Uyghur Short Text Classification Using Morphological Information

Batuer Aisha

College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, 830046, China

batur61@163.com

Abstract. In this paper, we propose a novel method for improving the classification performance of short text strings using conditional random fields (CRFs) that combine morphological information. Experimental results on three datasets (Uyghur, Chinese, and English) demonstrate that our method can yield higher classification accuracy than Support Vector Machine (SVM) classifier and Maximum Entropy Model (MEM) classifier. Moreover, we show that our method can greatly decrease error rates, particularly if the number of training texts or the size of the strings in the train set is small.

Key words: Uyghur, Short text, Morphological information

1 Introduction

With the rapid growth of the available information on the Internet, text classification is becoming one of the key techniques in organizing, filtering and handling a large amount of text data. Since constructing text classifiers by hand is difficult and time consuming, it is desirable to learn classifiers from training data that do the category assignments automatically. Text classification task has provoked much interest in machine learning. As an aspect of text classification, short text classification is a particular challenge.

Uyghur is an agglutinative language with a rich and complex morphology, which is very different from the morphology of other languages, such as Chinese (Isolating language) and English (Inflecting language). Uyghur is written from right to left and words are separated by a blank space. Observe the following examples:

Sen Kitabxanigha Mang (You go to bookstore) سەن كىتابخانىغا ماڭ

The first and the third words (counted from right to left) are kept unchanged whereas the second word كىتابخانىغا (to bookstore) is decomposed into two parts, غا (inflectional suffix) and, كىتابخانى (corresponding to bookstore), a "quasi-word" that can never appear in the lexicon. After phonetic harmonization [1], the sentence becomes: سەن كىتابخانا غا ماڭ . Note that كىتابخانى is now changed to كىتابخانا (bookstore) which is a word in a lexicon.

It is worth noting that "to bookstore" is a word, or a phrase-like unit in Uyghur, "to" within it is a case-marker-like inflectional suffix. This means that the morphological information of Uyghur words in a sentence is associated with the classification features.

This classification problem is usually viewed as supervised learning, where the goal is to assign predefined category labels to unlabeled documents based on the likelihood inferred from the training set of labelled documents. The text categorization algorithms are to assign each test document set to one or more pre-specified classes [2].

Uyghur information society is facing the challenge of handling massive volume of online documents, news, and so on. The amount of data and increment rate is so high that this process cannot be done by hand. Hence, erotic recognition, filtering of spam mails, monitoring ill gossips and evil messages, quick search of interesting topics from large databases, and retrieving the information based on user's preferences from information sources, are some other examples where text classification can play an important role. If an automatic classification engine is developed, classification task can be achieved with less cost and in less time, while improving analyst's productivity. To use the Internet more efficiently, it needs to be classified. When we classify, seek to group things that have a common structure or exhibit a common behavior.

On the contrary of other languages, there is not much study on Uyghur texts. In this study, a system is mainly developed for automatic short text classification of Uyghur texts. The articles are classified into 7 different classes and 68% success ratio is achieved.

As there are no previous researches devoted to short text classification and being short of related works that we could use as a base for our research; in this paper, we compare our CRFs-based approach [3] with Support Vector Machine (SVM) [4] classifier and Maximum Entropy Model[1] (MEM) [5] classifier.

Text classification task has drawn a large body of research in machine learning community [6][7].

As a special aspect of text classification, short text classification is a particular challenge in that, short text examples tend to share few terms. It is particularly difficult to classify new instances and common comparisons between texts because they often yield no useful results. Short texts are typically sparse and ambiguous [8].

Nevertheless, to classify these short texts into certain target categories is a difficult but important problem. It is crucial in many information systems to classify short text segments, such as titles of documents and queries from users, into a well-formed topic hierarchy [9].

For example, in some time-consuming task and because of resource limitations, we deal with the title or the keywords of the document instead of the full text.

On the other hand, many search engines provide users with the ability to use natural language queries to ask questions and search for manually prepared answers. By using query classifications, queries in similar topics can automatically be clustered. Therefore, short text classification can make the preparation of answers to the queries more efficient.

[1] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

Although widely used, classification of short text snippets, such as titles and search queries, work poorly. Traditional document classification measures are often few, in any terms common between two short text snippets.

Short lengths text segments do not provide enough word or shared context for a good similarity measure. That is the reason why normal machine learning classification approaches usually fail to achieve desired accuracy.

Previous work attempts to overcome the data sparseness to get a better classification performance and deal with the feature sparse problem, external resources are used in short text classification task.

In [10, 11, 12, 13], search engines are employed in order to expand and enrich the context of data. In [14, 15, 16], online data repositories, such as Wikipedia or Open Directory Project, are used as external knowledge sources. With the help of external resource, we can improve the performance. However, the use of external resource, such as the search engine, is quite time-consuming, not suitable for real-time applications.

In order to speed up our short text classification approach, external resources are not used. Instead, we take advantage of the characteristics of short texts to improve the classification algorithm. The underlying idea is that features have strong relationship in very short texts.

The contributions of this paper are twofold. First, we investigate the change of words under different lengths of texts; second, we propose a novel CRFs-based method to improve the short text classification performance.

The remainder of the paper is organized as follows. Section 2 provides our analysis about short text. Based on our analysis, a CRFs-based short text classification method is proposed in section 3. In Section 4, our experimental results are demonstrated. There are different kinds of short text classification applications, so in our experiment section we conduct experiments on different applications to get a comprehensive evaluation. The last section summarizes our contribution and outlines future goals.

2 Consistency of features in very short texts

In this section, we want to investigate the characteristics of short texts. The adjacent features of short texts tend to talk about the same topic. On the other hand, the topic might be changed along with the increase of text length. We will analyze these issues as follows.

2.1 Analysis of short texts

Text segments do not provide enough features in common in short length, so they suffer from the feature sparse problem.

Nonetheless, short texts have their advantage for classification task. Many short texts focus on one topic because of their short lengths. But manual labelling text may be too short to learn new features, and too short to obtain proper word statistics.

Instead, long texts usually contain segments that belong to different topics. For example, a full text about sports may have a segment talking about football, and another segment talking about history. In short texts, this kind of problem seldom happens, because short texts focus on one topic. The features of short texts have strong relationship. If a feature is related to a topic, the features in the near contexts also incline to this topic. In this work, we use tokenization to indicate this kind of characteristic.

In [17], the author proposes a text categorization method in which documents are split into fragments. And instead of classifying the full text, they classify the segments and use the segment category to yield the result. This method can yield some improvements of text classification performance. This work also gives us the clue that text segments have special characteristics. If properly used, it can help us to improve classification accuracy.

3 CRFs-based Classifier

Based on our observation in the previous section that short texts focus on one topic, and that the features of short texts have strong relationship; and that when a feature is related to a topic, the features in the near contexts also incline to this topic. In other words, features in very short texts are quite consistent. It is expected therefore that this kind of characteristic can be used properly, and we can improve the short text classification performance. CRFs can be used to add constraints among features.

CRFs are undirected graphical models trained to maximize a conditional probability first introduced by Lafferty, such as natural language processing to the indexing of the string to learning tasks.

However, CRFs cannot be used in short text classification directly, because it is a sequence labelling algorithm. Therefore, we need to convert short text classification problem to sequence labelling problem. In this section, we will propose our CRFs-based classifier. We borrow the character tagging approach, which is widely used in Chinese word segmentation task [18][19], to covert our classification problem to a sequence problem, so that CRFs can be used in our approach. Finally, we propose our novel short text classification method based on CRFs.

3.1 Proposed algorithm

In this subsection, we borrow the tagging approach to reformulate short text classification task as a sequence labelling problem.

This approach is used in both train and test step. Nevertheless, our algorithm tags every character with the category of the short text. We will use an example to illustrate our proposed algorithm.

Assume we have such a short text in train set,

Uyghur characters: ئۈچته ياخشى ئوقۇغۇچىلار تەقدىرلەندى

Uyghur Latin characters: Üchte Yaxshi Oqughuchilar Teqdirlendi

("Three good" student was praised).

This text belongs to Education class.

Uyghur characters: بۇ يىل ئۆي باھاسى ئىككى پىرسەنت ئۆستى

Uyghur Latin characters: Buyil Öy Baxasi Ikki Pirsent Östi

(House prices rose two percent in this year).

This text belongs to Business class.

Uyghur characters: خەلقئارا سەھىيە تەشكىلاتى زۇكامغا قارشى ۋاكسىنىنى نامرات دۆلەتلەرگە ئەۋەتتى

Uyghur Latin characters: Xeliqara Sehiye Teshkilati Zukamgha Qarshi Waksinini Namrat Döletlerge Ewetti

(WHO send swine flu vaccine to poor countries).

This text belongs to Health class.

Uyghur characters: چېگرادىن چىقىرىلغان ئۆزبەكلەر قىرغىزتانغا قايتۇرلىشتىن ئەنسىرەيدىغانلىغىنى بىرلەشكەن دۆلەتلەرتەشكىلاتىغا خەۋەر قىلدى

Uyghur Latin characters: CHëgradin CHiqirilghan Özbekler Qirghiz-tangha Qayturlishtin Ensireydhghanlighini Birleshken Döletler Teshkilatigha Xewer Qildi

(The deported Uzbek people told the UN fears returned to Kyrgyz-stan).

This text belongs to Politics class.

Uyghur characters: بارسېلۇنالىق مېسسىي بۇ يىل دۇنيا توپ چولپىنى دېگەن نامغا ئېرىشتى

Uyghur Latin characters: Barsëlunaliq Mëssiy Bu Yil Dunya Top CHolpini Dëgen Namgha Ërishti

(Barcelona's Messy named World Player of the Year).

This text belongs to Sports class.

Uyghur characters: جېكسوننىڭ ئۆلىمى دۇنيانى تەۋرەتتى

Uyghur Latin characters: Jëksonning Ölimi Dunyani Tewretti

(Jackson's death shocked all over the world).

This text belongs to Play class.

If we turn the short texts with category to labeled sequence, CRFs algorithm can be used to train models, which can be used in test step.

Assuming we have such a short text in test set,

Uyghur characters: ئىككى تالىبان ئافغانستاندا ئۆلدى

Uyghur Latin characters: Ikki Taliban Afghanistanda Öldi

(Two Taliban dead in Afghan city gun battle).

This text belongs to Military class. This short text is turned into a labeled sequence as shown in Fig. 1.

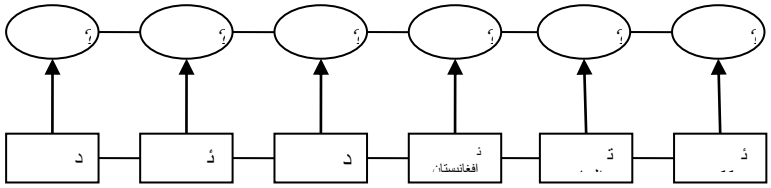

Fig. 1. Graphical structures of linear chain

In Fig. 1, square nodes indicate characters and circle nodes indicate tags according to the characters. We will use CRFs algorithm to infer its category.

We use the CRFs model obtained in the train step to infer the category of short text in test set.

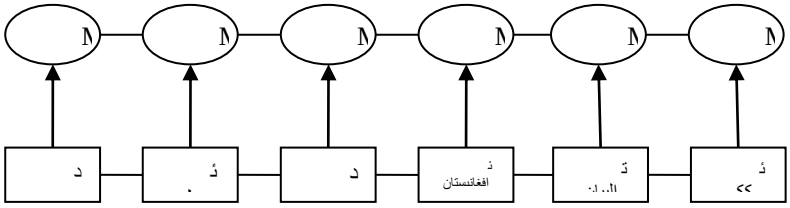

Fig. 2. Graphical structures of linear chain

In Figure 2, we get the category of the short text. This short text belongs to "Military" class.

Our analysis shows that tokenization changes under different lengths of texts, has a slight negative correlation with the length.

The problem of short text classification can be formally stated as follows: Given a sequence of token $w_i \ldots w_n$, we want to find the corresponding sequence of classification tags $t_1 \ldots t_n$, drawn from a set of tags T, which satisfies:

$$P(t_1 \ldots t_n | w_1 \ldots w_n) = \prod_{i=1}^{n} P(ti \mid Ci) \quad (1)$$

where Ci is the context for token w_i.

The best feature templates used in our short text classification experiments:

Unigram features:

U_{i-1}: previous word in sentence

U_i: current word in sentence

U_{i+1}: next word in sentence

Bigram features:

$U_{i-1}U_i$, U_iU_{i+1}

We incorporate morphological information in Uyghur by using a morpheme analyzer [20].The features of very short texts tend to be consistent. From the graphical

structure of linear chain CRFs, we can find constraints among the hidden variables. In other words, there are constraints among the topics. This method calculates similarities between two strings (e.g. texts or sequences) by matching the common substring in the strings. Therefore, our CRFs-based approach is expected to yield promising performance.

The main idea of the algorithm is to select feature words from each document; those words cover all the ideas in the document. The results of this algorithm are a list of the main subjects founded in the document. Also, in this paper the effects of the Uyghur text classification on Information Retrieval have been investigated. The goal was to improve the convenience and effectiveness of information access. We will quantitatively evaluate our method in the next section.

4 Experiments

In the data preparation step, only long titles are eliminated, noisy text such as stop words are still in our dataset. Therefore, the accuracies of both classifiers are relatively low. Nevertheless, our CRFs-based approach still outperforms SVM classifier. Three different datasets are used to evaluate our CRFs based approach.

Uyghur datasets compose titles of web documents. Sogou[2] short text corpora are Chinese datasets, which compose of titles of web documents. Ohsumed-all dataset is an English corpus, which consists of medical abstracts.

In our experiments, unigram and bigram features are used. On the other hand, we focus on very short texts. Therefore, the lengths of all the texts we used are less than twelve. It is just an empirical value, because there is no clear definition of "very short text". In Uyghur texts, the limit is twelve token and In Chinese texts, the limit is twenty characters, while in English texts, and the limit is ten words.

In the next subsections, we compare our CRFs-based approach with SVM classifier in the evaluation forms. Libsvm[3] is used as our basic classifier as it has been proved to be effective on many machine learning tasks especially text classification. We use probability SVM models in this paper.

4.1 Uyghur Short Text Classification

Uyghur Energy（UE）corpus is a relatively small dataset. Therefore, we conduct experiments on it to verify our results. UE short text corpora are used in this subsection.

Uyghur titles of web documents corpus has seven subject categories and 59,992 single-labeled documents. We use the titles of web documents as a short text corpus. All the short texts are used in seven class classification experiments. "sports", "military", "play", "health", "politics", "education", "business" classes are used in our experiments. The whole corpus is randomly divided into train set and test set with different proportion.

The classifying accuracy and train time (second) comparisons according to different proportions of train set and test set (7.5:2.5, 5:5, 3:7) are shown Figure3-5 (Table 1-3) and Table 4-5 as follows:

[2]http://www.sogou.com/labs/resources.html

[3]http://www.csie.ntu.edu.tw/~cjlin/libsvm

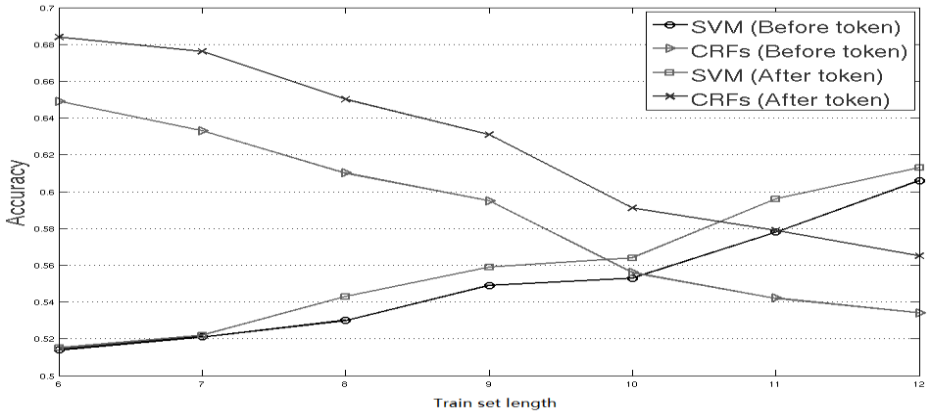

Fig. 3. The classifying accuracy comparisons on UE corpus (train: test=7.5:2.5)

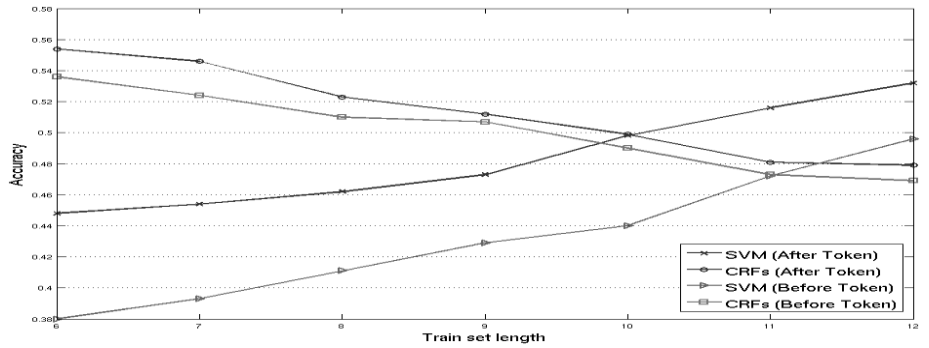

Fig. 4. The classifying accuracy comparisons on UE corpus (train: test=5:5)

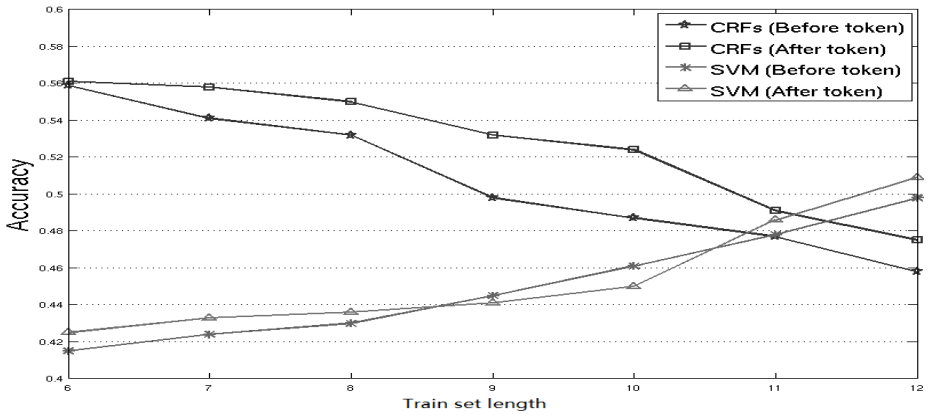

Fig. 5. The classifying accuracy comparisons on UE corpus (train: test=3:7)

Table 1: The classifying accuracy on UE corpus (train: test=7.5:2.5)

Train set length	MEM (Before token)	MEM (After token)
6	51.22	54.84
12	66.83	67.79

Table 2: The classifying accuracy on UE corpus (train: test=5:5)

Train set length	MEM (Before token)	MEM (After token)
6	42.11	41.27
12	54.78	53.96

Table 3: The classifying accuracy on UE corpus (train: test=3:7)

Train set length	MEM (Before token)	MEM (After token)
6	40.58	43.55
12	54.09	56.10

Table 4: The train time comparisons on UE corpus (train: test=7.5:2.5)

Before Token		
Train set length	SVM	CRFs
6	14100	2400
12	16140	4260
After Token		
Train set length	SVM	CRFs
6	14820	8760
12	16440	9900

Table 5: The train time comparisons on UE corpus (train: test=5:5)

Before Token		
Train set length	SVM	CRFs
6	5400	1740
12	5880	3000
After Token		
Train set length	SVM	CRFs
6	5580	4080
12	6240	5880

Experimental results in the tables above demonstrate our CRFs-based approach can yield promising performances. Two different datasets are used to evaluate our CRFs based approach.

UE and UE-tokenized short text corpora are Uyghur datasets, which are compose of titles of web documents.

On the experiments on UE corpus, our proposed CRFs-based approach outperforms SVM classifier and MEM classifier.

4.2 Chinese short text classification

Sogou web documents corpus is a public dataset. We use the titles of web document as a short text corpus. After eliminating long titles, we finally obtain 14 subject categories (sports, military, play, health, politics, education, business, women, culture, house, news, information, education, travel, and auto) and 40894 single-labeled documents.

Previous work [21] shows that Chinese character bigram has better performance than Chinese word unit. Besides, we don't need to take Chinese word segmentation into consideration. The classifying accuracy comparisons according to different proportions of train set and test set (7:3) are shown Table 6 as follows:

Table 6: The classifying accuracy comparisons on Sogou corpus

Train set length (characters)	SVM	CRFs
10	0.693	0.745
20	0.791	0.801

Experimental results in the tables above demonstrate our CRFs-based approach still outperforms SVM classifier. Our method is also valid on Chinese datasets.

4.3 English short text classification

We also conduct experiments on English corpora. Ohsumed (MEDLINE) is used in this subsection. Ohsumed-all dataset is composed of 50216 medical abstracts classified into 23 categories (C1,C2…C23). After eliminating long titles (longer than ten words), we finally obtain 28399 very short texts in 23 different classes. We conduct experiments on 4 class classification and 5 class classification tasks.

Table 7: Performance comparisons on Ohsumed corpus (four class classification)

	SVM	CRFs
C1~C4	0.560	0.627
C5~C8	0.457	0.494
C9~C12	0.501	0.564
C13~C16	0.590	0.619
C17~C20	0.432	0.475

Table 8: Performance comparisons on Ohsumed-all corpus (five class classification)

	SVM	CRFs
C1~C5	0.501	0.558
C6~C10	0.431	0.479
C11~C15	0.501	0.531
C16~C20	0.382	0.421

In the data preparation step, only long titles are eliminated, noisy text such as stop words are still in our dataset.

Therefore, the accuracies of both classifiers are relatively low. Nevertheless, our CRFs-based approach still outperforms SVM classifier. Our method is also valid on English datasets.

Experiment results demonstrate that our CRFs-based approach can yield promising performances. CRFs-based short text classification approach outperforms SVM classifier on all datasets.

5 Conclusion

In this paper, we presented a brief overview of the text classification task. We applied two supervised learning algorithms, SVM and CRFs, for Uyghur, Chinese and English short text classification. We also compared their performance with different proportion of train set and test set.

This paper presents enhanced, effective and simple approach to short text classification. The approach uses an algorithm to automatically classify documents.
The main idea of our work is that features have strong relationship in very short texts. The features of very short texts tend to be consistent. On the other hand, from the graphical structure of linear chain CRFs, we can find constraints among the hidden variables. In other words, there are constraints among the topics. Therefore, our CRFs-based approach is expected to yield promising performance.

Three different datasets are used to evaluate our CRFs based approach and that both datasets are used in our evaluation section.
Our method can greatly decrease error rates, particularly if the number of examples or the size of the strings in the training set is small. All experiment results demonstrate that our proposed algorithm can yield stable and significant improvements on different kinds of classification tasks. Therefore, our approach can also be used in regular text classification task.

In comparison with large train languages, such as Chinese and English, our system has less performance. This could be caused by the limited number of train words which are not enough to cover the UE. But our research will be very useful in the development of Turkic language families (include Turkish, Azeri, Uzbek, Kazakh, Turkmen, Tatar, Kyrgyz and others) text classification and other related research. How to use it properly will be our future work.

In future, we plan to conduct more experiments on SVM and CRFs with multi-class documents, which is a large number of single class documents (include micro blogs) and multi label documents.

Acknowledgements

We would like to thank anonymous reviewers for their useful comments and suggestions. This work is supported by Xinjiang University doctoral project.

References

1. Batuer Aisha , Maosong Sun.: A Statistical Method for Uyghur Tokenization, In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2009,pp.383-387.
2. Eibe Frank and Remco R. Bouckaert.: Naïve Bayes for text classification with unbalanced classes. In Proc. 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer, Berlin, Germany, 2006, pp.503-510.
3. John Lafferty, Andrew McCallum, and Fernando Pereira.: Conditional random fields: probabilistic models for segmenting and labeling sequence data, In Proceedings of ICML, 2001, pp. 591-598.
4. T. Joachims.: Text categorization with Support Vector Machines: Learning with many relevant features, In Proceedings of the 10th European Conference on Machine Learning, 1998, pp. 137-142.
5. A. Berger, S. Della Pietra, and V. Della Pietra.: A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22(1):39-71.
6. F.Sebastiani.: Machine learning in automated text categorization, ACM Computing Surveys, 2002, 34(1):1- 47 .
7. Yiming Yang, Xin Liu.: A re-examination of text categorization methods, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States, August 15-19, 1999, pp.42-49 .
8. Zelikovitz, S. and Hirsh, H.: Improving short-text classification using unlabeled background knowledge to assess document similarity, Proceedings of the Seventeenth International Conference on Machine Learning (2000) .
9. Shen, D., Pan, R., Sun, J.-T., Pan, J. J.;Wu, K., Yin, J., and Yang, Q.: Query enrichment for web-query classification, ACM Trans. Inf. Syst. 2006, 24(3):320-352.
10. M. Sahami and T. Heilman.: A Web based kernel function for measuring the similarity of short text snippets, In Proc. WWW (2006).
11. D. Bollegala, Y. Matsuo, and M. Ishizuka.: Measuring semantic similarity between words using Web search engines, Proc. WWW (2007).
12. W. Yih and C. Meek.: Improving similarity measures for short segments of text, In Proc. AAAI (2007).
13. Huanhuan Cao, Derek Hao Hu and Dou Shen et al.:Context-Aware Query Classification, Proc. ACM SIGIR (2009).
14. JS. Banerjee, K. Ramanathan, and A. Gupta.: Clustering short texts using Wikipedia, Proc. ACM SIGIR (2007).
15. X. Phan, L. Nguyen, and S. Horiguchi.: Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large-scale Data Collections, In Proc. WWW (2008).
16. P. Schonhofen.: Identifying document topics using the Wikipedia category network, Proc.

the IEEE/WIC/ACM International Conference on Web Intelligence, (2006).
17. Jan Blaˇt´ak, Eva Mr´akov´a and Luboˇs Popel´ınsk´.: Fragments and Text Categorization, Proceedings of the ACL 2004.
18. Fuchun Peng, Fangfang Feng, and Andrew McCallum.: Chinese segmentation and new word detection using conditional random fields, In Proceedings of the 20th international conference on Computational Linguistics, 2004,pp. 562-568.
19. Xue. Nianwen: Chinese word segmentation as character tagging, In Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48.
20. Batuer Aisha, Maosong Sun, A Uyghur Morpheme Analysis Method based on Conditional Random Fields, International Journal on Asian Language Processing, 19(2),2009,69-77.
21. J.Y. Li, Mao song Sun, Xian Zhang.: A Comparison and Semi-Quantitative Analysis of Words and Character-Bigrams as Features in Chinese Text Categorization, In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, 2006, pp. 545-552.