DomEx: Extraction of Sentiment Lexicons for Domains and Meta-Domains

Ilia Chetviorkin[1] *Natalia Loukachevitch*[2]
(1) Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University,
Moscow, Leninskiye Gory 1, Building 52
(2) Research Computing Center,
Lomonosov Moscow State University,
Moscow, Leninskiye Gory 1, Building 4
ilia.chetviorkin@gmail.com, louk_nat@mail.ru

ABSTRACT

In this paper we describe a DomEx sentiment lexicon extractor, where a new approach for domain-specific sentiment lexicon extraction is implemented. Sentiment lexicon extraction is based on the machine learning model comprising a set of statistical and linguistic features. The extraction model is trained in the movie domain and then can be utilized to other domains. The system can work with various domains and languages after part of speech tagging. Finally, the system gives possibility to combine the sentiment lexicons from similar domains to obtain one general lexicon for the corresponding meta-domain.

TITLE AND ABSTRACT IN RUSSIAN

DomEx: Извлечение Оценочной Лексики для Различных Предметных Областей и Мета-Областей

В данной работе мы описываем систему для извлечения оценочных слов DomEx, в которой реализован новый подход для формирования оценочного словаря. Извлечение оценочной лексики основано на машинном обучении с использованием набора статистических и лингвистических признаков. Модель для извлечения обучается в предметной области о фильмах и затем может быть использована в других предметных областях. Система может работать с различными предметными областями и языками после этапа морфологической обработки. Наконец, система дает возможность комбинировать списки оценочных слов из похожих предметных областей для формирования одного, общего словаря для соответствующей мета-области.

KEYWORDS : Sentiment Analysis, Sentiment Lexicon, Domain Adaptation
KEYWORDS IN RUSSIAN: Анализ Тональности, Оценочные слова, Адаптация к Предметной Области

Proceedings of COLING 2012: Demonstration Papers, pages 77–86,
COLING 2012, Mumbai, December 2012.

В последнее время большие усилия были направлены на решение задачи анализа мнений в различных предметных областях. Автоматизированные подходы к анализу тональности могут быть полезны для государственных органов и политиков, компаний и простых пользователей. Одной из важнейших задач, являющейся основой для анализа мнений в текстах, написанных на различных языках, является создание словарей оценочных слов.

В данной демонстрационной работе мы представляем ***DomEx***, систему по извлечению оценочных слов, которая использует обученную модель для извлечения оценочных слов в различных предметных областях и на различных языках, а также позволяет пользователям создавать общий словарь оценочной лексики для группы похожих областей.

Работа системы основана на нескольких текстовых коллекциях: коллекции отзывов о продуктах с оценками пользователей, коллекции описаний продуктов и контрастной коллекции (например, новостная коллекция). Такие коллекции могут быть автоматически сформированы для разных предметных областей. Кроме того, мы предположили, что можно выделить некоторые части корпуса мнений (например, о фильмах), в которых концентрация оценочных слов выше: предложения, заканчивающиеся на «!» или «…»; короткие предложения не более чем из 7 слов; предложения, содержащие слово «фильм» без других существительных. Условно назовем это корпус – малый корпус.

Для каждого слова в коллекции отзывов мы вычисляем набор лингвистических и статистических признаков:

- Частотные: частотность в коллекции (т.е. сколько раз слово встретилось во всей коллекции); документная частотность; частотность слов с большой буквы; «странность» (Ahmad et al., 2009); TFIDF.
- На основании оценки пользователя: отклонение от средней оценки; дисперсия оценки слова; вероятность встретить заданное слово с каждой из оценок.

Также был добавлен набор из лингвистических признаков, так как они играют важную роль в улучшении качества извлечения оценочных слов:

- Четыре бинарных признака для частей речи (существительное, прилагательное, глагол и наречие).
- Два бинарных признака, первый, отражающий неоднозначность части речи (т.е. слово может употребляться в разных частях речи, в зависимости от контекста), и второй, отражающий присутствие слова в словаре морфологического анализатора.
- Заранее заданный список приставок (например приставки “*не*”, “*бес*”, “*без*” и т.д.). Этот признак являтеся важным индикатором оценочных слов, начинающихся с отрицания.

Для обучения алгоритмов нам необходимо было размеченное множество слов. Для этого мы вручную разметили множество всех слов с частотой выше трех из предметной области о фильмах (18362 слова). Мы относили слово к категории

оценочных в случае если могли представить его в каком-либо оценочном контексте.

Мы решали задачу классификации на два класса: разделение всех слов на оценочные и неоценочные. Для этих целей использовались следующие алгоритмы: *Logistic Regression*, *LogitBoost* и *Random Forest*. Все параметры алгоритмов были выставлены в соответствии с их значениями по умолчанию.

Используя данные алгоритмы, мы получили списки слов, упорядоченные по вероятности оценочности слов. Для оценки качества этих списков использовалась мера *Precision@n*. Для сравнения качества работы системы в разных предметных областях мы использовали значение n = 1000.

Мы заметили, что извлеченные списки оценочных слов существенно различаются в зависимости от алгоритма. Поэтому мы решили вычислить среднее от значений вероятностей в каждом из списков. В результате качество автоматического извлечения оценочных слов в области фильмов *Precision@1000* составило 81.5%.

Для использования системы в новой предметной области необходимо собрать аналогичный набор коллекций, как и предметной области о фильмах. Наши эксперименты в адаптации модели к другим предметным областям (книги, компьютерные игры) описаны в (Chetviorkin & Loukachevitch, 2011). Оценка качества переноса показала, что модель достаточно устойчива для использования в других областях.

Для использования нашей системы с другими языками нужно сделать несколько небольших изменений:

1. Все входные данные должны быть обработаны морфологическим анализатором для соответствующего языка. Все соответствующие тэги должны быть изменены в системе.
2. Необходимо изменить ключевое слово для извлечения потенциальных оценочных предложений при составлении малого корпуса.
3. Необходимо изменить список приставок в соответствии с обрабатываемым языком.

После таких изменений система без какого-либо дополнительного обучения может быть использована для обработки других языков.

Мы применили нашу систему для предметной области о фильмах на английском языке. Для этого использовались отзывы и описания с IMDb и новостная коллекция Reuters-21578. Используя эти коллекции, DomEx формирует вектора признаков для каждого слова и применяет модель, обученную на русскоязычных отзывах о фильмах. Качество работы системы после оценки составило 70.5% в соответствии с метрикой P@1000. Наиболее вероятными извлеченными английскими словами являются: *remarkable, recommended, overdo, understated, respected, overlook, lame*, и др. Некоторые из этих слов (например, *overlook*) являются оценочными словами только в предметной области о фильмах.

1 Introduction

Over the last few years a lot of efforts were made to solve sentiment analysis tasks in different domains. Automated approaches to sentiment analysis can be useful for state bodies and politicians, companies, and ordinary users. One of the important tasks, considered as a basis for sentiment analysis of documents written in a specific language, is a creation of its sentiment lexicon (Abdul-Mageed et al., 2011; Peres-Rosas et al., 2012).

Usually authors try to gather general sentiment lexicons for their languages (Mihalcea et al., 2007; Banea et al., 2008; Clematide & Klenner, 2010; Steinberger et al., 2011). However a lot of researchers stress the differences between sentiment lexicons in specific domains. For example, "must-see" is a strongly opinionated word in the movie domain, but neutral in the digital camera domain (Blitzer et al., 2007). For these reasons, supervised learning algorithms trained in one domain and applied to other domains demonstrate considerable decrease in the performance (Ponomareva & Thelwall, 2012; Read & Carroll, 2009; Taboada et al., 2011).

In many studies domain-specific sentiment lexicons are created using various types of propagation from a seed set of words, usually a general sentiment lexicon (Kanayama & Nasukawa, 2007; Lau et al., 2011; Qiu et al., 2011). In such approaches an important problem is to determine an appropriate seed lexicon for propagation, which can heavily influence the quality of the results. Besides, the propagation often lead to unclear for a human sentiment lists. So, for example, in (Lau et al., 2011) only 100 first obtained sentiment words are evaluated by experts, *precision@100* was around 80%, what means that the intrinsic quality of the extracted 4000 lexicon (as announced in the paper) can be quite low.

The sentiment lexicon extraction system presented in this demo exploits a set of statistical and linguistic measures, which can characterize domain-specific sentiment words from different sides. We combined these features into a single model using machine learning methods and trained it in the movie domain. We argue that this model incorporated into our system can be effectively transferred to other domains for extraction of their sentiment lexicons.

Stressing the differences in sentiment lexicons between domains, one should understand that domains can form clusters of similar domains. So a lot of sentiment words relevant to various product domains are not relevant to the political domain or the general news domain and vice versa. For example, such words as *evil* or *villain* are not applicable to all product domains. Therefore we suppose that gathering a specialized sentiment lexicon for meta-domains comprising several similar domains can be useful for researchers and practitioners.

In this demo paper we present ***DomEx*** sentiment lexicon extractor, which utilizes the trained extraction model to different domains and different languages and allows users to create a joint sentiment lexicon for a group of similar domains.

2 Training Model for Extraction of Sentiment Lexicon in a Specific Domain

Training of the sentiment lexicon model is based on several text collections, which can be automatically formed for many domains, such as: a collection of product reviews with authors' evaluation scores, a text collection of product descriptions and a contrast corpus (for example, a general news collection). For each word in the review collection we calculate a set of linguistic and statistical features using the aforementioned collections and then apply machine learning algorithms for term classification.

Our method does not require any seed words, and is rather language-independent, however, lemmatization (or stemming) and part-of speech tagging are desirable. Working with Russian language, we use a dictionary-based morphological processor, including unknown word processing. Below in the text we will say only about lemmatized words.

The basic model is constructed for the movie domain. We collected 28, 773 movie reviews of various genres from the online recommendation service *www.imhonet.ru*. For each review, user's score on a ten-point scale was extracted. We called this collection the **review collection**. We also required a contrast collection of texts for our experiments. In this collection the concentration of opinions should be as little as possible. For this purpose, we collected 17, 680 movie descriptions. This collection was named the **description collection**. One more contrast corpus was a collection of two million news documents. We had calculated a document frequency of each word in this collection and used only this frequency list further. This list was named the **news corpus**.

We also suggested that it was possible to extract some fragments of reviews from the review collection, which had higher concentration of sentiment words. These fragments may include: sentences ending with a "!"; sentences ending with a "…"; short sentences (no more than seven word length); sentences containing the word «movie» without any other nouns. We called this collection the **small collection**.

Our aim is to create a high quality list of sentiment words based on the combination of various discriminative features. We utilize the following set of features for each word:

- Frequency-based: collection frequency $f(w)$ (i.e. number of occurrences in all documents in the collection); document frequency; frequency of capitalized words; weirdness (Ahmad et al., 2009); TFIDF.
- Rating-based: deviation from the average score; word score variance; sentiment category likelihood for each *(word, category)* pair.

Some linguistic features were also added to our system because they can play crucial role in improving the sentiment lexicon extraction.

- Four binary features indicating word part of speech (noun, verb, adjective and adverb).

- Two binary features reflecting POS ambiguity (i.e. word can have various parts of speech depending on a context) and feature indicationg if this word is recognized by the POS tagger.
- Predefined list of prefixes of a word (for example, Russian prefixes "*ne*", "*bes*", "*bez*" etc. similar to English "*un*", "*in*", "*im*" etc.). This feature is a strong predictor for words starting with negation.

To train supervised machine learning algorithms we needed a set of labeled sentiment words. For our experiments we manually labeled words with the frequency greater than three in the movie review collection (18362 words). We marked up a word as a sentiment one in case we could imagine it in any opinion context in the movie domain.

We solved the two class classification problem: to separate all words into sentiment and neutral categories. For this purpose Weka[1] data mining tool was used. We considered the following algorithms: *Logistic Regression*, *LogitBoost* and *Random Forest*. All parameters in the algorithms were set to their default values.

Using this algorithms we obtained word lists, ordered by the predicted probability of their opinion orientation. To measure the quality of these lists the *Precision@n* metric was used. This metric was very convenient for measuring the quality of list combinations and it could be used with different thresholds. To compare quality of the algorithms in different domains we chose n = 1000. This level was not too large for the manual labeling and demonstrated the quality in an appropriate way. We noticed that the lists of sentiment words extracted by the algorithms differ significantly. So we decided to average word probability values in these three lists. Combining three classifiers we obtained Precision@1000 = 81.5%

As the baseline for our experiments we used the lists ordered by frequency in the review collection and deviation from the average score. *Precision@1000* in these lists was 26.9% and 35.5% accordingly. Thus our algorithms gave significant improvements over the baselines.

3 Model Adaptation. Meta-Domain Sentiment Lexicon

To adapt the model to a new domain it is necessary to collect similar data as for the movie domain. Our experiments in adaptation of the model to other domains (books, computer games) are described in (Chetviorkin & Loukachevitch. 2011). For all words in a particular field (excluding low frequent ones) we compute feature vectors and construct a domain word-feature matrix using them. We applied our classification model, which was trained in the movie domain, to these word-feature matrixes and manually evaluated the first thousand of the most probable sentiment words in each domain. The results of the evaluation showed that the sentiment lexicon extraction model is robust enough to be transferred to other domains.

Many domains can form groups with similar lexicons. So many similar sentiment words can be applied to various products. Therefore it is useful to generate sentiment lexicon for such a joint domain – meta-domain.

[1] http://www.cs.waikato.ac.nz/ml/weka/

Constructing the general sentiment lexicon from several extracted domain-specific lexicons we want to boost words that occur in many different domains and have high weights in each of them. We propose the following function for the word weight in the resulting list:

$$R(w) = \max_{d \in D}(prob_d(w)) \cdot \sum_{d \in D} \frac{1}{|D|} \cdot \left(1 - \frac{pos_d(w)}{|d|}\right)$$

where D – is the domain set with five domains, d is the sentiment word list for a particular domain and $|d|$ is the total number of words in this list. Functions $prob_d(w)$ and $pos_d(w)$ are the sentiment probability and position of the word in the list d.

The meta-domain list of sentiment words created in such a way consists of words really used in users' reviews and its creation does not require any dictionary resources.

4 System Functionality and Architecture

Thus ***DomEx*** extractor has the following functionality:

- Extraction of domain-specific sentiment lexicon (see Figure 1).
- Construction of a joint lexicon for several similar domains.
- Application of the model to other languages.

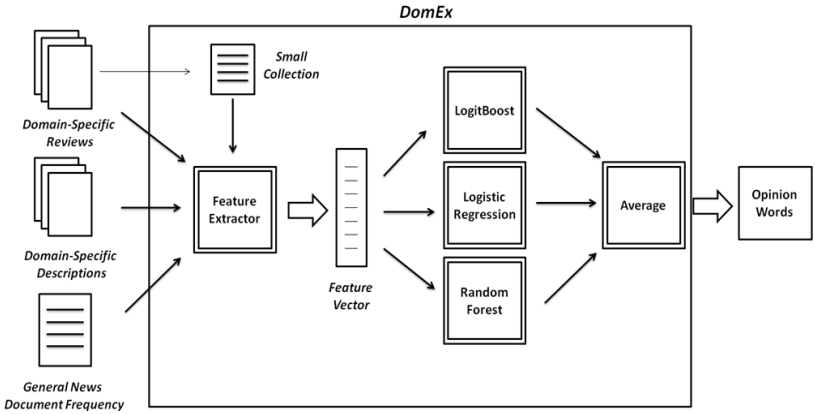

FIGURE 1 –System Overview. Double boxed items are system components and single boxed items are text files

To utilize our system for another language some minor changes should be made:

1. All input data collections should be pre-processed with corresponding POS tagger and change appropriate tags in the system.

2. Changing the key word for extracting potential opinion-bearing sentences during the construction of the small collection (see Section 2).

3. The list of sentiment-bearing prefixes should be specified for a specific language. These prefixes should indicate potential opinion words, for example "*un*", "*in*", "*im*" in English.

After such changes our system without any additional learning can be transformed to process texts in other languages. The most difficult part is to collect appropriate amount of reviews and descriptions of entities in the specific domain and language.

As an example, we utilized our system for English language in the movie domain. We use the review dataset from (Blitzer et al., 2007), but take only reviews from the movie domain. As contrast collections we used plot dataset freely available on the IMDb[2] and Reuters-21578[3] news collection. Using these datasets DomEx computed the word-feature matrix following the previously described procedure and applied our model trained on the Russian movie reviews. The evaluated quality of obtained lexicon was 70.5% according to P@1000 measure.

The most probable extracted English sentiment words in the movie domain were as follows: *remarkable, recommended, overdo, understated, respected, overlook, lame*, etc. Some of these words (for example *overlook*) are opinion words only in the movie domain.

Conclusion and Perspectives

In this paper we presented DomEx sentiment lexicon extractor, in which a new approach for domain-specific sentiment lexicon extraction is implemented. Sentiment lexicon extraction is based on the machine learning model comprising a set of statistical and linguistic features. The extraction model is trained in the movie domain and then can be utilized to other domains. The experiments showed that the model can be transferred to other domains and had good generalization abilities. The system can work with various domains and languages after a part of speech tagging.

Finally, the system gives possibility to combine the sentiment lexicons from similar domains to obtain one general lexicon for the corresponding meta-domain.

Acknowledgments

This work is partially supported by RFBR grant N11-07-00588-a.

References

Abdul-Mageed M., Diab M., Korayem M. (2011). Subjectivity and Sentiment Analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, number 3, pp. 587-591.

Ahmad K., Gillam L., Tostevin L. (1999). University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval *In the Proceedings of Eigth Text Retrieval Conference (Trec-8).*

[2] Information courtesy of The Internet Movie Database (http://www.imdb.com). Used with permission.
[3] http://www.daviddlewis.com/resources/testcollections/reuters21578

Banea C., Mihalcea R., Wiebe J. and Hassan S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Blitzer J., Dredze M., Pereira F. (2007) Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*, pp. 440–447.

Chetviorkin I. and Loukachevitch N. (2011). Extraction of Domain-specific Opinion Words for Similar Domains. In *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition held in conjunction with RANLP 2011*, pp. 7–12.

Clematide S., Klenner S. (2010) Evaluation and extension of a polarity lexicon for German. In *WASSA-workshop held in conjunction with ECAI-2010*, pp 7-13.

Kanayama H. and Nasukawa T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *EMNLP '06*, pp. 355–363, Morristown, NJ, USA.

Lau R., Lai C., Bruza P. and Wong K. (2011). Pseudo Labeling for Scalable Semi-supervised Learning of Domain-specific Sentiment Lexicons. In *20th ACM Conference on Information and Knowledge Management*.

Mihalcea R., Banea C. and Wiebe J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 976–983, Prague, Czech Republic.

Perez-Rosas V., Banea C. and Mihalcea R. (2012). Learning Sentiment Lexicons in Spanish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Ponomareva N. and Thelwall M. (2012): Bibliographies or blenders: Which resource is best for cross-domain sentiment analysis? In *Proceedings of the 13th Conference on Intelligent Text Processing and Computational Linguistics*.

Qiu G., Liu B., Bu J. and Chen C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1).

Read J., Carroll J. (2009). Weakly Supervised techniques for domain independent sentiment classification. In *Proceedings of the first International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, pp. 45-52.

Steinberger J., Lenkova P., Ebrahim M., Ehrmann M., Hurriyetogly A., Kabadjov M., Steinberger R., Tanev H., Zavarella V. and Vazquez S. (2011). Creating Sentiment Dictionaries via Triangulation. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011*, pp. 28–36,

Taboada M., Brooke J., Tofiloski M., Voll K. and Stede M. (2011). Lexicon-based methods for Sentiment Analysis. *Computational linguistics*, 37(2).