**Construction of unsupervised sentiment classifier on idioms resources**

XIE Song-Xian(谢松县),WANG Ting(王挺)

Department of Computer Science and technology, School of Computer, National University of Defense Technology ChangSha, 410073, China

|  |
| --- |
| Abstract: Sentiment analysis has high value from both research and practical perspectives. This paper focuses on the difficulties in the construction of sentiment classifiers which normally need tremendous labeled domain training data. A novel method has been proposed to make use of the idioms resources to developing a general sentiment classifier. Furthermore, the general classifier has also been taken as the base of a self-training procedure to get a domain-specific sentiment classifier. The experiments on Chinese online reviews dataset show that the proposed work achieved encouraging result.  Keywords: senmentiment classification; self-training; idioms resources; machine learning |

**1 Introduction**

The amount of user-generated content (UGC) on the Internet has risen exponentially over the last decade, with the emergence and advance of web 2.0 technology, and such content is now always at our fingertips. User-generated contents, in particular, become an ever-growing source of opinions and sentiments which are spread worldwide through blogs, wikis, chats and diverse social networks such as Twitter[[1]](#footnote-1) and Facebook[[2]](#footnote-2)[1]. The distillation of subjective knowledge from such abundant sources is an important part of applications in fields such as commerce, tourism, politics and health, but the quantity and the nature of User-generated contents make it a very difficult and challenging task.

As a live example of our daily life, more and more review sites continue to grow in popularity as more and more people begin to refer the advice of fellow users regarding services and products before they make their dealing decision. However, with the explosion of such information, we are often forced through large quantities of and sometimes low quality reviews in order to find the useful information we want. This has led to increasing research interest in the areas of opinion mining and sentiment analysis, with the goal of finding effective methods and techniques that can automatically analyze reviews and extract the subjective information to be summarized for us[2].

Sentiment analysis is the computational study of how opinions, attitudes, emotions, and perspectives are expressed in language (especially in written text), so as to provide tools and techniques for extracting this kind of evaluative information from large datasets and summarizing it[3]. With the growing need of identifying opinions and sentiments automatically from text data on the web, sentiment analysis tasks have received considerable attention recently, and been applied to Business Intelligence, Public Opinion Analysis, Election Prediction, etc.

Sentiment classification, which deals with determining sentiment orientation of target text, is one of sentiment analysis tasks[4]. The task can be viewed as a specific text categorization problem. Given an instance of opinionated text (may be document, sentence or words, since we investigate document sentiment classification, text stands for document in this paper ), the goal is to classify it as positive or negative (or neutral in multi-class classification)[5]. In fact, sentiment classification is a more challenging task than text classification. Firstly, methods and techniques developed in traditional text classification usually do not work well on this task, since they tend to take frequent-occurring words(also called keywords) as good indicators of the class a document belongs to, however, for sentiment classification of an opinionated document, words indicating sentiment are usually ambiguous and maybe infrequent. Secondly and most importantly, sentiment expressions critically depend on domains and contexts[6], and opinions are often hidden in a large amount of domain dataset, so there are no universal resources available for sentiment classification. However acquiring human-labeled data for each domain is costly and difficult, for manual annotation is very expensive and time-consuming.

Since sentiment classification can be viewed as specific text categorization[4], many researchers cast their eyes on all kinds of machine learning techniques. With more and more work on document-level polarity classification using machine learning methods, various classifiers and feature sets have been explored[4][7], which can be categorized into supervised, semi-supervised and unsupervised approach.

Supervised approaches were firstly applied to sentiment classification by Pang, et al.[4] by comparing multiple supervised machine learning algorithms (Naive Bayes, maximum entropy, support vector machines) for the task of sentiment classification of movie reviews, and afterwards various classifiers and features selection has manifested in many other researches[8–13]. Performance of supervised approaches is reasonably satisfying because of the requirement that test data should be similar to manually annotated training data, and the high accuracy often is the upper bound for other methods to compare with. But collecting annotated data in the new domain and retraining the classifier are unavoidable for moving a supervised sentiment classifier to another domain. The dependency on domain annotated training datasets is one major shortcoming of all supervised methods.

Semi-supervised approaches try to improve the performance of classification by fitting labeled and unlabeled datasets together[14] with various methods such as EM on Naïve Bayes, co-training, transductive SVMs, and co-regularization, etc.[15–18].But just as supervised methods, labeled training datasets are needed for semi-supervised approaches, which are mainly annotated by hand, whereas there are some automated means, because reliability of training data is the main consideration for learning methods. And most importantly, training datasets are critically domain-dependent whether annotated automatically or manually, so different training dataset is needed to be labeled for different domain, which means supervised and semi-supervised sentiment classification is very hard for the domain without any training dataset. Although recent years many researchers have tried to solve the problem of shortage of training dataset by adaptive techniques (such as transfer learning)[19–21] to realize cross-domain (or cross-language)[20], [22–24] sentiment classification, most are inefficient in that adaptive learning is needed for the changing of target domain, and the low accuracy is another problem of adaptive methods because of disambiguation of sentiment word in different domain.

At the same time, many unsupervised approaches are brought forth to tackle the problem of annotated training data shortage and domain-dependence[25–32], most of which are domain-independent rules based, and try to get some highly confident examples produced by the rule-classifier as the training data for bootstrapping learning. These rules rely on specific sentiment expressions of linguistics analyzed by expertise, and there are two problems of these methods: firstly, linguistic expertise knowledge are needed for the manually produced rules; secondly, discovering all sentiment expressions in the text by a few rules is impossible because human intuition may not be correct or comprehensive for the complication of language[4], so limited and biased examples are returned as the low coverage of the rules, which will influence the performance of pipeline supervised classifier of the bootstrapping procedure.

The problem of domain dependence of sentiment classification was focused on in this work, and to overcome such constraint, methods based on general resources out of domains was put forward, of which a general classifier is trained on the off-the-shelf resources without the need of laborious labeling training data. The main contributions of this work include:

(1) A novel perspective for sentiment classification was propose by assuming that feature space could be divided into two parts, which is composed of: the domain-independent part, i.e. general sentiment features, which is identified by domain-independent resources used to train an general classifier; domain-dependent part, i.e. specific sentiment features, which is embedded in the target domain dataset and identified by highly confident outputs of the general classifier used to train in-domain classifier.

(2) The proposed approach was realized by training a general classifier on Chinese idioms resources to get the domain-independent classification model with general sentiment features, which can be applied to any domain without the need of labeled dataset.

(3) An in-depth analysis of experiment results was provided, revealing that general domain-general classifier can achieve better performance than cross-domain classifier. It is also showed that resources independent of domain and context are in general highly valuable, and that general classifier trained on these resources are particularly robust, showing promise for sentiment classification tasks with no labeled data available.

As a whole, the main goal of this work was to overcome the problem of domain dependency in sentiment classification, and the approaches distinguishes from all other techniques in that it is completely unsupervised and domain-independent, and needs no extra specialist knowledge. All that needed are some general resources which are highly opinionated but do not depend on any context and domain, the proposed general classifier trained on these resources can produce highly confident instances as training data for the in-domain self-training supervised classifier. So no manually annotated training data and no expertise knowledge based rules are needed any more for any domain, but robust and dependable classifier got.

**2 The proposed approach**

**2.1 Problem formulation**

2.1.1 Sentiment classification

Sentiment classification aims to automatically classify document as predefined classes (binary or multi-class, we study binary classification here for simplicity). Formally, Given documents and predefined category set , the task of sentiment classification is to classify each in , with a label expressed in . To be along with text categorization, each document can be represented as a vector of bag-of-words features , where is the size of a pre-specified vocabulary . The weight of each entry in this vector usually is specified as binary, with weight equals to 1 for terms present in the vector. Given a training dataset , we can build a binary classifier , and employ it to predict label for an unseen instance by computing , with each instance represent as a vector in which is the ith feature’s weight.

2.1.2 Features division

Often there is an assumption underlying previous sentiment classification researches which consider features appearing in a document representing the document’s sentiment polarity equally with different weights, for example in the following book review:

|  |
| --- |
| Chinese:这本书是我一个好朋友推荐并且送给我的，并且说就把它放在枕边，随时阅读，我抱着这样的想法一气呵成读完，觉得译者的语言翻译的很准确，没有多少翻译痕迹，并且语言形象生动，故事易于理解，令人身临其境，有不少借鉴意义。正打算读第二遍。 推荐!  English: The book is recommended and sent to me by one of my good friends, and he told me to put it beside my pillow so as to read it anytime available, I read it through without any letup with the same idea, and I feel the translation very accurate without any mark of translation, the language very lively and vividly and the story easily understood to make the reader personally on the scene, and there is much use for reference in the book. I am going to read it again. Recommend! |

**Fig. 1** Book review example

When classifying the review, all words extracted as features are considered potentially indicators of positive evaluation for the book, but after careful examination, It can be found that words “推荐(recommend)” and “准确 (accurate)” are positive indicators of reviews across all domains, while “形象生动(lively and vividly)”, “身临其境(personally on the scene) ” and “借鉴意义(use for reference)” are more frequently used in the book reviews to express positive evaluation. With this intuition, we propose an assumption as:

|  |
| --- |
| Assumption 1: In the feature space of sentiment classification, features can be divided into two parts: domain-independent part, i.e. general sentiment features which are indicators of polarity across all domains and independent of any context; domain-dependent part, i.e. specific sentiment features whose polarity depend on special context of each domain. |

**Fig. 2** Assumption about feature division

Based on assumption 1, feature vector representing a document of the sentiment classification task can be expressed as , where denote the weights of general part of features, and denote the weights of specific part of features.

Questions may arise about assumption 1: firstly, what’s the meaning of the division of feature space, and how to identify each part of feature space? Now imagine such a scenario, when reading a review about a professional book, that one could still distinguish which polarity (recommend or not) the reviewer prefer even if he knows nothing about the in-domain knowledge what the book describes, as long as “good”, “accurate”, “recommend” appear in the review. Intuitively, this kind of phenomenon may be explained by the general part of the text which is used to express the holistic sentiment polarity of the author, and the polarity of general sentiment words are prone to be recognized by anyone independent of in-domain knowledge. Comparably, in sentiment classification, we put forward that, the sentiment polarity of a document could still be recognized with only the general part of feature space , that is to say, theoretically, if we could model general sentiment polarity knowledge, we can still classify what polarity a review prefers for. Then another question is how to create such model. In fact, many researcher has tried to establish all kinds of sentimental ontology lexicons to represent this knowledge such as Sentiwordnet[33] and General Inquiry[34] in English, Hownet[[3]](#footnote-3) and NTUSD(Chinese Network Sentiment Dictionary)[35] in Chinese, etc. However, they all failed in modeling the universal sentiment knowledge in that many of entries with multiply senses in these lexical resources are unavoidably dependent of domain and context. Actually such knowledge exists in such cases as one word or combination of words could represent, so the way that models the universal knowledge could be changed by training a classifier on instances with sentimental features independent of domains. Then another question is: how to find such kind of instances? In fact, this research is motivated by the question.

2.1.3 Chinese Idioms

Many linguistic resources are highly valuable for sentiment classification, of which idioms resources attract interests of this work. Idioms are common part of languages beside Chinese, such as “castles in the air”, “a bed of thorns”, “bring down the house” in English. The form of idioms is succinct and the meaning is penetrating, which is a quintessence part in the language. Generally speaking, the structure is fixed and can’t be changed at will; idioms have semantic intactness, which is not generally the simple summation of the literal meaning of each component; they chiefly use metaphor, exaggerator and comparison in the rhetoric to express its real meaning; and most importantly, the semantic orientation of idiom is independent and unchangeable under any context. There are many lexical resources in Chinese with entries takes the form as:

|  |
| --- |
| Chinese: 空中楼阁：贬义词，比喻虚幻的事物或脱离实际的空想。  English: castles in the air: a derogatory term, indicate the illusive things or impractical fanciness metaphorically. |

**Fig. 3** Example entry of idiom

The entry is composed of three parts: the idiom “空中楼阁”(castles in the air), the semantic orientation “贬义词”(a derogatory term, negative polarity) and a short paraphrase with three general negative words (“虚幻的(illusive)”, “脱离实际(impractical)” and “空想(fanciness)”). The entry provides us with an illustrational universally-labeled instance with general sentimental features and negative label, which is independent of any domain just as the idiom it explains. Based on the observation, we make another assumption as follows:

|  |
| --- |
| Assumption 2: The sentiment polarity of paraphrase of an idiom is as independent of domains as idiom. |

**Fig. 4** Paraphrase independency assumption

With assumption 2 admitted, the training dataset could be set about constructing which is used to train the domain-independent classifier to model the universal sentiment knowledge, with the paraphrases of idioms as train instances, the semantic orientation values as sentimental labels, and the paraphrases being represented as vectors of general sentiment features.

**2.2 Algorithms**

2.2.1 General classifying algorithm

The same methods as Pang et al. [4] were adopt, because they have applied Naive Bayes, Maximum Entropy (Maxent) and Support Vector Machine (SVM) classification techniques to identify the effectiveness of machine learning on sentiment classification of movie reviews, and they got satisfying result (accuracy 82.9%) by using unigrams as features. Also unigrams were taken as features to train a general classifier on the training dataset constructed on Chinese idioms resources, and in order to prove its general capability of sentiment classification, it is necessary to construct multi-domains dataset to be used as test data.

* Naïve Bayesian classifier

Naïve Bayesian method is one of the most popular techniques for text classification. Given a set of training documents , each document is considered as an ordered list of words. Entry is used to denote the word in position of document , where each word is from the vocabulary . The vocabulary is the set of all words considered for classification. A set of pre-defined classes, is used to label a document. To perform classification, It is needed to compute the posterior probability, , where is a class and is a document. Based on the Bayesian probability and the multinomial model, the following equation is got:

(1)

With the assumption that the probabilities of document words are independent given the class:

(2)

In the Naïve Bayes classifier, the class with the highest probability is assigned as the class of the document.

* Maximum Entropy classifier

Maximum entropy (MaxEnt) classifier has been widely applied in many Natural Language Processing tasks[36][37]. The classifier assigns the conditional probability of the class label given the document as follows:

(3)

Where is a vector of feature weights and is a feature function that maps pairs to a nonnegative real-valued feature vector. Each feature has an associated parameter , called its weight; and is the corresponding normalization factor.

With a set of labeled documents, Maximum likelihood parameter estimation (training) for such a model is trying so solve such a optimization problem:

(4)

* Support Vector Machines Classifier

Support vector machines(SVM) classifier is a kind of discriminative method of machine learning techniques. Based on the structural risk minimization principle of the computational learning theory, SVM tries to find a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective discriminative points from the training dataset.

Multiple variants of SVM have been developed for different tasks. In this research, linear SVM is considered due to its popularity and high performance in sentiment classification. The optimization of SVM (dual form) is to minimize:

(5)

Subject to:

2.2.2 Domain-specific classifying algorithm

As the general sentiment features are only part of all features of feature space, the domain-dependent part must be considered in order to capture the subtle clues embedded in the specific sentiment expressions using specific sentiment features, so that in-domain instances can be classified accurately. Usually unsupervised approaches improve its performance by combining with supervised learning in target domain, and such ways were adopted to verify the rationality of the hypothesis about feature space division. Specifically, initial labels and the confidence of classification were identified for instances of each domain after applying general classifier to the multi-domain test dataset, so an in-domain supervised classifier could be trained on the labeled instances with high confidence by adopting semi-supervised self-training framework. The main idea of this bootstrapping method was to use a domain-independent classifier to label an in-domain dataset, and the instances labeled by the general classifier with high confidence served as training data for a supervised machine learning classifier. Generally speaking, the resulting supervised classifier should be more effect in domain than the general classifier as long as the general classifier were dependable enough, because too many errors passed into iterative training process would bring down the domain-specific classifier. The figure 5 illustrates the self-training algorithm being used.

|  |
| --- |
| Algorithm 1: in-domain classifier self-training algorithm |
| Input:  idiom data containing and samples;  domain test data  Output:  automatically labeled sample set  Procedure:  (a) train general classifier on dataset  (b) classify each instance in with classifier  (c) initiate in-domain training dataset  c1) sort according to confidence  c2) initiate with high confidence instances  c3)remove instances from  (e) Loop:  e1) train in-domain classifier on dataset  e2) classify each instance in with classifier  e3) if accuracy is falling, end the loop  e4) extend  e4-1) sort according to confidence given by classifier  e4-2) extend with high confidence instances  e4-3) remove instances from  (f) get by classifying with the final |

**Fig.5** In-domain classifier self-training algorithm

**3 Experiments**

In this section, systematically evaluation of the proposed approaches designed for document-level sentiment classification is introduced.

**3.1 Dataset Description**

Because most existent Chinese idioms dictionaries have limited entries and some of them do not contain sentiment polarity labels, at the same time laborious manual annotation are trying to be avoided, the resources on the web compose the main source of training dataset. After crawling from the online idioms dictionary of “中教网(China Education Network)”[[4]](#footnote-4), an idiom dataset of 24,395 entries with each entry labeled with “褒义”(positive), “贬义”(negative) and “中性” (neutral) are achieved. Since the research focus on binary classification, the neutral entries are removed from the dataset. As a result, a dataset of sizes of 8,160 examples is used for training general classifier. To evaluate the performance of general classifier, three Chinese reviews corpus[[5]](#footnote-5) on domains of book, hotel and notebook PC are used as test dataset, each of which consisted of 4,000 reviews (2,000 positives and 2,000 negatives).

**3.2 Packages and classifiers**

All Chinese text of train and test dataset is segmented with mmseg Chinese segmentation package[[6]](#footnote-6). Naïve Bayes classifier and Maximum Entropy classifier of NLTK (Natural Language ToolKits) [[7]](#footnote-7) package and Support Vector Machine classifier of libsvm package[38] are used for classification. All the parameters and settings are optimized by cross-validation, and details are not described in the paper.

**3.3 Feature selection**

Feature selection methods are used to pick out discriminating features for training and testing. The statistic measures calculate the association value between features and categories[39], define as:

(7)

Where is the number of times and co-occur; is the number of times occurs without ; is the number of times occurs without ; is the number of times neither nor occurs; is the total number of documents. Features with high values are selected and the last 5% low value features are removed.

**3.4 Baseline and upper bound**

3.4.1 Baselines

Two baselines are used to compare with the proposed method, one is naïve 50% baseline since the test corpus are all balanced with respect to the sentiment classes, the other one is the cross-domain classifier with the same algorithms and settings as general classifier, and the latter baseline is used to demonstrate the superiority of the proposed method in that not only can it be applied to any domain without any labeled training dataset but can be more robust and dependable than using labeled dataset of other domains.

3.4.2 Upper bound

As mentioned in introduction, domain-specific supervised machine learning methods are often used as upper bound which is used to be challenged by semi-supervised and unsupervised methods. In the experiments, an upper bound is also setup by training domain-specific supervised classifiers with the same algorithms and settings as general classifier except for the dataset settings, of which each domain dataset is split five-folded with one fifth for testing and others for training, and the performance is measured by averaging the results of five iterative.

**3.5 Performance measurement**

Of various complicated measurements of machine learning, the simplest accuracy measurement is chosen to evaluate the performance, as the comparison between measurements is not the important points of this research. In this paper, accuracy is defined as:

(8)

**4 Experiment Results**

**4.1 General classifier versus cross-domain classifier**

Because of the imbalance of the idiom dataset with 5,611 negative instances and only 2,549 positive instances, over-sampling techniques are adopted which specifically aim to balance the class populations through replicating the minority class samples[40]. After over-sampling the positive data, a balanced train dataset of 11,222 instances is achieved. For the cross-domain classifier, three classifiers are trained separately on full dataset of each domain and tested on the other two domains. The results are shown in table 1, and from the table the following results can be observed.

Table 1 Results for general and cross-domain classifier

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Book classifier | Hotel classifier | Notebook classifier | General classifier |
| book | NB |  | 48.425 | 59.350 | **69.995** |
| MX | 47.525 | 57.850 | **63.675** |
| SVM | 32.500 | 61.225 | **80.700** |
| hotel | NB | 50.312 |  | **76.344** | 66.425 |
| MX | 50.837 | **76.341** | 61.640 |
| SVM | 43.685 | 59.665 | **71.775** |
| notebook | NB | 50.100 | 62.675 |  | **65.100** |
| MX | 50.400 | 63.050 | **70.500** |
| SVM | 50.075 | 63.250 | **63.850** |

NB denotes Naïve Bayes classifier, MX denotes Maximum Entropy classifier and SVM denotes Support Vector Machine classifier

* Firstly, as it can be seen from the table that the accuracies of general classifier tested on three domains all surpass the naïve baseline (50 percent), with the least gap (11.640 percent) of Maximum Entropy classifier tested on the hotel corpus and the largest gap (30.7percent) of SVM classifier tested on the book corpus. The result proves that the general classifier is superior to random selection and may be better choice when there are no labeled dataset available for sentiment classification.
* Secondly, as for cross-domain classifiers, the performance of each classifier is diverse, from the worst hotel SVM classifier tested on book corpus (17.5 percent below 50%) to best notebook Naïve Bayes classifier tested on hotel corpus (26.344 percent above 50%). The result shows unstableness of cross-domain classifier and the reason will be discussed in the conclusion and discussion section. Besides, labeled dataset are needed for cross-domain classifier despite of out-domain dataset.
* Finally, for comparison between two kinds of classifier, the general classifier outperforms the cross-domain classifier with 5 largest accuracies versus 4 largest accuracies marked boldly in table 1. Though not obvious, the average accuracy of general classifier is much larger than the cross-domain classifier, and performs more stable, and in addition, no labeled data are needed for general classifier. Overall, it is more robust and dependable for general classifier.

**4.2 Domain-specific versus self-training classifier**

Three domain-specific classifiers are trained in five-folded cross-validation mode and the average accuracies are reported as the upper bound for self-training classifier, and as illustrated in table 1, the independent SVM classifier shows the best performance, so we decide to apply the self-training process using SVM classifier according to algorithm 1 described in last section. Table 2 illustrates the result.

Table 2 Results for in-domain and self-training classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NB | MX | SVM | Self-training |
| book | 55.087 | 57.487 | **93.375** | 86.200 |
| hotel | 72.407 | 77.910 | 87.479 | **87.521** |
| notebook | 90.337 | 90.212 | **90.700** | 81.500 |

From table 2, it is obvious that the in-domain SVM classifier performs best in accordance with conclusion of other researchers[4]. The performance of the general classifier with self-training improves a lot by bootstrapping training procedure in each domain, with 5.5 percent in book domain, 15.146 percent in hotel domain and 17.65 percent in notebook domain. But only in hotel domain, does self-training classifier surpass all three in-domain classifier. In book and hotel domain the self-training classifier also surpass in-domain Naïve Bayes and Maximum Entropy classifier. As a whole, the performance of the self-training classifier based on general classifier approximates to the upper bound domain-specific classifier, which demonstrates the efficiency of the proposed approach.

**5 Discussion and future work**

In this paper, a novel perspective for sentiment classification is introduced by dividing feature space into two part: the domain-independent part (general sentiment features) and the domain-dependent part (specific sentiment features). Elicited by human’s identification of sentiment polarity reading reviews, we suggest the general sentiment features play important role in sentiment classification, and propose to train general classifier with the general part features on domain-independent resources.

As for domain-independent resources, there are many linguistic sources with semantic orientation independent of domain and context, and can be found almost in every language. In this paper, we adopt Chinese idioms resources and make a reasonable assumption that the sentiment polarity of paraphrase is independent of any domain just as idiom. Based on the assumption, we construct a domain-independent dataset with resources crawled from web and train a general classifier.

It turns out that our method outperforms baselines in the comparing experiment with cross-domain classifier, and the general classifier is more robust and dependable especially when no labeled dataset are available.

A self-training supervised classifier which takes the high confident output of general classifier as input is trained in each domain to exploit domain-dependent features, and the performance is much like the upper bound domain-specific classifier. The output of general classifier has an impact on the performance of the self-training classifier. Usually, the more accurate the general classifier is, the better the resulting self-training classifier is.

Therefore, further improvement of general classifiers is one of our future works, and there are some ways to do so, first of all, more domain-independent dataset are needed to improve the coverage of features and reduce sparseness. Table 3 illustrates feature coverage between domains and coverage between idioms dataset and each domain. It’s obvious the coverage of idiom features is relatively small, and improvement of coverage would improve performance of general classifier. Sparseness is another factor that influence performance, according to our statistic, the paraphrase of each idiom is very short (19 characters of mean length) and the abandonment of words caused by segmenting error also causes sparseness. From table 3 we can see, although some coverage between domains is high, the cross-domain classification performance is not as good in that the different distribution of common part of features will make cross-domain classifier unsuitable in another domain.

Table 3 Feature coverage between domains and idioms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | book | hotel | notebook | idiom |
| Book  (20219) |  | 0.413  (8365) | 0.243  (4933) | 0.358  (7251) |
| Hotel  (16220) | 0.515  (8365) |  | 0.307  (4994) | 0.329  (5345) |
| Notebook  (7595) | 0.649  (4933) | 0.657  (4994) |  | 0.408  (3100) |

At present we only model unigram to construct the polarity classifiers, many researches in supervised learning sentiment classification suggest that more advanced linguistic modeling is likely to improve classifier. So our next future work will focus on mining more useful linguistic features to improve performance of general classifier.

**6 Conclusions**

1) With the need of sentiment classification in various domains, domain dependency has become the bottleneck of machine learning techniques with the shortage of labeled domain data.

2) General sentiment classifiers can be obtained by making use of all kinds of linguistic resources independent of domains such as Chinese Idioms resources with the assumption that feature space be divided into the general part and the specific part.

3) The performance of general classifier can be improved by adopting self-training procedure in the target domain which can approximate the upper bound supervised learning in-domain classifier.

**References**

[1] M. Miller, C. Sathi, D. Wiesenthal, J. Leskovec, and C. Potts, “Sentiment Flow Through Hyperlink Networks,” 2011.

[2] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.

[3] A. Agarwal and P. Bhattacharyya, “Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified,” in Proceedings of the International Conference on Natural Language Processing (ICON), 2005.

[4] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification Using Machine Learning Techniques,” in emnlp2002, 2002, pp. 79-86.

[5] B. Yang, “Semi-supervised Learning for Sentiment Classification,” Learning, pp. 1-8.

[6] S. Owsley, S. Sood, and K. J. Hammond, “Domain Specific Affective Classification of Documents,” in AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), 2006, pp. 181-183.

[7] V. Ng, S. Dasgupta, and S. Arifin, “Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews,” in Proceedings of the COLING/ACL on Main conference poster sessions, 2006, vol. pp, pp. 611–618.

[8] A. Stepinski and V. Mittal, “A fact/opinion classifier for news articles,” in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 807-808.

[9] E. Boiy and M.-F. Moens, “A machine learning approach to sentiment analysis in multilingual Web texts,” Information Retrieval, vol. 12, no. 5, pp. 526-558, 2008.

[10] O. T\"ackström and R. McDonald, “Discovering fine-grained sentiment with latent variable structured prediction models,” Advances in Information Retrieval, pp. 368–374, 2011.

[11] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, “Target-dependent Twitter Sentiment Classification,” in Computational Linguistics, 2011, pp. 151-160.

[12] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, “Structured Models for Fine-to-Coarse Sentiment Analysis,” in Proceedings of the Association for Computational Linguistics (ACL), 2007, pp. 432-439.

[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning Word Vectors for Sentiment Analysis,” in Proceedings of the ACL, 2011.

[14] S. Somasundaran, G. Namata, L. Getoor, and J. Wiebe, “Opinion Graphs for Polarity and Discourse Classification,” in Proceedings of the 2009 Workshop on Graphbased Methods for Natural Language Processing TextGraphs4, 2009, pp. 66-74.

[15] O. Täckström, “Semi-supervised latent variable models for sentence-level sentiment analysis,” in Computational Linguistics, 2011, pp. 569-574.

[16] N. Yu and S. Kubler, “Semi-supervised Learning for Opinion Detection,” in 2010 IEEEWICACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, pp. 249-252.

[17] V. Sindhwani and P. Melville, “Document-Word Co-regularization for Semi-supervised Sentiment Analysis,” 2008 Eighth IEEE International Conference on Data Mining, pp. 1025-1030, 2008.

[18] A. B. Goldberg and J. Zhu, “Seeing stars when there aren’t many stars: Graph-based semi-supervised learning for sentiment categorization,” in TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing, 2006.

[19] S. Tan, G. Wu, H. Tang, and X. Cheng, “A novel scheme for domain-transfer problem in the context of sentiment analysis,” Proceedings of the sixteenth ACM conference on Conference on information and knowledge management CIKM 07, no. 2, p. 979, 2007.

[20] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, “Cross-domain sentiment classification via spectral feature alignment,” in Proceedings of the 19th international conference on World wide web WWW 10, 2010, p. 751.

[21] T. Li, V. Sindhwani, C. Ding, and Y. Zhang, “Knowledge transformation for cross-domain sentiment classification,” in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval SIGIR 09, 2009, pp. 716-717.

[22] B. Lu, C. Tan, C. Cardie, and B. K Tsou, “Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011, pp. 320-330.

[23] R. Mihalcea, C. Banea, and J. Wiebe, “Learning multilingual subjective language via cross-lingual projections,” in ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2007, vol. 45, no. 1, p. 976.

[24] N. Yu and S. Kübler, “Filling the Gap: Semi-Supervised Learning for Opinion Detection Across Domains,” in Proceedings of the Fifteenth Conference on Computational Natural Language Learning, 2011, pp. 200–209.

[25] S. Brody and N. Elhadad, “An Unsupervised Aspect-Sentiment Model for Online Reviews,” Human Language Technologies The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, no. June, pp. 804-812, 2010.

[26] T. Zagibalov and J. Carroll, “Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text,” Proceedings of the Conference on Computational Linguistics COLING 2008, no. August, pp. 1073-1080, 2008.

[27] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, no. July, pp. 417–424.

[28] S. Tan, Y. Wang, and X. Cheng, “Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples,” Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 08, p. 743, 2008.

[29] C. Lin and Y. He, “Joint sentiment/topic model for sentiment analysis,” in CIKM 09 Proceeding of the 18th ACM conference on Information and knowledge management, 2009, pp. 375-384.

[30] A. Kennedy and D, “Sentiment classification of movie reviews using contextual valence shifters,” Computational Intelligence, vol. 22, no. 2, pp. 110-125, 2006.

[31] C. Whitelaw, N. Garg, and S. Argamon, “Using appraisal groups for sentiment analysis,” in Proceedings of the 14th ACM international conference on Information and knowledge management CIKM 05, 2005, p. 625.

[32] J. Liu and S. Seneff, “Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm,” in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, no. August, pp. 161-169.

[33] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in Proceedings of the 7th Conference on Language Resources and Evaluation (LREC’10), 2010, vol. 25, pp. 2200-2204.

[34] P. J. Stone, The General Inquirer: A Computer Approach to Content Analysis. M.I.T. Press; First Edition edition (January 1, 1966), 1966, p. 651.

[35] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, “Opinion Extraction, Summarization and Tracking in News and Blog Corpora,” in AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), 2006, no. 2001, pp. 100-107.

[36] A. L. Berger, S. A. Della, V. J. Della, and V. J. D. Pietra, “A Maximum Entropy Approach to Natural Language Processing,” Computational Linguistics, no. 1992, pp. 1-36, 1996.

[37] S. Li, H. He, W. R. Xu, and J. Guo, “Automatic Chinese Sentiment Word Extraction Based on Maximum Entropy,” Proceedings of 2009 International Conference on Wavelet Analysis and Pattern Recognition, pp. 437-441 501, 2009.

[38] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, p. 27:1----27:27, 2011.

[39] L. Galavotti, F. Sebastiani, and M. Simi, “Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization Research and Advanced Technology for Digital Libraries,” vol. 1923, J. Borbinha and T. Baker, Eds. Springer Berlin / Heidelberg, 2000, pp. 59-68.

[40] A. Blum and S. Chawla, “Learning from Labeled and Unlabeled Data Using Graph Mincuts,” 2001, pp. 19-26.

1. **Foundation item:** Project(61170156) is supported by National Natural Science Foundation of China; Project(60933005) is supported by National Natural Science Foundation of China.

   **Received Date:** 0000－00－00; **Acceptd date:** 0000－00－00

   **Corresponding author:** XIE Song-xian, PhD; Tel: +86−731−84574627; Email:xsongx@nudt.edu.cn

   http://www.twitter.com [↑](#footnote-ref-1)
2. http://www.facebook.com [↑](#footnote-ref-2)
3. http://www.keenage.com/ [↑](#footnote-ref-3)
4. http://chengyu.teachercn.com [↑](#footnote-ref-4)
5. http://www.searchforum.org.cn/tansongbo/corpus/ [↑](#footnote-ref-5)
6. http://code.google.com/p/pymmseg-cpp/ [↑](#footnote-ref-6)
7. http://nltk.org [↑](#footnote-ref-7)