

Abstract Sentence Classification for Scientific Papers Based on Transductive SVM

Yuanchao Liu¹, Feng Wu¹, Ming Liu¹ & Bingquan Liu¹

¹ School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Correspondence: Yuanchao Liu, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. E-mail: lyc@insun.hit.edu.cn

Received: May 19, 2013 Accepted: June 19, 2013 Online Published: September 29, 2013

doi:10.5539/cis.v6n4p125

URL: <http://dx.doi.org/10.5539/cis.v6n4p125>

This research was supported by “the Fundamental Research Funds for the Central Universities” (Grant No. HIT.NSRIF.2009065) and Key Laboratory Opening Funding of China MOE-MS Key Laboratory of Natural Language Processing and Speech (HIT.KLOF.2009022)

Abstract

Presently, sentence-level researches are very significant in fields like natural language processing, information retrieval, machine translation etc. In this paper we present a practical task on sentence classification. The main purpose of this work is to classify the abstract sentences of scientific papers in the corpus built by ourselves into four categories- the background, the goal, the method and the result- which differ from each other in common usage, so that we can do further researches such as frequent pattern mining, information extraction and making a corpus for writing assistant system of scientific paper with these results. The main method of the classification is the Support Vector Machine, which is acknowledged among the best machine learning methods in the common text classification tasks. A semi-supervised method, Transductive Support Vector Machine, is also introduced into this four-class classification task to improve the accuracy. The experiments are conducted upon the corpus made by ourselves that consists of abstract sentences of scientific papers. The accuracy of the classifier finally reaches 75.86% with the semi-supervised method.

Keywords: abstract sentence, scientific paper, sentence classification, SVM, Transductive SVM

1. Introduction

Over the recent years, the study about short text becomes more and more significant in many fields such as natural language processing, information retrieval, machine learning etc. Our work aims at classifying abstract sentences in scientific papers into four categories according to the content of the sentences. In general, the abstract usually consists of the background part, the goal part, the method part and the conclusion part. In addition, they are also the most important parts even in the text of a scientific paper. So we consider it very significant to classify the sentences into the four categories for further usage like frequent pattern mining, information extraction and writing assistant of scientific paper with a large number of predicted sentences.

The abstract almost covers the whole content of a paper clearly and concisely, and sentences in it usually have a large amount of information in spite of their short length. In the traditional text classification tasks, statistic machine learning methods usually have good performance along with the vector space model. Several studies suggest that Support Vector Machine is generally acknowledged to have the best performance in text classification among these statistical machine learning methods. However, short texts have quite sparse features, usually couples of words. So the traditional Bag-of-words method meets some problems which leads to low efficiency. In our task, an abstract sentence usually has tens of words. In addition, our aim is to classify the abstract sentences into “background”, “goal”, “method” and “result”, so some uncertain semantic factors may also influence the classification.

Because of the lack of pre-existing corpus, we build the corpus and mark the instances ourselves. We carry out the experiments in both supervised and semi-supervised methods upon the scale-limited data set. And finally, we improve the accuracy of classification after several steps. After the classification, we may then choose the high-confidence predicted abstract sentences in each category for further researches such as frequent pattern

mining, information extraction and writing assistant of scientific papers.

Figure 1 displays the process of our work. First, we build the corpus of abstract sentences ourselves in order to carry out the experiments, including acquiring the corpus, analyzing the corpus and tagging the instances in it. Second, we conduct a group of experiments on feature selection in order to get a better feature vector for classification in our task. Third, we carry out the classification task with both supervised method and semi-supervised method.

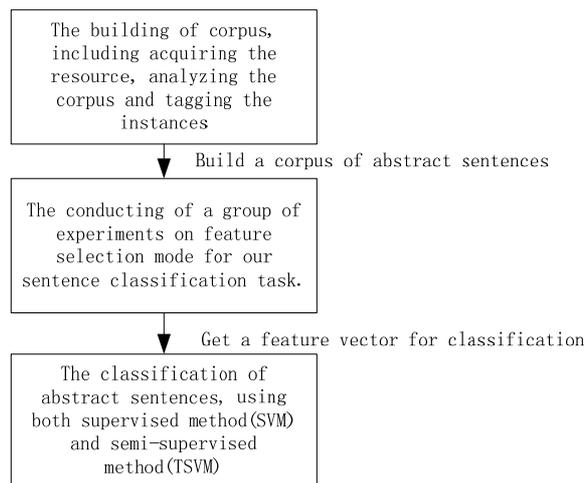


Figure 1. The process of our work

This paper is organized as follows: in Section 1, we introduce the background and aim of this work; in Section 2, we talk about the related work; in Section 3, we introduce the data set on which we conduct the experiment; in Section 4, we discuss the feature selection step of our work; in Section 5, we discuss the processing of the classification and the results of the experiments; in Section 6, we demonstrate the conclusion of our work.

2. Related Work

In this section we review some recent related literature on sentence classification and Transductive Support Vector Machine.

2.1 Sentence Classification

Research on sentence classification has been carried out over the recent years. In an earlier research, Naughton and Stokes et al. (2008) treated event detection as a sentence level text classification problem. They concluded that SVM consistently outperform the Language Model technique in their task, and that the manual rule based classification system was a powerful baseline that outperformed the SVM on half of the six event types. Lui (2012) used feature stacking to combine a variety of feature sets drawn from lexical and structural information at sentence level as well as the sequential information at the abstract level. Their system attained a ROC area-under-curve of 0.972 and 0.963 on two subsets of test data and produced the winning entry to the ALTA2012 Shared Task (Note 1). Molla (2012) found that the cluster-based feature improved the results for Naive-Bayes classifiers but not for better-informed classifiers such as Max-Entropy and Logistic Regression in their participation to the ALTA 2012 Shared Tasks.

2.2 Transductive Support Vector Machine

After Joachims (1999) proposed the Transductive Support Vector Machine, lots of machine learning researches and tasks were carried out based on it. Chen and Wang et al. (2008) introduced an application of TSVM in Chinese Semantic Role Labeling. They designed some heuristics from the semantic perspective to improve the performance of TSVM and the results showed that TSVM outperformed SVM in small tagged data and that after using heuristics TSVM performed further better. Miceli-Barone and Attardi (2012) presented a shift/reduce dependency parser that could handle unlabeled sentences in its training set using a transductive SVM as its action selection classifier. They performed the experiments with this parser on a domain adaptation task for the Italian language. TSVM was also used in fields like pixel classification (Chakraborty & Maulik, 2011; Maulik & Chakraborty, 2013).

3. The Building of Date Set

Our study object is mainly abstract sentences in scientific papers, because the abstract usually covers the whole content of a paper. The scientific papers are downloaded from the web site "<http://www.sciencedirect.com>", involving several fields. And then the abstract parts are extracted from the pages and split into sentences. By several studies and investigating of our corpus, we find that the abstract sentences could be classified into four categories as follows.

Class 1 is the background, usually including the general areas of the research, the specific research direction, the background and related work of the research, the significance and importance of the research, the usual research methods and basis of the area. For instance, the sentence "*Feature selection is an indispensable preprocessing step for effective analysis of high dimensional data.*" belongs to this category. So does the following sentence "*Finding an optimal feature subset for a problem in an outsized domain becomes intractable and many such feature selection problems have been shown to be NP-hard.*".

Class 2 is the goal, usually including the sentences that directly point out the proposed idea, methods, concepts, etc., but not the simple repetition of the aim. In addition, it can also contain the structure of the paper. For instance, the sentence "*This paper formulates the text feature selection problem as a combinatorial problem and proposes an Ant Colony Optimization (ACO) algorithm to find the nearly optimal solution for the same.*" belongs to this category. So does another sentence "*In this study, we focused on pathway figures that illustrate signaling or metabolic pathways, because many of these are important in understanding disease mechanism(s).*".

Class 3 is the method, usually including the process of the research, what method and data is used, as well as the principle and the conditions of the experiment. For instance, the sentence "*Documents from 20 newsgroup benchmark dataset were used for experimentation.*" belongs to this category, so does another sentence "*Multivariate analyses were performed to analyze the subject's perceptions and to build conceptual models for telephone design.*".

Class 4 is the result, usually including the observed results of the experiment, the analysis of the results, the conclusion obtained from the results, comparison with other results and the prospect of further work. For instance, the sentence "*An F -score of 0.78 is obtained for labeling relevant coordinating constructions in an independent test set.*" belongs to this category. So does another sentence "*Experiments showed that the performance of classifiers improved through adopting the proposed methodology.*".

We then tagged the corpus that has a scale of 4550 abstract sentences with the four labels. After splitting, we get 127718 words (including words, tokens, names, other special tokens) and a vocabulary book of 10346 words. Table 1 displays the statistics of the labeled corpus for each class. The "#sentence" refers to the number of the sentences for each class. The "proportion" refers to the proportion of the sentences in each class. The "#word" refers to the number of words in each class. The "vocabulary" refers to the scale of vocabulary for each class. And Figure 2 displays the distribution of the words after lowercasing, stemming, lemmatizing and stop words removal.

Table 1. Statistics for each category

	Label 1	Label 2	Label 3	Label 4
#sentence	1054	783	1716	997
proportion	0.232	0.172	0.377	0.219
#word	29035	23125	47162	28396
vocabulary	4542	3969	6418	4674

Description: The "#sentence" refers to the number of the sentences for each class. The "proportion" refers to the proportion of the sentences in each class. The "#word" refers to the number of words in each class. The "vocabulary" refers to the scale of vocabulary for each class.

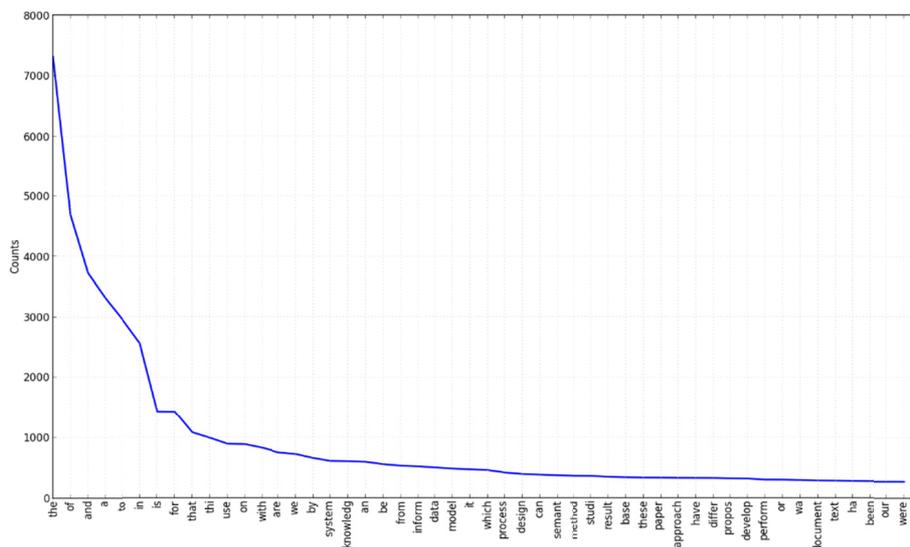


Figure 2. The Distribution of the Words

Description: This figure shows the distribution of the words after stop words removal, lowercasing, stemming, lemmatizing and stop words removal. The X-axis refers to the terms ordered by the counts of appearance and the Y-axis refers to the counts of the terms. It approximately follows the Zipf’s law.

4. Feature Selection

Feature selection is an essential step in classification tasks. In an earlier research (Yang & Pederson, 1997), term selection methods based on document frequency (TF), information gain (IG), mutual information (MI), a χ^2 -test (CHI), and term strength (TS) are compared in a text classification task. The conclusion is that IG and CHI are most effective in their experiment and simple methods with low cost such as DF can be reliably used instead of IG and CHI in case of expensive computation because of the strong correlation between them.

However, this sentence classification task meets the same problem of sparse feature that the common short text classification usually does. So we conduct a couple of experiments to find an appropriate feature selection method. We use the toolkit “libshorttext-1.0” to carry out the experiments (Note 2). We divide the prepared corpus into two parts, a training set with 2404 labeled sentences and a testing set with 2146 labeled sentences. The machine learning method of the toolkit is the linear Support Vector Machine. The following tables (Table 2, Table 3 and Table 4) show the results that indicate that stop words play an important role in abstract sentence classification and can significantly improve the accuracy, whereas both bi-gram and stemming do slightly. On the other hand, the four modes of feature’s value (binary, word count, TF, TF-IDF) hardly differ from each other. The explanation for the outstanding contribution of stop words is that some certain collocations that contain a stop word is usually used to write a certain category of abstract sentence. For instance, a sentence that begins with “In this paper ...” or “The aim of ...” has a high probability to be talking about the goal. In addition, even punctuations like “%” is likely to be a symbol of a sentence talking about the result or conclusion.

Table 2. Accuracy for each feature selection mode using linear Support Vector Machine

P \ F	0	1	2	3
0	62.2554	62.4418	62.3952	60.1118
1	64.9581	65.3774	65.3774	64.3057
2	61.2768	61.5564	61.3700	59.4595
3	65.0047	64.9115	64.9115	64.9581
4	56.4306	55.0326	55.0326	55.7316
5	57.2693	58.2013	58.1547	58.2479
6	55.2190	53.7745	53.7745	54.1938
7	57.2227	55.7782	55.7782	57.8285

Table 3. Feature selection mode for parameter P

P	Feature selection mode
0	No stop word removal, no stemming, uni-gram
1	No stop word removal, no stemming, bi-gram
2	No stop word removal, stemming, uni-gram
3	No stop word removal, stemming, bi-gram
4	stop word removal, no stemming, uni-gram
5	stop word removal, no stemming, bi-gram
6	stop word removal, stemming, uni-gram
7	stop word removal, stemming, bi-gram

Table 4. Feature selection mode for parameter F

F	Feature selection mode
0	binary
1	Word count
2	TF
3	TF-IDF

5. Classification of Abstract Sentences

5.1 Supervised Learning Method

Supervised statistical machine learning methods are widely used in text classification tasks, such as KNN method, Naive Bayes method, ME model, Support Vector Machine model, Decision Tree method, ANN model et al. And among them Support Vector Machine has the best effect in most common text classification tasks. However, for short text classification, Support Vector Machine also meets the problem of sparse feature. In our task, lack of training data is another factor that can affect the performance of the classifier.

We determine the feature selection modes based on the investigation above, namely lower letter formula, no stop word removal, stemming, lemmatizing and bi-gram. We make a CHI-test for the four categories in the training set and select the Top-1000 words that have the highest CHI-value from each category. Then we merge the four sets of words and finally get a 3758-dimension feature vector. We train the Support Vector Machine model on the training set with libsvm-3.16 (Note 3). With the RBF kernel trick and the grid technique, the accuracy rises to 70.2579%.

5.2 Semi-Supervised Learning Method

Semi-supervised learning method can be used in a task in the situation that the labeled training data is not enough to fit the distribution while a large number of unlabeled data is available. This situation should usually rely on the cluster assumption that the decision hyperplane should cross the area in which few spots are located. Joachims (1999) proposed a semi-supervised learning method based on Support Vector Machine model, the Transductive Support Vector Machine (TSVM). In his work he conducted several experiments on text classification and gave an explanation why TSVMs are especially well suited for text classification. Based on the theory he proposed, he developed the SVMlight to solve the optimization problem. Generally speaking, a TSVM trains a classifier both on the labeled training data and unlabeled data.

In our task, we carry out a group of experiment based on SVMlight-5.00 (Note 4). The training set contains the 2404 labeled instances and large amount of unlabeled data. The testing set is still the 2146 labeled instances. The feature selection mode is the same as Section 2 has described. On bringing in 8821 unlabeled instances totally, the accuracy rises to 75.8621% on the test dataset. The following figures (Figure 3, Figure 4, Figure 5 and Figure 6) show the trend of accuracy, precision, recall, F-score for each class after bringing in unlabeled instances step by step. The X-axis refers to the number of training data, 1 represents 2404 labeled instances, 2 represents 2404 labeled instances with 2246 unlabeled instances, 3 represents 2404 labeled instances with 4368 unlabeled instances, 4 represents 2404 labeled instances with 6657 unlabeled instances and 5 represents 2404 labeled

instances with 8821 unlabeled instances. The Y-axis refers to the percentage of four evaluation measures, the blue represents the accuracy, the pink represents the precision, the yellow represents the recall and the green represents the F-score. From the figures we discover that the accuracy remains almost the same, and the precision declines while the recall rises so as to lead to a slight rising in F-score. On one hand, the addition of unlabeled instances brings in extra prior knowledge thus increasing the recall. On the other hand, unlabeled instances make extra constraint on the decision hyperplane thus the overfit leads to the decline of the precision. Meanwhile the accuracy remains almost the same. Totally, the F-score (the harmonic mean of precision and recall) increases after bringing in the unlabeled instances, which indicates that the effect of classification is improved.

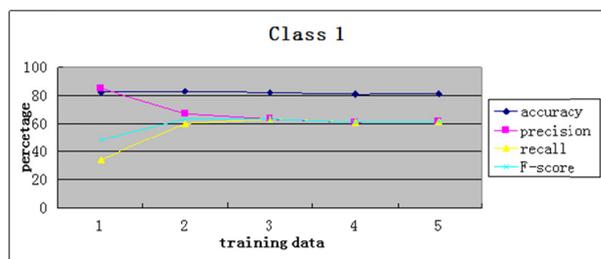


Figure 3. Trend of Evaluation Measures for Class 1

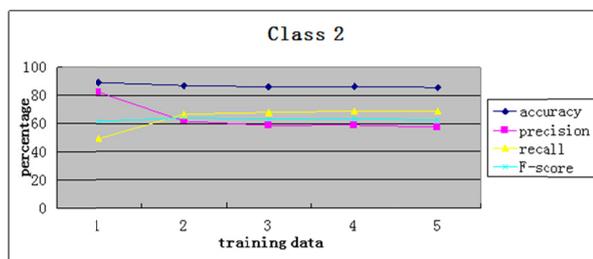


Figure 4. Trend of Evaluation Measures for Class 2

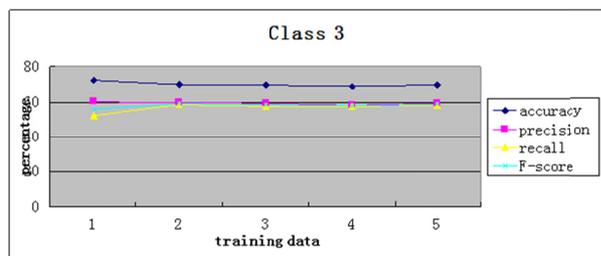


Figure 5. Trend of Evaluation Measures for Class 3

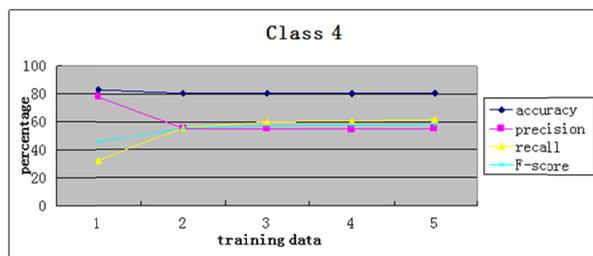


Figure 6. Trend of Evaluation Measures for Class 4

6. Conclusion

We conclude from our work as follows: First, in abstract sentence classification stop words make a contribution to improving the effect of a classifier, stemming do quite slightly, and bi-gram mode is a little better than the uni-gram mode. Second, different feature value modes do not differ obviously from each other; however, words have high CHI value can tell their category from the others so as to make up a better feature vector. Third, on bringing in unlabeled data and using semi-supervised learning method, the shortcoming of lack train data can be overcome, and the performance of the classifier is improved with a decline in precision and a rise in recall and overall a rise in F-score.

Finally we train a classifier that classifies abstract sentences into the four categories with the accuracy around 75%. And then highly-confident predicted instances are collected separately for further research and usage.

References

- Chakraborty, D., & Maulik, U. (2011, December). Semisupervised pixel classification of remote sensing imagery using transductive SVM. In *Recent Trends in Information Systems (ReTIS), 2011 International Conference on* (pp. 30-35). IEEE.
- Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27. <http://dx.doi.org/10.1145/1961189.1961199>
- Chen, Y., Wang, T., Chen, H., & Xu, X. (2008). Semantic Role Labeling of Chinese Using Transductive SVM and Semantic Heuristics. In *IJCNLP* (pp. 919-924).
- Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. In *ICML* (Vol. 99, pp. 200-209).
- Joachims, T. (1999). Making large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, & A. Smola

- (eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.
- Maulik, U., & Chakraborty, D. (2013). Learning with transductive SVM for semisupervised pixel classification of remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 77, 66-78. <http://dx.doi.org/10.1016/j.isprsjprs.2012.12.003>
- Miceli-Barone, A. V., & Attardi, G. (2012, April). Dependency Parsing domain adaptation using transductive SVM. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP* (pp. 55-59). Association for Computational Linguistics.
- Mollá, D. (2012). Experiments with Clustering-based Features for Sentence Classification in Medical Publications: Macquarie Test's participation in the ALTA 2012 shared task. In *Australasian Language Technology Workshop*.
- Naughton, M., Stokes, N., & Carthy, J. (2008, August). Investigating statistical techniques for sentence-level event classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 617-624). Association for Computational Linguistics.
- VRL, N. (2012). Feature Stacking for Sentence Classification in Evidence-Based Medicine. In *Australasian Language Technology Association Workshop 2012* (p. 134).
- Wong, S. M. J., & Dras, M. (2010, December). Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop* (pp. 67-75).
- Xue, N., & Yang, Y. (2011, June). Chinese sentence segmentation as comma classification. In *ACL (Short Papers)* (pp. 631-635).
- Yu, H. F., Ho, C. H., Arunachalam, P., Somaiya, M., & Lin, C. J. (2012). *Product Title Classification versus Text Classification*. Technical report, 2012. Retrieved from <http://www.csie.ntu.edu.tw/~cjlin/papers/title.pdf>
- Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *ICML* (Vol. 97, pp. 412-420).

Notes

- Note 1. <http://alta.asn.au/events/sharedtask2012/>
- Note 2. <http://www.csie.ntu.edu.tw/~cjlin/libshorttext/>
- Note 3. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Note 4. http://www.cs.cornell.edu/people/tj/svm_light/

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).