A Comparison and Semi-Quantitative Analysis of Words and Character-Bigrams as Features in Chinese Text Categorization

Jingyang Li **Maosong Sun** **Xian Zhang**
National Lab. of Intelligent Technology & Systems, Department of Computer Sci. & Tech.
Tsinghua University, Beijing 100084, China
lijingyang@gmail.com sms@tsinghua.edu.cn kevinn9@gmail.com

Abstract

Words and character-bigrams are both used as features in Chinese text processing tasks, but no systematic comparison or analysis of their values as features for Chinese text categorization has been reported heretofore. We carry out here a full performance comparison between them by experiments on various document collections (including a manually word-segmented corpus as a golden standard), and a semi-quantitative analysis to elucidate the characteristics of their behavior; and try to provide some preliminary clue for feature term choice (in most cases, character-bigrams are better than words) and dimensionality setting in text categorization systems.

1 Introduction[1]

Because of the popularity of the Vector Space Model (VSM) in text information processing, document indexing (term extraction) acts as a pre-requisite step in most text information processing tasks such as Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1999) and Text Categorization (Sebastiani, 2002). It is empirically known that the indexing scheme is a non-trivial complication to system performance, especially for some Asian languages in which there are no explicit word margins and even no natural semantic unit. Concretely, in Chinese Text Categorization tasks, the two most important indexing units (feature terms) are word and character-bigram, so the problem is: which kind of terms[2] should be chosen as the feature terms, words or character-bigrams?

To obtain an all-sided idea about feature choice beforehand, we review here the possible feature variants (or, options). First, at the word level, we can do stemming, do stop-word pruning, include POS (Part of Speech) information, etc. Second, term combinations (such as "word-bigram", "word + word-bigram", "character-bigram + character-trigram"[3], etc.) can also be used as features (Nie et al., 2000). But, for Chinese Text Categorization, the "word or bigram" question is fundamental. They have quite different characteristics (e.g. bigrams overlap each other in text, but words do not) and influence the classification performance in different ways.

In Information Retrieval, it is reported that bigram indexing schemes outperforms word schemes to some or little extent (Luk and Kwok, 1997; Leong and Zhou 1998; Nie et al., 2000). Few similar comparative studies have been reported for Text Categorization (Li et al., 2003) so far in literature.

Text categorization and Information Retrieval are tasks that sometimes share identical aspects (Sebastiani, 2002) apart from term extraction (document indexing), such as *tfidf* term weighting and performance evaluation. Nevertheless, they are different tasks. One of the generally accepted connections between Information Retrieval and Text Categorization is that an information retrieval task could be partially taken as a binary classification problem with the query as the only positive training document. From this

[1] This research is supported by the National Natural Science Foundation of China under grant number 60573187 and 60321002, and the Tsinghua-ALVIS Project co-sponsored by the National Natural Science Foundation of China under grant number 60520130299 and EU FP6.

[2] The terminology "term" stands for both word and character-bigram. Term or combination of terms (in word-bigram or other forms) might be chosen as "feature".

[3] The terminology "character" stands for Chinese character, and "bigram" stands for character-bigram in this paper.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 545–552,
Sydney, July 2006. ©2006 Association for Computational Linguistics

viewpoint, an IR task and a general TC task have a large difference in granularity. To better illustrate this difference, an example is present here. The words "制片人(film producer)" and "译制片(dubbed film)" should be taken as different terms in an IR task because a document with one would not necessarily be a good match for a query with the other, so the bigram "制片(film production)" is semantically not a shared part of these two words, i.e. not an appropriate feature term. But in a Text Categorization task, both words might have a similar meaning at the category level ("film" category, generally), which enables us to regard the bigram "制片" as a semantically acceptable representative word snippet for them, or for the category.

There are also differences in some other aspects of IR and TC. So it is significant to make a detailed comparison and analysis here on the relative value of words and bigrams as features in Text Categorization. The organization of this paper is as follows: Section 2 shows some experiments on different document collections to observe the common trends in the performance curves of the word-scheme and bigram-scheme; Section 3 qualitatively analyses these trends; Section 4 makes some statistical analysis to corroborate the issues addressed in Section 3; Section 5 summarizes the results and concludes.

2 Performance Comparison

Three document collections in Chinese language are used in this study.

The electronic version of *Chinese Encyclopedia* ("CE"): It has 55 subject categories and 71674 single-labeled documents (entries). It is randomly split by a proportion of 9:1 into a training set with 64533 documents and a test set with 7141 documents. Every document has the full-text. This data collection does not have much of a sparseness problem.

The training data from a national Chinese text categorization evaluation[4] ("CTC"): It has 36 subject categories and 3600 single-labeled[5] documents. It is randomly split by a proportion of 4:1 into a training set with 2800 documents and a test set with 720 documents. Documents in this data collection are from various sources including news websites, and some documents may be very short. This data collection has a moderate sparseness problem.

A manually word-segmented corpus from the State Language Affairs Commission ("LC"): It has more than 100 categories and more than 20000 single-labeled documents[6]. In this study, we choose a subset of 12 categories with the most documents (totally 2022 documents). It is randomly split by a proportion of 2:1 into a training set and a test set. Every document has the full-text and has been entirely word-segmented[7] by hand (which could be regarded as a golden standard of segmentation).

All experiments in this study are carried out at various feature space dimensionalities to show the scalability. Classifiers used in this study are Rocchio and SVM. All experiments here are multi-class tasks and each document is assigned a single category label.

The outline of this section is as follows: Subsection 2.1 shows experiments based on the Rocchio classifier, feature selection schemes besides *Chi* and term weighting schemes besides *tfidf* to compare the automatic segmented word features with bigram features on CE and CTC, and both document collections lead to similar behaviors; Subsection 2.2 shows experiments on CE by a SVM classifier, in which, unlike with the Rocchio method, *Chi* feature selection scheme and *tfidf* term weighting scheme outperform other schemes; Subsection 2.3 shows experiments by a SVM classifier with *Chi* feature selection and *tfidf* term weighting on LC (manual word segmentation) to compare the best word features with bigram features.

2.1 The Rocchio Method and Various Settings

The Rocchio method is rooted in the IR tradition, and is very different from machine learning ones (such as SVM) (Joachims, 1997; Sebastiani, 2002). Therefore, we choose it here as one of the representative classifiers to be examined. In the experiment, the control parameter of negative examples is set to 0, so this Rocchio based classifier is in fact a centroid-based classifier.

$Chi_{\max}$ is a state-of-the-art feature selection criterion for dimensionality reduction (Yang and Peterson, 1997; Rogati and Yang, 2002). $Chi_{\max*}CIG$ (Xue and Sun, 2003a) is reported to be better in Chinese text categorization by a cen-

[4] The Annual Evaluation of Chinese Text Categorization 2004, by 863 National Natural Science Foundation.

[5] In the original document collection, a document might have a secondary category label. In this study, only the primary category label is reserved.

[6] Not completed.

[7] And POS (part-of-speech) tagged as well. But POS tags are not used in this study.

troid based classifier, so we choose it as another representative feature selection criterion besides $Chi_{\max}$.

Likewise, as for term weighting schemes, in addition to *tfidf*, the state of the art (Baeza-Yates and Ribeiro-Neto, 1999), we also choose *tfidf*CIG* (Xue and Sun, 2003b).

Two word segmentation schemes are used for the word-indexing of documents. One is the m*aximum match* algorithm ("mmword" in the figures), which is a representative of simple and fast word segmentation algorithms. The other is ICTCLAS[8] ("lqword" in the figures). ICTCLAS is one of the best word segmentation systems (SIGHAN 2003) and reaches a segmentation precision of more than 97%, so we choose it as a representative of state-of-the-art schemes for automatic word-indexing of document).

For evaluation of single-label classifications, F_1-measure, *precision*, *recall* and *accuracy* (Baeza-Yates and Ribeiro-Neto, 1999; Sebastiani, 2002) have the same value by microaveraging[9], and are labeled with "performance" in the following figures.

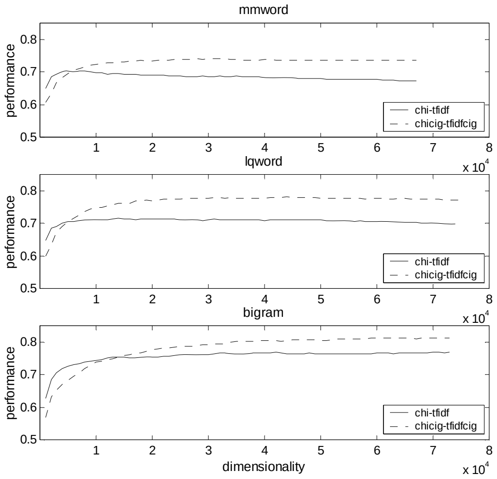

Figure 1. *chi-tfidf* and *chicig-tfidfcig* on CE

Figure 1 shows the performance-dimensionality curves of the *chi-tfidf* approach and the approach with *CIG*, by *mmword*, *lqword* and *bigram* document indexing, on the CE document collection. We can see that the original *chi-tfidf* approach is better at low dimensionalities (less than 10000 dimensions), while the *CIG* version is better at high dimensionalities and reaches a higher limit.[10]

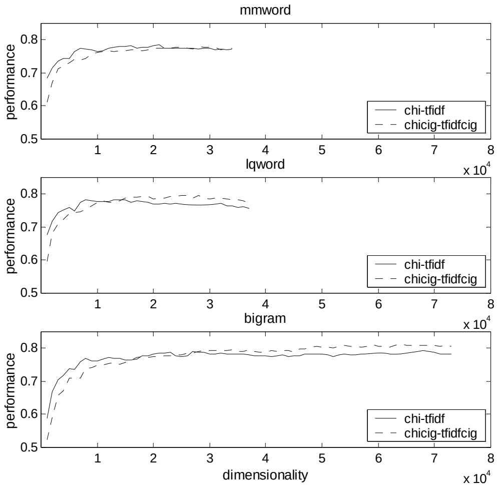

Figure 2. *chi-tfidf* and *chicig-tfidfcig* on CTC

Figure 2 shows the same group of curves for the CTC document collection. The curves fluctuate more than the curves for the CE collection because of sparseness; The CE collection is more sensitive to the additions of terms that come with the increase of dimensionality. The CE curves in the following figures show similar fluctuations for the same reason.

For a parallel comparison among *mmword*, *lqword* and *bigram* schemes, the curves in Figure 1 and Figure 2 are regrouped and shown in Figure 3 and Figure 4.

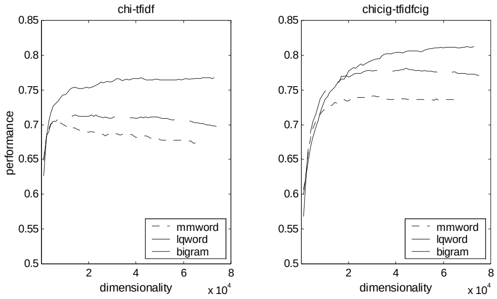

Figure 3. *mmword*, *lqword* and *bigram* on CE

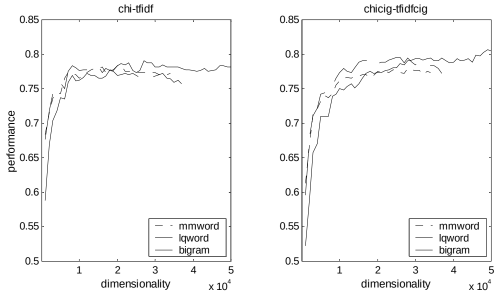

Figure 4. *mmword*, *lqword* and *bigram* on CTC

[8] http://www.nlp.org.cn/project/project.php?proj_id=6

[9] Microaveraging is more prefered in most cases than macroaveraging (Sebastiani 2002).

[10] In all figures in this paper, curves might be truncated due to the large scale of dimensionality, especially the curves of bigram scheme. For these kinds of figures, at least one of the following is satisfied: (a) every curve has shown its zenith; (b) only one curve is not complete and has shown a higher zenith than other curves; (c) a margin line is shown to indicate the limit of the incomplete curve.

We can see that the *lqword* scheme outperforms the *mmword* scheme at almost any dimensionality, which means the more precise the word segmentation the better the classification performance. At the same time, the *bigram* scheme outperforms both of the word schemes on a high dimensionality, wherea the word schemes might outperform the *bigram* scheme on a low dimensionality.

Till now, the experiments on CE and CTC show the same characteristics despite the performance fluctuation on CTC caused by sparseness. Hence in the next subsections CE is used instead of both of them because its curves are smoother.

2.2 SVM on Words and Bigrams

As stated in the previous subsection, the *lqword* scheme always outperforms the *mmword* scheme; we compare here only the *lqword* scheme with the *bigram* scheme.

Support Vector Machine (SVM) is one of the best classifiers at present (Vapnik, 1995; Joachims, 1998), so we choose it as the main classifier in this study. The SVM implementation used here is LIBSVM (Chang, 2001); the type of SVM is set to "C-SVC" and the kernel type is set to linear, which means a one-with-one scheme is used in the multi-class classification.

Because the *CIG*'s effectiveness on a SVM classifier is not examined in Xue and Sun (2003a, 2003b)'s report, we make here the four combinations of schemes with and without *CIG* in feature selection and term weighting. The experiment results are shown in Figure 5. The collection used is CE.

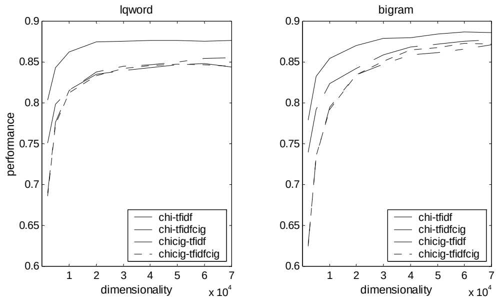

Figure 5. *chi-tfidf* and *cig*-involved approaches on *lqword* and *bigram*

Here we find that the *chi-tfidf* combination outperforms any approach with *CIG*, which is the opposite of the results with the Rocchio method. And the results with SVM are all better than the results with the Rocchio method. So we find that the feature selection scheme and the term weighting scheme are related to the classifier, which is worth noting. In other words, no feature selection scheme or term weighting scheme is absolutely the best for all classifiers. Therefore, a reasonable choice is to select the best performing combination of feature selection scheme, term weighting scheme and classifier, i.e. *chi-tfidf* and SVM. The curves for the *lqword* scheme and the *bigram* scheme are redrawn in Figure 6 to make them clearer.

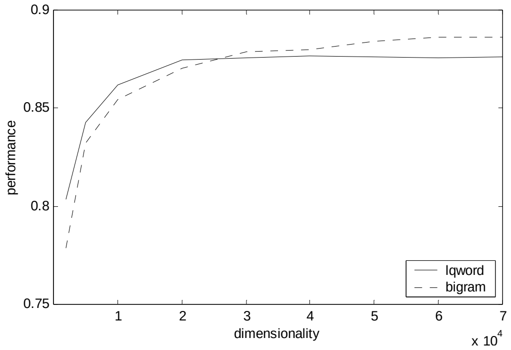

Figure 6. *lqword* and *bigram* on CE

The curves shown in Figure 6 are similar to those in Figure 3. The differences are: (a) a larger dimensionality is needed for the *bigram* scheme to start outperforming the *lqword* scheme; (b) the two schemes have a smaller performance gap.

The *lqword* scheme reaches its top performance at a dimensionality of around 40000, and the *bigram* scheme reaches its top performance at a dimensionality of around 60000 to 70000, after which both schemes' performances slowly decrease. The reason is that the low ranked terms in feature selection are in fact noise and do not help to classification, which is why the feature selection phase is necessary.

2.3 Comparing Manually Segmented Words and Bigrams

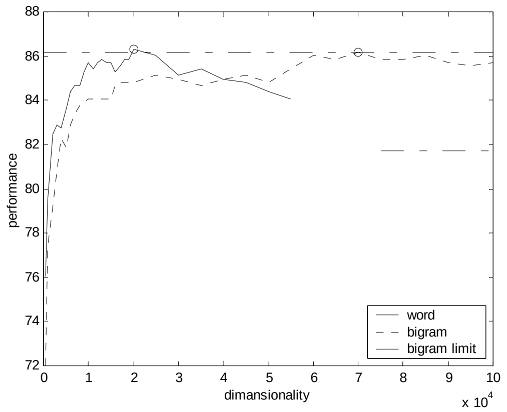

Figure 7. *word* and *bigram* on LC

Up to now, bigram features seem to be better than word ones for fairly large dimensionalities. But it appears that word segmentation precision impacts classification performance. So we choose here a fully manually segmented document collection to detect the best performance a word scheme could reach and compare it with the bigram scheme.

Figure 7 shows such an experiment result on the LC document collection (the circles indicate the maximums and the dash-dot lines indicate the superior limit and the asymptotic interior limit of the bigram scheme). The word scheme reaches a top performance around the dimensionality of 20000, which is a little higher than the bigram scheme's zenith around 70000.

Besides this experiment on 12 categories of the LC document collection, some experiments on fewer (2 to 6) categories of this subset were also done, and showed similar behaviors. The word scheme shows a better performance than the bigram scheme and needs a much lower dimensionality. The simpler the classification task is, the more distinct this behavior is.

3 Qualitative Analysis

To analyze the performance of words and bigrams as feature terms in Chinese text categorization, we need to investigate two aspects as follows.

3.1 An Individual Feature Perspective

The word is a natural semantic unit in Chinese language and expresses a complete meaning in text. The bigram is not a natural semantic unit and might not express a complete meaning in text, but there are also reasons for the bigram to be a good feature term.

First, two-character words and three-character words account for most of all multi-character Chinese words (Liu and Liang, 1986). A two-character word can be substituted by the same bigram. At the granularity of most categorization tasks, a three-character words can often be substituted by one of its sub-bigrams (namely the "intraword bigram" in the next section) without a change of meaning. For instance, "标赛" is a sub-bigram of the word "锦标赛(tournament)" and could represent it without ambiguity.

Second, a bigram may overlap on two successive words (namely the "interword bigram" in the next section), and thus to some extent fills the role of a word-bigram. The word-bigram as a more definite (although more sparse) feature surely helps the classification. For instance, "气预" is a bigram overlapping on the two successive words "天气(weather)" and "预报(forecast)", and could almost replace the word-bigram (also a phrase) "天气预报(weather forecast)", which is more likely to be a representative feature of the category "气象学(meteorology)" than either word.

Third, due to the first issue, bigram features have some capability of identifying OOV (out-of-vocabulary) words[11], and help improve the *recall* of classification.

The above issues state the advantages of bigrams compared with words. But in the first and second issue, the equivalence between bigram and word or word-bigram is not perfect. For instance, the word "文学(literature)" is a also sub-bigram of the word "天文学(astronomy)", but their meanings are completely different. So the loss and distortion of semantic information is a disadvantage of bigram features over word features.

Furthermore, one-character words cover about 7% of words and more than 30% of word occurrences in the Chinese language; they are effevtive in the word scheme and are not involved in the above issues. Note that the impact of effective one-character words on the classification is not as large as their total frequency, because the high frequency ones are often too common to have a good classification power, for instance, the word "的 (of, 's)".

3.2 A Mass Feature Perspective

Features are not independently acting in text classification. They are assembled together to constitute a feature space. Except for a few models such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990), most models assume the feature space to be orthogonal. This assumption might not affect the effectiveness of the models, but the semantic redundancy and complementation among the feature terms do impact on the classification efficiency at a given dimensionality.

According to the first issue addressed in the previous subsection, a bigram might cover for more than one word. For instance, the bigram "织物" is a sub-bigram of the words "织物(fabric)", "棉织物(cotton fabric)", "针织物(knitted fabric)", and also a good substitute of

[11] The "OOV words" in this paper stand for the words that occur in the test documents but not in the training document.

them. So, to a certain extent, word features are redundant with regard to the bigram features associated to them. Similarly, according to the second issue addressed, a bigram might cover for more than one word-bigram. For instance, the bigram “篇小” is a sub-bigram of the word-bigrams (phrases) “短篇小说(short story)”, “中篇小说(novelette)”, “长篇小说(novel)” and also a good substitute for them. So, as an addition to the second issue stated in the previous subsection, a bigram feature might even cover for more than one word-bigram.

On the other hand, bigrams features are also redundant with regard to word features associated with them. For instance, the “锦标” and “标赛” are both sub-bigrams of the previously mentioned word “锦标赛”. In some cases, more than one sub-bigram can be a good representative of a word.

We make a word list and a bigram list sorted by the feature selection criterion in a descending order. We now try to find how the relative redundancy degrees of the word list and the bigram list vary with the dimensionality. Following issues are elicited by an observation on the two lists (not shown here due to space limitations).

The relative redundancy rate in the word list keeps even while the dimensionality varies to a certain extent, because words that share a common sub-bigram might not have similar statistics and thus be scattered in the word feature list. Note that these words are possibly ranked lower in the list than the sub-bigram because feature selection criteria (such as *Chi*) often prefer higher frequency terms to lower frequency ones, and every word containing the bigram certainly has a lower frequency than the bigram itself.

The relative redundancy in the bigram list might be not as even as in the word list. Good (representative) sub-bigrams of a word are quite likely to be ranked close to the word itself. For instance, “作曲” and “曲家” are sub-bigrams of the word “作曲家(music composer)”, both the bigrams and the word are on the top of the lists. Theretofore, the bigram list has a relatively large redundancy rate at low dimensionalities. The redundancy rate should decrease along with the increas of dimensionality for: (a) the relative redundancy in the word list counteracts the redundancy in the bigram list, because the words that contain a same bigram are gradually included as the dimensionality increases; (b) the proportion of interword bigrams increases in the bigram list and there is generally no redundancy between interword bigrams and intraword bigrams.

Last, there are more bigram features than word features because bigrams can overlap each other in the text but words can not. Thus the bigrams as a whole should theoretically contain more information than the words as a whole.

From the above analysis and observations, bigram features are expected to outperform word features at high dimensionalities. And word features are expected to outperform bigram features at low dimensionalities.

4 Semi-Quantitative Analysis

In this section, a preliminary statistical analysis is presented to corroborate the statements in the above qualitative analysis and expected to be identical with the experiment results shown in Section 1. All statistics in this section are based on the CE document collection and the *lqword* segmentation scheme (because the CE document collection is large enough to provide good statistical characteristics).

4.1 Intraword Bigrams and Interword Bigrams

In the previous section, only the intraword bigrams were discussed together with the words. But every bigram may have both intraword occurrences and interword occurrences. Therefore we need to distinguish these two kinds of bigrams at a statistical level. For every bigram, the number of intraword occurrences and the number of interword occurrences are counted and we can use

$$\log\left(\frac{interword\# + 1}{intraword\# + 1}\right)$$

as a metric to indicate its natual propensity to be a intraword bigram. The probability density of bigrams about on this metric is shown in Figure 8.

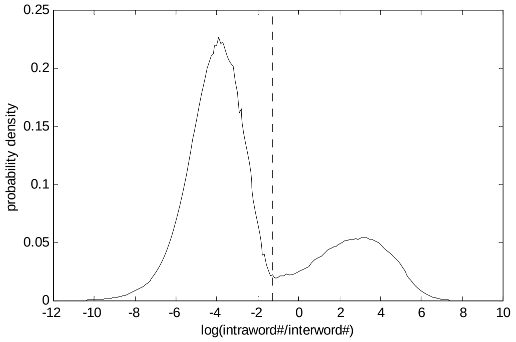

Figure 8. Bigram Probability Density on log(*intraword#*/*interword#*)

The figure shows a mixture of two Gaussian distributions, the left one for "natural interword bigrams" and the right one for "natural intraword bigrams". We can moderately distinguish these two kinds of bigrams by a division at -1.4.

4.2 Overall Information Quantity of a Feature Space

The performance limit of a classification is related to the quantity of information used. So a quantitative metric of the information a feature space can provide is need. *Feature Quantity* (Aizawa, 2000) is suitable for this purpose because it comes from information theory and is additive; *tfidf* was also reported as an appropriate metric of feature quantity (defined as "*probability* · *information*"). Because of the probability involved as a factor, the overall information provided by a feature space can be calculated on training data by summation.

The redundancy and complementation mentioned in Subsection 3.2 must be taken into account in the calculation of overall information quantity. For bigrams, the redundancy with regard to words associated with them between two intraword bigrams is given by

$$\sum_{b_{1,2}\subset w} tf(w)\cdot\min\left\{idf(b_1), idf(b_2)\right\}$$

in which b_1 and b_2 stand for the two bigrams and w stands for any word containing both of them. The overall information quantity is obtained by subtracting the redundancy between each pair of bigrams from the sum of all features' *feature quantity* (*tfidf*). Redundancy among more than two bigrams is ignored. For words, there is only complementation among words but not redundancy, the complementation with regard to bigrams associated with them is given by

$$\begin{cases} tf(w)\cdot\min\limits_{b\subset w}\left\{idf(b)\right\}, & \text{if } b \text{ exists;} \\ tf(w)\cdot idf(w), & \text{if } b \text{ does not exists.} \end{cases}$$

in which b is an intraword bigram contained by w. The overall information is calculated by summing the complementations of all words.

4.3 Statistics and Discussion

Figure 9 shows the variation of these overall information metrics on the CE document collection. It corroborates the characteristics analyzed in Section 3 and corresponds with the performance curves in Section 2.

Figure 10 shows the proportion of interword bigrams at different dimensionalities, which also corresponds with the analysis in Section 3.

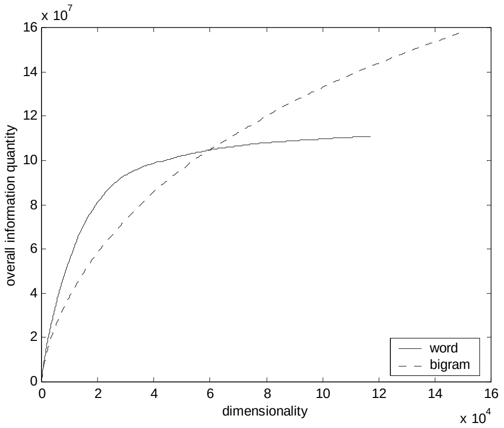

Figure 9. Overall Information Quantity on CE

The curves do not cross at exactly the same dimensionality as in the figures in Section 1, because other complications impact on the classification performance: (a) OOV word identifying capability, as stated in Subsection 3.1; (b) word segmentation precision; (c) granularity of the categories (words have more definite semantic meaning than bigrams and lead to a better performance for small category granularities); (d) noise terms, introduced in the feature space during the increase of dimensionality. With these factors, the actual curves would not keep increasing as they do in Figure 9.

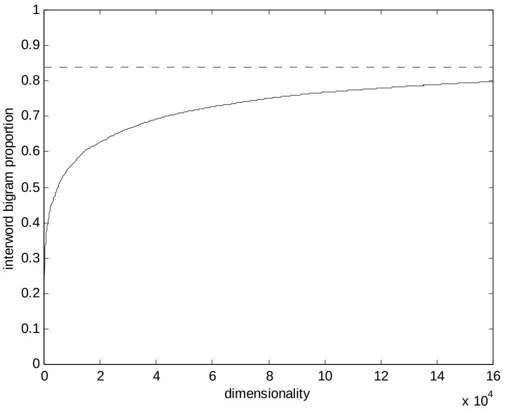

Figure 10. Interword Bigram Proportion on CE

5 Conclusion

In this paper, we aimed to thoroughly compare the value of words and bigrams as feature terms in text categorization, and make the implicit mechanism explicit.

Experimental comparison showed that the *Chi* feature selection scheme and the *tfidf* term weighting scheme are still the best choices for (Chinese) text categorization on a SVM classifier. In most cases, the bigram scheme outperforms the word scheme at high dimensionalities and usually reaches its top performance at a dimensionality of around 70000. The word scheme often outperforms the bigram scheme at low dimensionalities and reaches its top performance at a dimensionality of less than 40000.

Whether the best performance of the word scheme is higher than the best performance scheme depends considerably on the word segmentation precision and the number of categories. The word scheme performs better with a higher word segmentation precision and fewer (<10) categories.

A word scheme costs more document indexing time than a bigram scheme does; however a bigram scheme costs more training time and classification time than a word scheme does at the same performance level due to its higher dimensionality. Considering that the document indexing is needed in both the training phase and the classification phase, a high precision word scheme is more time consuming as a whole than a bigram scheme.

As a concluding suggestion: a word scheme is more fit for small-scale tasks (with no more than 10 categories and no strict classification speed requirements) and needs a high precision word segmentation system; a bigram scheme is more fit for large-scale tasks (with dozens of categories or even more) without too strict training speed requirements (because a high dimensionality and a large number of categories lead to a long training time).

Reference

Akiko Aizawa. 2000. *The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures, Proceedings of ACM SIGIR 2000,* 104-111.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval,* Addison-Wesley

Chih-Chung Chang, Chih-Jen Lin. 2001. *LIBSVM: A Library for Support Vector Machines*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Steve Deerwester, Sue T. Dumais, George W. Furnas, Richard Harshman. 1990. *Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science,* 41:391-407.

Thorsten Joachims. 1997. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Proceedings of 14th International Conference on Machine Learning* (Nashville, TN, 1997), 143-151.

Thorsten Joachims. 1998. *Text Categorization with Support Vector Machine: Learning with Many Relevant Features, Proceedings of the 10th European Conference on Machine Learning,* 137-142.

Mun-Kew Leong, Hong Zhou. 1998. *Preliminary Qualitative Analysis of Segmented vs. Bigram Indexing in Chinese, The 6th Text Retrieval Conference (TREC-6), NIST Special Publication 500-240,* 551-557.

Baoli Li, Yuzhong Chen, Xiaojing Bai, Shiwen Yu. 2003. *Experimental Study on Representing Units in Chinese Text Categorization, Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2003),* 602-614.

Yuan Liu, Nanyuan Liang. 1986. *Basic Engineering for Chinese Processing – Contemporary Chinese Words Frequency Count, Journal of Chinese Information Processing,* 1(1):17-25.

Robert W.P. Luk, K.L. Kwok. 1997. *Comparing representations in Chinese information retrieval. Proceedings of ACM SIGIR 1997*, 34-41.

Jianyun Nie, Fuji Ren. 1999. *Chinese Information Retrieval: Using Characters or Words? Information Processing and Management*, 35:443-462.

Jianyun Nie, Jianfeng Gao, Jian Zhang, Ming Zhou. 2000. *On the Use of Words and N-grams for Chinese Information Retrieval, Proceedings of 5th International Workshop on Information Retrieval with Asian Languages*

Monica Rogati, Yiming Yang. 2002. *High-performing Feature Selection for Text Classification, Proceedings of ACM Conference on Information and Knowledge Management 2002,* 659-661.

Gerard Salton, Christopher Buckley. 1988. *Term Weighting Approaches in Automatic Text Retrieval, Information Processing and Management,* 24(5):513-523.

Fabrizio Sebastiani. 2002. *Machine Learning in Automated Text Categorization, ACM Computing Surveys*, 34(1):1-47

Dejun Xue, Maosong Sun. 2003a. *Select Strong Information Features to Improve Text Categorization Effectiveness, Journal of Intelligent Systems,* Special Issue.

Dejun Xue, Maosong Sun. 2003b. *A Study on Feature Weighting in Chinese Text Categorization, Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2003),* 594-604.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory,* Springer.

Yiming Yang, Jan O. Pederson. 1997. *A Comparative Study on Feature Selection in Text Categorization, Proceedings of ICML 1997*, 412-420.