. Generating focused topic-specific sentiment lexicons // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010. P. 585–594.

10. Kanayama H., Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. 2006. P. 355–363.

11. Lu Y., Castellanos M., Dayal U., Zhai C.X. Automatic construction of a context-aware sentiment lexicon: an optimization approach // Proceedings of the 20th international conference on World wide web. 2011. P. 347–356.

12. Neviarouskaya A., Prendinger H., Ishizuka, M. Sentiful: Generating a reliable lexicon for sentiment analysis // Affective Computing and Intelligent Interaction and Workshops. 2009. P. 1–6.

13. Pan S.J., Ni X., Sun J.T., Yang Q., Chen Z. Cross-domain sentiment classification via spectral feature alignment // Proceedings of the 19th international conference on World wide web. 2010. P. 751–760.

14. Pang B., Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. Hanover, Massachusetts: Now Publishers, 2008.

15. Ponomareva N., Thelwall M. Biographies or blenders: which resource is best for cross-domain sentiment analysis? // Computational Linguistics and Intelligent Text Processing. 2012. P. 488–499.

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

17. Velikovich L., Blair-Goldensohn S., Hannan K., McDonald R. The viability of web-derived polarity lexicons // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010. P. 777–785.

18. Wu Q., Tan S., Cheng X. Graph ranking for sentiment transfer // Proceedings of the ACL-IJCNLP. 2009. P. 317–320

Construction of a Model for the Cross-Domain Opinion Word Extraction

Loukachevitch N.V., Chetviorkin I.I.

Lomonosov Moscow State University
Leninskiye Gory, 1, Build. 4, Research Computing Center, Moscow, GSP-1, 119991, Russia
Leninskiye Gory 1, Build. 52, Faculty of Computational Mathematics and Cybernetics, Moscow, GSP-1, 119991, Russia

Keywords: sentiment analysis, opinion words, domain adaptation

In this paper we consider a new approach for domain-specific opinion word extraction in the Russian language. We propose a set of statistical features and an algorithm combination that can extract opinion words in a particular domain. The extraction model was trained in the movie domain and then applied to four other domains. The quality of the obtained sentiment lexicons was evaluated intrinsically on the base of an expert markup and remained on the high level during the model transfer to various domains. Finally, our method is adapted to the movie domain in English and it demonstrated good results.

Сведения об авторах:
Лукашевич Наталья Валентиновна, МГУ, Научно-исследовательский вычислительный центр, ведущий научный сотрудник;
Четвёркин Илья Игоревич, МГУ, Факультет вычислительной математики и кибернетики, аспирант.