ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 627-631
doi: 10.18178/wcse.2017.06.109

Measuring Interrogative and Negative Opinion Expression in Social Media Sentences

Qiang Gu[1], Xuejun Yu [1], Ye Li [1], Yiqun Zhu[1], Guorui Cao[1] and Bo Wang[2+]

[1] State Grid Tianjin Electric Power Research Institute, China
[2] Tianjin University, China

Abstract. Measuring of interrogative and negative expression in social media sentences is a task to judge users' opinion about objects. With the emerging of big data, it becomes difficult to obtain accurate opinion judgement with low cost. To solve this problem, we present a method synthetizing linguistic rules and statistical learning. Both single-word and multi-word patterns are used in rules and machine learning process. The synsthized method succeed the rule-based and statistical method independetaly in a real task on large scale microblog corpus.

Keywords: Opinion measuring, Interrogative expression, Negative expression, Social media.

1. Introduction

The goal of opinion mining is to identify individuals' views on objects. Models of opinion mining based on social media content enable the simulation and prediction of the evolution of opinions [1]. The focus is on the automatic identification and extraction of opinions from text and multimedia [2-3]. For example, opinion detection has been proposed as a key technology allowing the automatic analysis of the opinions [4-5] such as 'positive opinion', 'neutral opinion', 'negative opinion' and 'information'. Sentiment analysis investigates the opinion with the emotional part using various categorization algorithms [6-7, 11-12]. These approaches are effective on formal corpus [8-10] but is not very satisfying on informal short texts with considerable noise. In this work, as a case study, we focus on two important kinds of opinion expression in Chinese social media sentences: interrogative and negative. A novel solution is proposed to recognize these two kinds of opinions on large scale and complicated microblog sentences synthetizing statistical and rule based methods.

2. Problem Definition and Analysis

2.1. Problem Definition

The problem requires a sentence category judgment of a given natural language sentence from social media. Target categories include: "interrogative sentence", "negative sentence" and "the others". For a given sentence, the available information includes:

Prior linguistic knowledge: mainly include the linguistic knowledge related to the interrogative, and negative expression, on the lexical, syntactic and semantic level.

Labeled samples: samples which have been labeled with one of the three categories.

Linguistic features of unlabeled sentence. The linguistic features involving lexical, syntactic, and semantic information of the sentences.

+ Corresponding author. Tel.: +86 13512828426.
E-mail address: *2944440@qq.com*.

With above knowledge, the problem can not only be modeled as a supervised or semi-supervised classification problem from the view of statistical learning. The choice depends the quality and quantity of the labeled data.

2.2. Challenges on Expression Measuring of Social Media Sentences

Large scale of corpus

In the real task of this work, the scale of social corpus is 1.2 million short sentences. As a result, the automatic linguistic analysis of the corpus requires considerable time and space consumption, which is a major challenge on expression measuring of social media sentences.

Scale difference between training and test set

In this real task, training set contains only 2000 labeled sentences while test set contains 1.2 million sentences. In machine learning, it is an unconventional task to label large scale unkonwn samples with the model trained on very small training set. The fundamental challenge is that the small training set has significant low coverage of the patterns in test data.

Distribution bias of different kinds of samples

In this real task, the distribution ratio of the samples of three categories is unbalanced. The proportion of the interrogative sentences is 10%-20%, the negative sentences do not exceed 10%, and the proportion of the other sentences is about 80%. This is another typical challenge for classification.

The informality and complexity of the social media texts

The informality and ambiguity is especially significant for spoken language in social media than other kinds of corpus. This also makes it challengeable to understand the expression of the sentences.

Balance of precision and recall

In natural language, there are always ambiguous sentences whose types are difficult to be determined even by artificial judgment. In practice, the judgment of the ambiguous samples is actually a task to balance the precision and recall in classification which is another challenge.

2.3. Our Solution

To deal with these challenges, we adopt a combination of rule-based and statistical methods on the typical feature words. We construct linguistic rules for high precision judgment using typical feature words according to linguistic knowledge. We also build statistical classification model with a small set of labeled samples and typical feature words as well. When we make the decision, we synthetize the rule-based and statistical methods with the principle of rules first. It is noted that, we also adopt a semi-supervised strategy to interatively lablel the large amount of unkown samples starting with a small training set. The advantage of this solution mainly comes from following aspects:

Make the use of rich prior knowledge.

From the view of knowledge utilization, the rule-based method introduces the prior knowledge of linguistics and statistical method introduces the prior knowledge annotation. The combination of the rules and statistical method can makes full use of two kinds of available prior knowledge.

Cope with the fuzziness of semantics and reduce the problem of noise.

Rule-based method has high precision, but it is often too rigid in the case of the semantic ambiguity. And the method based on statistics has better flexibility for ambiguous cases.

Relieve the scale gap between training and testing set.

The scale gap between training and testing set brings challenges to statistical learning. However, rules-based approach can effectively deal with part of the patterns which cannot be covered by training set. Furthermore, the semi-supervised strategy in our solution can also deal with the challenge of scale differnce between training and testing set with an interative labeling process.

Precision and recall balance.

Rule-based approach has a relatively higher precision and lower recall. On the contrary, statistical learning method has a relatively higher recall and lower precision. Therefore, the combination of the two methods can help to balance the precision and recall.

Relieve the problem of colloquialism and informal expression.

By extracting the typical feature words, we can effectively eliminate the irrelevant factors in sentences so as to reduce the problems caused by colloquialism and informal expression. Then, on the basis of typical feature words, we can construct the rules-based and statistical models which can capture the language patterns for expression judgement more effectively.

3. Recognition of Interrogative and Negative Expression in Microblog Sentences

3.1. Data Clean

In order to deal with the noise and to remove the influence on the linguistic analysis made by the specific symbols we firstly clean the social media sentences with following steps:

Remove duplicate expressions. We firstly remove and merge the repeated and similar sentences with similarity measuring, respectively.

Relabel the merged sentences. The original labels of repeated and similar sentences may be conflict to each other. Therefore, after merging, we relabel the merged sentences automatically.

Correct the labeling noise. For the noise existing in the sentences, we automatically identify the abnormal samples as noise and then correct according to the common label with similar samples.

3.2. Single-Word Features Selection based on Linguistic Knowledge

In social media corpus, we have many challenges especially the large scale of corpus and the informality of content. To overcome these challenges, we firstly propose to use the single-word features which is relative simple and can be effectively handled. A single-word feature refers to a single word which has important meaning for an interrogative or negative sentence. Single-word features are essential in our technical scheme. On one hand, the rules which are formed by single-word features are the basis of the rule-based method. On the other hand, single-word features are also a feature expression form for statistical classification method.

In this task, we summarized some classical Chinese linguistic knowledge of interrogative sentence and negative sentence analysis, observed the specific data labels and then screened a number of important single-word features.

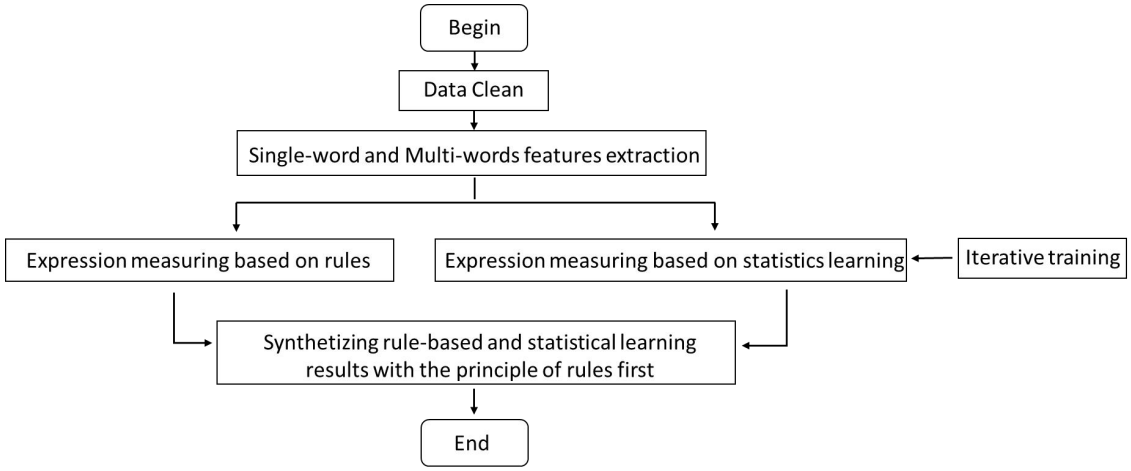

Fig. 1: Solution of measuring opinion expression in social media sentences synthetizing rules and statistical learning

3.3. Multi-Words Features and Concise Syntactic Pattern Identification

Parsing is widely used to measure the expression of language. However, parsing has a very high cost on spoken language and large corpus. Multi-Words features are the skipped combination of words. We construct multi-words features regarding them as “concise syntactic pattern”. The advantages of the concise

syntactic pattern are that it not only expresses the context of the feature words, but also records the order and the position of the feature words in sentences.

Sentence Expression Measuring based on Rules

Although statistical methods are widely used, rule-based methods are still common in NLP tasks. In despite of its high cost and relative low recall, rule-based method usually gets a high precision. In our scheme, we construct the decision rules based on the single-word, multi-words features and the relationship between feature words and their contexts, i.e., there context patterns.

Sentence Expression Measuring based on Statistical Learning

Due to the noise of training set labels and the scale difference between training set and unknown samples, we use maximum entropy model as the main statistical learning model which has a good robustness. In this task, we have used the popular maximum entropy toolkit developed by Zhang Le. For features setting, both single word features and multi words features are integrated.

Sentence Expression Measuring Synthetizing Linguistic Rules and Statistical Learning

Rules-based method has a higher precision and statistical method has a higher recall. We use rule-based methods first and label the rest of the samples with statistical model. The whole process can be illustrated with Figure 1, Algorithm 1 and Algorithm 2.

4. Experiments (CCF National Big Data Innovation Contest 2015)

The task and data of experiments are from the Chinese National Big Data Innovation contest 2015 . The traning data contains only 2000 labelled Chinese microblog sentences, and the test data contains 1.2 million microblog sentences. Based on the above ideas, we gradually tried following schemes.

Rules only: Measurement only based on single-word features rules.

Statistics only: Measurement only based on single-word features and maximum entropy based classification model.

Semi-supervised strategy with iteration: measure the 1.2 million test samples interactively with 2000 original labelled training samples using proposed semi-supervised strategy.

Rules + Statistics: combine the rules and statistical method following rules first principle.

Rules + Statistics + Multiple words features: introduce multi-words features to the model

As the results shown in Table 1, the performance of the combination method exceeds the independet rules or statistics based method. Iterative method and multi-words features can further improve the performance.

Algorithm 1

Input: test data S, training data T, training data labels TL
Output: test data labels SL
Initialize: null
For each sentence in S
divide sentences into clauses
data clean
recognize interrogative expression
recognize negative expression
merge results of negative expression recognition and results of interrogative expression recognition

Algorithm 2

Input: new test data after dividing and cleaning S1, new training data after deleting repeat, dividing and cleaning T1, new training data labels corresponding to the new training data TL1
Output: results file of interrogative expression recognition IF
Initialize: null
The algorithm do:
change TL1 with three categories into a result file with two categories: interrogative sentence and others
get the machine learning training file
get the machine learning test file

Use the maximum entropy tool to get the model through training of the training file, and then use the model file to predict the test file to get the forecast results

For each clause in S1 do

If c statisfy one of the rules then

c.label = interrogative

Table 1: Experimental results figure

Methods	F-score
Rules only	0.86643
Statistics only	0.86725
Semi- supervised strategy with iteration	0.87954
Rules + Statistics	0.88673
Rules + Statistics + Multiple words features	**0.90303**

5. Conclusions

It is a widely known challenge to measure the opinion expression in social media texts because of the large scale of data, informal statements, strong ambiguity and small labelled set. In order to solve these problems, we propose a solution containing a set of key technologies, including data cleaning, extraction of single-word and multi-words features, interactive learning strategy and the combination of rules and statistics for judgement. In a real contest task, our method obtained 0.90303 F-score and won the second position. The proposed solution is proved to be effective in social language processing with practical problems including low signal-to-noise ratio, training sample sparse, uneven distribution of categories and informal expression.

6. References

[1] Rainie, L. Election 2006 online. Pew internet & American life project report. (2007)

[2] Chesley, P, et al. Using verbs and adjectives to automatically classify blog sentiment. Proceedings of AAAI, the Spring Symposia on Computational Approaches. (2006)

[3] Lin, W. H., et al., A. Which side are you on? Identifying perspectives at the document and sentence levels. IJCNLP. (2006).

[4] Allen, C., et al. A place for emotion in attitude models. Journal of Business Research. (2005).

[5] Kwon, N., et al.. Multidimensional text analysis for eRulemaking. Proceedings of dg.o. (2006).

[6] DeSteno, D., et al. Discrete emotions and persuasion: The role of emotion-induced expectancies. Journal of Personality and Social Psychology, 86, 43. (2004).

[7] Mitrović, M., et al., B. Networks and emotion-driven user communities at popular blogs. The European Physical Journal B: Condensed Matter and Complex Systems, 77, 597–609. (2010).

[8] Thelwall, M., et al. Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology, 61. (2010).

[9] Chmiel, A., et al. Negative emotions boost users activity at BBC Forum. Physica A (2011).

[10] Ding, F., and Liu, Y. Modeling opinion interactions in a BBS community. The European Physical Journal B. (2010).

[11] Poria, Soujanya, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based Systems (2016).

[12] Li P, Yan Y, Wang C, et al. Customer voice sensor: A comprehensive opinion mining system for call center conversation.Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on. IEEE(2016: 324-329)