The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

Sentiment Analysis for Domain-Specific Texts

Hidekazu Yanagimoto[1], and Michifumi Yoshioka[1]

[1] Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
(Tel: 81-72-254-9279, Fax: 81-72-254-9909)
{hidekazu, yoshioka}@cs.osakafu-u.ac.jp

Abstract: We develop a sentiment analysis system for domain-specific texts. The sentiment analysis estimates a polarity of a text, for example positive or negative. To develop such a system a dictionary, which consists of words and their polarities, is needed. The dictionary reflects specialists ' knowledge and affects a sentiment analysis precision. Hence, it is important to construct an appropriate dictionary. And making a dictionary needs much human cost generally. As a good dictionary is constructed, the human cost must be decreased. To achieve the goal our proposed approach uses a bootstrap method to decrease human cost and a χ^2 statistic to estimate a polarity of a word correctly. To evaluate a performance of the proposed approach we carried out an evaluation experiment using real stock market news, T&C news. We confirmed that the proposed approach could construct a dictionary but had some problems which were inappropriate words and that the dictionary did not have enough words.

Keywords: Sentiment Analysis, Bootstrap, Natural Language Processing

1 INTRODUCTION

Many reviews are posted in shopping sites and opinion sites and shared in the Internet. The reviews include opinions of users and are read by many users and affect a consumer activity. Hence, opinion mining and sentiment analysis is paid attention by many researchers. Turney[1] reported because there often were positive words around a positive word, a polarity of a word was determined using a cooccurrence frequency. Hatzivassiloglou et al[2] determined a polarity of a word using a conjunction. Nasukawa et al[3] estimated a polarity of a word using context coherence. Pan et al[4] focuses domain-specific words and domain-nonspecific words and developed a cross-domain sentiment classifier.

In the paper we focus a sentiment analysis for domain-specific texts. The domain-specific texts include a lot of technical terms because only users shared with the same interests read them. Hence, to develop the sentiment analysis system we must deal with the technical terms effectively. However, it is difficult to determine polarities for the terms because specialists for the specific domain need to determine them. This means that the specialists pay much effort to construct a dictionary that represents polarities of the technical terms. To solve the problems a dictionary construction approach must decrease human costs.

Previous works used a cooccurrence frequency to determine a polarity that a word has not attached yet. It is important what kind of evaluation criterion is used to evaluate how often the word occurs with positive words or negative words. The simple cooccurrence frequency includes much noise and is not a reliable criterion. Hence, a new evaluation method, which is hard to be disturbed by noise, must be proposed.

The χ^2 statistic is one of criteria that evaluates a bias of an occurrence probability. As the χ^2 statistic is big, it means that occurrence of the word has statistical significant. Hence, if words have the bigger χ^2 statistic than a threshold that is enough big, a pseudo cooccurrence can be neglected. Using the χ^2 statistic, we construct a dictionary that includes many words attaching correct polarities.

In the paper we presents a dictionary construction approach that uses a bootstrap method and a χ^2 statistic to evaluate the polarity of a word. The bootstrap method starts a dictionary construction from a small dictionary and expands it using predefined rules. Hence, the initial small dictionary must be constructed manually but human costs are decreased because of the small number of words in the initial dictionary.

Finally we check our proposed approach using real stock market news, T&C news. At first the polarity dictionary is constructed using all the articles in T&C news. We discuss the dictionary constructed with our proposed approach and found some characteristics.

1. The dictionary consists of the similar number of positive and negative words.

2. The dictionary includes words attaching correct polarities form the viewpoint of human judgment.

3. The dictionary includes some inappropriate words. For example, some words have opposite polarities.

4. The number of words in the dictionary is too small comparing with the number of all adjectives in all the articles.

© ISAROB 2012

In the next section we describe the proposed approach and the details of our approach are presented. In Section 3 we explain experiment and discuss the proposed approach. And we conclude our work in Section 4.

2 PROPOSED METHOD

We explain a dictionary construction method using a bootstrap method and χ^2 statistic for domain-specific texts. To determine a polarity for a domain-specific text it is important to use technical terms appropriately. For example, though "hooked" is often appeared in video game reviews[4] and is a positive word, it is a neutral word in some cases and a negative word in other cases. So it is important to construct a specific dictionary that assigns correct polarities with words.

In making the dictionary it is important to decrease human cost. Because a dictionary is constructed using a lot of specialists' knowledge generally, it denotes that making a dictionary needs a lot of human costs. Hence, we develop an automatic dictionary construction method using the least human costs.

To achieve our aim we use a bootstrap approach to construct a dictionary. The bootstrap approach stars from a small dictionary where some words are registered according to specialists' knowledge. A dictionary construct cost is very small because the initial dictionary is small. The approach expands the dictionary using rules that are used to select words that are assigned with polarities. In a proposed approach we use a cooccurrence frequency with words that have positive or negative polarities. We show a bootstrap flow below.

1. Select seed words and assign their correct polarities manually.
2. Count cooccurrence frequency with polarity-attached words.
3. Select words that polarities are estimated using the cooccurrence frequencies.
4. Retuen step 2 if a polarity is assigned with a new word in step3.

In the flow it is important what kind of criterion is used. Words that occur frequently with positive words are assigned with positive polarities. Hence, we must evaluate a bias of the cooccurrence frequency. The proposed approach uses a χ^2 statistic to evaluate the bias because a raw cooccurrence frequency is so noisy. Equation (1) shows how to calculate χ^2 statistic from cooccurrence frequencies.

$$\chi^2(w) = \sum_{w^s \in D_s} \frac{(\mathrm{freq}(w^s, w) \quad \frac{\mathrm{freq}(w^s)\mathrm{freq}(w)}{N})^2}{\frac{\mathrm{freq}(w^s)\mathrm{freq}(w)}{N}} \tag{1}$$

The $\mathrm{freq}(w_1, w_2)$ is cooccurrence frequency in a sentence and the $\mathrm{freq}(w)$ is occurrence frequency in a text corpus. The N denotes the number of all the sentences in the text corpus and the D_s denotes a set of seed words.

In this approach we use the χ^2 statistic to evaluate how often a word occurs with positive words or negative words. In constructing a dictionary the number of words attached with positive polarities are different from the number of words attached with negative polarities. Because the number of polarity-attached words have an affect on the χ^2 statistic, we must deny the affect in calculating the χ^2 statistic. The improved χ^2 statistics is defined below.

$$\chi^2_P(w) = \frac{1}{|D^P_S|} \sum_{w^s \in D^P_s} \frac{(\mathrm{freq}(w^s, w) \quad \frac{\mathrm{freq}(w^s)\mathrm{freq}(w)}{N})^2}{\frac{\mathrm{freq}(w^s)\mathrm{freq}(w)}{N}} \tag{2}$$

$$\chi^2_N(w) = \frac{1}{|D^N_S|} \sum_{w^s \in D^N_s} \frac{(\mathrm{freq}(w^s, w) \quad \frac{\mathrm{freq}(w^s)\mathrm{freq}(w)}{N})^2}{\frac{\mathrm{freq}(w^s)\mathrm{freq}(w)}{N}} \tag{3}$$

The D^P_S and the D^N_S are a set of words with positive polarities and a set of words with negative polarities respectively.

The bigger a χ^2 static of a word is, the more reliable a polarity of the word is. Hence, when the χ^2 static is over a threshold, which are defined previously, the word is registered in the dictionary.

$$w = \begin{cases} \text{positive} & (\chi^2_P(w) > \text{threshold}) \\ \text{negative} & (\chi^2_N(w) > \text{threshold}) \end{cases} \tag{4}$$

As the threshold is big, it is difficult to register new words in the dictionary and to expand the dictionary. On the other hand, as the threshold is small, new words are registered easily and the dictionary includes many meaningless words. Hence, because we must adopt the appropriate threshold.

We describe a sentiment analysis using the constructed dictionary. Because the dictionary consists of words and their polarities, a polarity of an article is determined according to how many positive or negative words are included. The above judgement uses a assumption that is that negative articles include many negative words and positive articles include many positive words. Hence, when an article includes more positive words than negative words, the article is estimated as a positive one. One the other hand, when an article includes more negative words than positive words, the article is estimated negative one. When there are the same number of positive words and negative words in a article or there are no polarity-attached word in a article, the article is estimated as neutral one.

The proposed approach does not consider negation and adversative conjunction at all. This decreases a performance of the proposed approach. This improvement is a future work.

© ISAROB 2012

3 EXPERIMENT

3.1 Data Set

In this experiments we used T&C news, which was one of stock market news delivery services, as data for dictionary construction. T&C news is written in Japanese and consists of 62,478 articles in 2010, which are stock price news, business performance reports, comments of specialists and so on.

We used all articles to construct a dictionary including word polarities. Using some articles to evaluate a performance of the proposed method is not a severe drawback, although the articles is shared in a construction step and an evaluation step. Because the proposed approach does not need labels of articles, which denotes whether the articles is positive or negative.

We extracted sentences from the all articles. The T&C news includes many tables and charts to explain a stock market situation for readers. The non-sentence elements tend to overestimate statistics to determine whether words are registered in a dictionary. After this preprocess we got 2,084,105 sentences from our data set. Hence, we made a polarity dictionary using all the sentences.

To construct the dictionary we used only adjectives in the sentences because adjective includes stronger polarity than other part of speech, for example, noun, verb, and adverb. In this experiment we used MeCab, which is a morphological anlyzer for Japanese, to extract adverbs from the sentences.

3.2 Dictionary Construction

We constructed a dictionary, which define words' polarities, using the all sentences, which are extracted from T&C news.

The proposed approach needs seed words to make the dictionary with a bootstrap method. Table 1 shows the seed words in this experiment. In this case we used 5 negative polarity words and 5 negative polarity words. Because the seed words are selected manually according to domain knowledge, they are correct polarities.

Table 1. Seed words to construct the dictionary

Positive polarity	Negative polarity
powerful	obscure
accessible	heavy
strong	low
successful	poor
positive	difficult

Using the seed words, we carried out the proposed method and constructed the dictionary. In this process we decreased the threshold, which was used to determine whether new words were registered in the dictionary. At first the registration criterion is so strict but the criterion is not strict finally. Because polarity-attached words are few At the beginning of the process, χ^2 statistic is not reliable, the threshold should be high. On the other hand after the polarity-attached words are enough, we can trust χ^2 statistic. So the threshold is small at the end of process. In this experiment the threshold was equal to 300.0 and is discounted at the rate of 0.95 times per loop.

Table 2 shows the number of words that are registered in the dictionary after executing the propose approach. Because there are 1,166 unique adjectives in T&C news, we think the number of words is so small. It is future work to increase the number of words that polarity we can estimate. The number of word with positive polarity is similar to the number of words with negative polarity. We found it was favorable that the number of positive words was similar to the number of negative words. If we use a dictionary that included more negative words, almost all articles are judged as negative article in a sentiment analysis step because negative words occurs in almost all articles easily. So the dictionary constructed in the experiment has a good feature.

Table 2. The number of positive polarity words and negative polarity words in the dictionary

Words with positive polarity	48
Words with negative polarity	47

We discuss the words that are registered in the dictionary using the propose method. Table 3 shows polarity-attached words in the dictionary. Some words, for example "appropriate" and "inappropriate", "necessary" and "unnecessary", "unstable" and "hard", and "thick" and "hard", has correct polarities. The "appropriate" and the "necessary" are common words and their polarities are estimated easily. The "thick" and "hard" is not common words but technical terms in stock market. It is difficult to estimate their polarities without using stock market news. Hence, the proposed approach is good because it can adapt easily using domain-specific texts and human load is small.

Table 3. The dictionary constructed with the proposed method

Words with positive polarity	Words with negative polarity
appropriate	inappropriate
necessary	unnecessary
unstalbe	stable
thick	hard
annoying	comfortable

The proposed approach has some drawbacks. In Table 3 some words are incorrect opposite to human judgement. The "annoying" is usually a negative word and the "com-

© ISAROB 2012

fortable" is usually a negative word. Because the proposed method does not consider negation and adversative conjunction in a sentence, the problems happen. It is a future work to improve the proposed method considering negation and adversative conjunction. Now we illustrate a problem on adversative conjunction. An article includes a sentence, "it is positive in C-duration and it is negative in D-duration." The sentence does not include adversative conjunction but the former clause is opposite to the latter clause. It is difficult to solve the problem from viewpoint of syntax level processing.

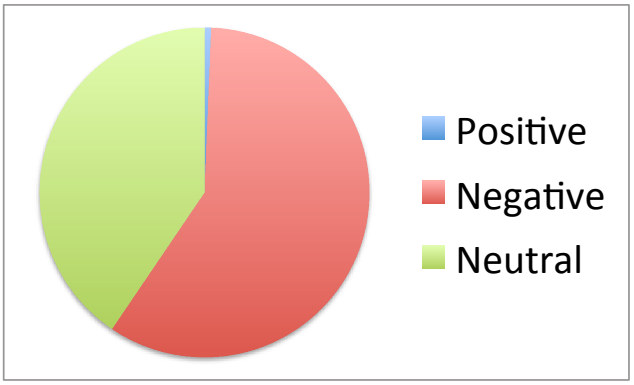

Fig. 1. The distribution of sentiment in all articles

Figure 1 shows the distribution of positive articles, negative ones, and neutral ones, which are not assigned with polarities. This figure shows negative articles occupy over half of articles. Because the articles are created after Lehman Shock, they include many negative articles. The number of positive articles, however, is too small in the result. Hence, the dictionary constructed with the proposed approach is difficult to capture sentiments of articles. There are many neutral articles in the result. This means that the number of words in the dictionary is too small.

4 CONCLUSION

We proposed a dictionary construction method to estimate correct polarities for domain-specific texts. The proposed approach has some features: (1)low human costs, and (2)easy adaptation for specific domain. The approach uses (1)a boostrap method and (2)a χ^2 statistic to construct the dictionary automatically. We confirmed that the proposed approach could make a dictionary in assign correct polarities with few words manually in an evaluation experiment.

The proposed approach, however, needs some improvements. The approach can not assign polarities with enough words in a data set. To improve the drawback we must discuss a χ^2 statistic and a criterion, which determine whether a word is registered in a dictionary or not. The approach assign incorrect polarities with some words. To improve the drawback we must consider negation and adversative conjunction in a sentence.

ACKNOWLEDGEMENT

We thank Centillion Co. for giving us data sets and many supports.

REFERENCES

[1] Peter D. (2002), Thumbs up? thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp.417-424.

[2] Htzivassiloglou V. and McKeown K. R. (1997), Predicting the Semantic Orientation of Adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp.174-181.

[3] Nasukawa T. and Kanayama H. (2004), Acquisition of Sentiment Lexicon by Using COntext Coherence (in Japanese). NL-162-16, pp.109-116.

[4] Pan S.J. et al. (2010), Cross-Domain Sentiment Classification via Spectral Feature Alignment. Proceedings of WWW2010, pp.751-760.

© ISAROB 2012