Extraction of Chinese Compound Words – An Experimental Study on a Very Large Corpus

Jian Zhang
Department of Computer Science and Technology of Tsinghua University, China
ajian@s1000e.cs.tsinghua.edu.cn

Jianfeng Gao, Ming Zhou
Microsoft Research China
{jfgao, mingzhou}@microsoft.com

Abstract

This paper is to introduce a statistical method to extract Chinese compound words from a very large corpus[1]. This method is based on *mutual information* and *context dependency*. Experimental results show that this method is efficient and robust compared with other approaches. We also examined the impact of different parameter settings, corpus size and heterogeneousness on the extraction results. We finally present results on information retrieval to show the usefulness of extracted compounds.

1 Introduction

Almost all techniques to statistical language processing, including speech recognition, machine translation and information retrieval, are based on words. Although word-based approaches work very well for western languages, where words are well defined, it is difficult to apply to Chinese. Chinese sentences are written as characters strings with no spaces between words. Therefore, words in Chinese are actually not well marked in sentences, and there does not exist a commonly accepted Chinese lexicon.

Furthermore, since new compounds (words formed with at least two characters) are constantly created, it is impossible to list them exhaustively in a lexicon. Therefore, automatic extraction of compounds is an important issue. Traditional extraction approaches used rules. However, compounds extracted in this way are not always desirable. So, human effort is still required to find the preferred compounds from a large compound candidate list. Some statistical approaches to extract Chinese compounds from corpus have been proposed (Lee-Feng Chien 1997, WU Dekai and Xuanyin XIA 1995, Ming-Wen Wu and Keh-Yih Su 1993) as well, but almost all experiments are based on relatively small corpus, it is not clear whether these methods still work well with large corpus.

In this paper, we investigate statistical approaches to Chinese compound extraction from very large corpus by using statistical features, namely *mutual information* and *context dependency*. There are three main contributions in this paper. First, we apply our procedure on a very large corpus while other experiments were based on small or medium size corpora. We show that better results can be obtained with a large corpus. Second, we examine how the results can be influenced by parameter settings including *mutual information* and *context dependency* restrictions. It turns out that *mutual information mainly affects precision while context dependency affects the count of extracted items.* Third, we test the usefulness of the extracted compounds for information retrieval. Our experimental results on IR show that the new compounds have a positive effect on IR.

The rest of this paper is structured as follows. In section 2, we describe the techniques we used. In section 3, we present several sets of experimental results. In section 4, we outline the related works as well as their results. Finally, we give our conclusions in section 5.

2 Technique description

Statistical extraction of Chinese compounds has been used in (Lee-Feng Chien 1997)(WU Dekai and Xuanyin XIA 1995) and (Ming-Wen Wu and Keh-Yih Su 1993). The basic idea is that a

[1] This work was done while the author worked for Microsoft Research China as a visiting student.

Chinese compound should appear as a stable sequence in corpus. That is, the components in the compound are strongly correlated, while the components lie at both ends should have low correlations with outer words.

The method consists of two steps. At first, a list of candidate compounds is extracted from a very large corpus by using *mutual information*. Then, *context dependency* is used to remove undesirable compounds. In what follows, we will describe them in more detail.

2.1 Mutual Information

According to our study on Chinese corpora, most compounds are of length less than 5 characters. The average length of words in the segmented-corpus is of approximately 1.6 characters. Therefore, only word *bi-gram*, *tri-gram*, and *quad-gram* in the corpus are of interest to us in compound extraction.

We use a criterion, called *mutual information*, to evaluate the correlation of different components in the compound. *Mutual information MI(x,y)* of a *bi-gram (x, y)* is estimated by:

$$MI(x,y) = \frac{f(x,y)}{f(x)+f(y)-f(x,y)}$$

Where *f(x)* is the occurrence frequency of word *x* in the corpus, and *f(x,y)* is the occurrence frequency of the word pair *(x,y)* in the corpus. The higher the value of *MI* is, the more likely *x* and *y* are to form a compound.

The mutual information MI(x,y,z) of tri-gram (x,y,z) is estimated by:

$$MI(x,y,z) = \frac{f(x,y,z)}{f(x)+f(y)+f(z)-f(x,y,z)}$$

The estimation of *mutual information* of *quad-grams* is similar to that of tri-grams. The extracted compounds should be of higher value of *MI* than a pre-set threshold.

2.2 Context Dependency

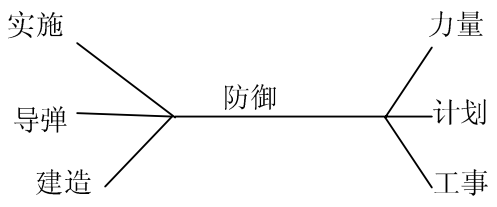

Figure 1

The extracted Chinese compounds should be complete. That is, we should generate a whole word, not a part of it. For example, 导弹防御计划(missile defense plan) is a complete word, and 导弹防御 (missile defense) is not, although both have relatively high value of mutual information.

Therefore, we use another feature, called *context dependency*. The contexts of the word 防御(defense) are illustrated by figure 1.

A compound *X* has NO left context dependency if

$$LSize = |L| > t1 \text{ or}$$

$$MaxL = MAX_{\alpha} \frac{f(\alpha X)}{f(X)} < t2$$

Where *t1*, *t2* are threshold value, *f(.)* is frequency, *L* is the set of left adjacent strings of *X*, $\alpha \in L$ and |*L*| means the number of unique left adjacent strings. Similarly, a compound *X* has NO right context dependency if

$$RSize = |R| > t3 \text{ or}$$

$$MaxR = MAX_{\beta} \frac{f(\beta X)}{f(X)} < t4$$

Where *t1*, *t2*, *t3*, *t4* are threshold value, *f(.)* is frequency, *R* is the set of right adjacent strings of *X*, $\beta \in R$ and |*R*| means the number of unique left adjacent strings.

The extracted complete compounds should have neither left nor right *context dependency*.

3 Experimental results

In our experiments, three corpora were used to test the performance of the presented approach. These corpora are described in table 1. Corpus A consists of local news with more than 325 million characters. Corpus B consists of documents from different domains of novel, news, technique report, etc., with approximately 650 million characters. Corpus C consists of People's Daily news and Xinhua news from TREC5 and TREC6 (Harman and Voorhees, 1996) with 75 million characters.

Table 1: Characteristics of Corpora

Corpus	Source	Size(char#)
Corpus A	political, economic news	325 M
Corpus B	Corpus A + novels + technique reports, etc.	650 M

Corpus C	TREC 5/6 Chinese corpus	75 M

In the first experiment, we test the performance of our method on corpus A, which is homogeneity in style. We then use corpus B in the second experiment to test if the method works as well on the corpus that is heterogeneity in style. We also use different parameter settings in order to figure out the best combination of the two statistical features, i.e. *mutual information* and *context dependency*. In the third experiment, we apply the results of the method to information retrieval system. We extract new compounds on corpus C, and add them to the indexing lexicon, and we achieve a higher average precision-recall. In all experiments, corpora are segmented automatically into words using a lexicon consisting of 65,502 entries.

3.1 Compounds Extraction from Homogeneous Corpus

Corpus A contains political and economic news. In this series of tests, we gradually loosen the conditions to form a compound, i.e. *MI* threshold becomes smaller and *MaxL/MaxR* becomes larger. Results for *quad-grams*, *tri-grams* and *bi-grams* are shown in tables 2,3,4. Some compounds extracted are listed in table 5.

Table2: Performance of *quad-gram* compounds extraction

	Parameter setting (*MI, LSize, MaxL, RSize, MaxR*)	Number of New compounds found	Precision (correct compounds/compounds checked)
1	0.01 1 0.75 1 0.75	27	100%（27/27）
2	0.005 1 0.85 1 0.85	92	98.9%（91/92）
3	0.002 1 0.90 1 0.90	513	95.8%（113/118）
4	0.001 1 0.95 1 0.95	1648	96.2%（179/186）
5	0.0005 1 0.95 1 0.95	4707	96.7%（206/213）

Table3: Performance of *tri-gram* compounds extraction

	Parameter setting (*MI, LSize, MaxL, RSize, MaxR*)	Number of New compounds found	Precision (correct compounds/compounds checked)
1	0.02 2 0.70 2 0.70	167	100%（167/167）
2	0.01 2 0.75 2 0.75	538	100%（205/205）
3	0.005 2 0.80 2 0.80	1607	100%（262/262）
4	0.003 2 0.80 2 0.80	3532	98.3%（341/347）
5	0.001 2 0.80 2 0.80	16849	96.6%（488/501）

Table4: Performance of *bi-gram* compounds extraction

	Parameter setting (*MI, LSize, MaxL, RSize, MaxR*)	Number of New compounds found	Precision (correct compounds/compounds checked)
1	0.05 3 0.5 3 0.5	1622	98.9%（184/186）
2	0.05 3 0.6 3 0.6	1904	98.6%（309/212）
3	0.03 3 0.6 3 0.6	3938	97.8%（218/223）
4	0.01 3 0.5 3 0.5	14666	97.5%（354/363）
5	0.005 3 0.5 3 0.5	32899	97.3%（404/415）

Table 5: Some *N-gram* compounds found by our method

N-gram	Extracted Compounds
N=2	粮|库(grain depot)、光盘|驱动器(CD-ROM Driver)、盖|茨(Bill Gates)
N=3	宣|武|门(XuanWu Gate)、异步|传输|模式(asynchronous transfer model)、亚|马|逊(Amazon)
N=4	爱|丽|舍|宫(Elysee)、俄|亥|俄|州(Ohio)、董|建|华|先生(Mr.Dong Jianhua)

It turns out that our algorithm successfully extracted a large number of new compounds (>50000) from raw texts. Compared with previous methods described in the next section, the precision is very high. We can also find that there is little precision loss when we loose restriction. The result may be due to three reasons. First, the two statistical features really characterize the nature of compounds, and provide a simple and efficient way to estimate the possibility of a word sequence being a compound. Second, the corpus we use is very large. It is always true that more data leads to better results. Third, the corpus we used in this experiment is homogeneity in style. The raw corpus is composed of news on politics, economy, science and technology. These are formal articles, and the sentences and compounds are well normalized and strict. This is very helpful for compound extraction.

3.2 Compounds Extraction from Heterogeneous Corpus

In this experiment, we use a heterogeneous corpus. It is a combination of corpus A, and some other novels, technique reports, etc. For simplicity, we discuss the extraction of *bi-gram* compounds only. In comparison with the first experiment, we find that the precision is strongly affected by the corpus we used. As shown in table 6, for each corpus, we use the same parameter setting, say *MI* >0.005, *LSize* >3, *MaxL* <0.5, *RSize*>3 and *MaxR*<0.5.

Table 6: Impact of heterogeneousness of corpora

Corpus	Compounds extracted	Extract precision
Corpus A	32899	97.3% (404/415)
Corpus B	36383	88.3% (362/410)

As we mentioned early, the larger the corpus we use, the better results we obtain. Therefore, we intuitively expect better result on corpus B, which is larger than corpus A. But, the result shown in table 6 is just the opposite.

There are mainly two reasons for this. The first one is that our method works better on homogeneous corpus than on heterogeneous corpus. The second one is that it might not be suitable to use the same parameter settings on two different corpora. We then try different parameter settings on corpus B.

There are two groups of parameters. *MI* measures the correlation between adjacent words, and other four parameters, namely *LSize, RSize, MaxL,* and *MaxR,* measure the context dependency. Therefore, each time, we fix one parameter, and relax another from tight to loose to see what happens. The Number of extracted compounds and precision of each parameter setting are shown in table 7.

Table 7: Extraction results with different parameter settings
(*MI=Mutual Information, CD = Context Dependency=(LSize, MaxL, RSize, MaxR)*)

MI\CD	(2, 0.8, 2, 0.8)	(6, 0.7, 6, 0.7)	(10, 0.6, 10, 0.6)	(14, 0.5, 4, 0.5)	(18, 0.4, 18, 0.4)	(22, 0.3, 22, 0.3)	(26, 0.2, 6, 0.2)
0.0002	1457781 (39.06%)	809502 (42.24%)	570601 (43.98%)	426223 (44.67%)	314810 (43.96%)	209910 (43.38%)	96383 (40.93%)
0.0004	784082 (48.98%)	485143 (46.84%)	359499 (52.53%)	277673 (49.25%)	209634 (53.92%)	141215 (49.55%)	63907 (52.53%)
0.0006	530723 (51.28%)	349882 (53.96%)	266068 (60.39%)	208921 (52.48%)	159363 (49.49%)	108120 (63.35%)	48683 (61.65%)
0.0008	396602 (54.63%)	273231 (58.00%)	211044 (55.19%)	167660 (65.24%)	128819 (60.54%)	87869 (64.40%)	39502 (54.86%)
0.0010	313868 (59.11%)	223827 (66.51%)	175050 (61.14%)	140197 (57.66%)	108322 (67.38%)	74104 (63.08%)	33354 (67.50%)
0.0012	257990 (58.94%)	189014 (59.50%)	149315 (60.98%)	120312 (65.28%)	93323 (70.47%)	64079 (65.32%)	28879 (64.65%)
0.0014	217766 (58.93%)	163189 (67.91%)	129978 (60.19%)	105334 (65.84%)	82083 (66.83%)	56582 (67.50%)	25486 (65.46%)

Table 7 shows the extraction results with different parameters. These results fit our intuition. While parameters become more and more strict, less and less compounds are found and precisions become higher. This phenomena is also illustrated in figure 2 and 3, in which the "correct compounds extracted" is an estimation from table7, i.e. *number of compounds found* × *precision*. (These two figures are very useful for one who wants to automatically extract a new lexicon with pre-defined size from a large corpus.)

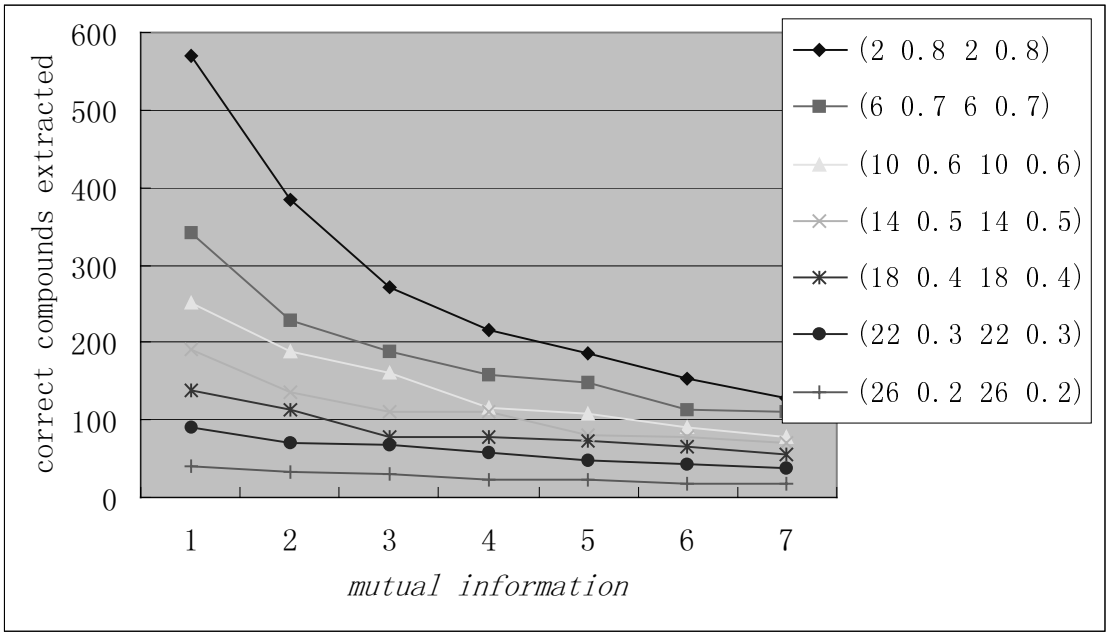

Figure 2 Impact of Parameter *Mutual Information*

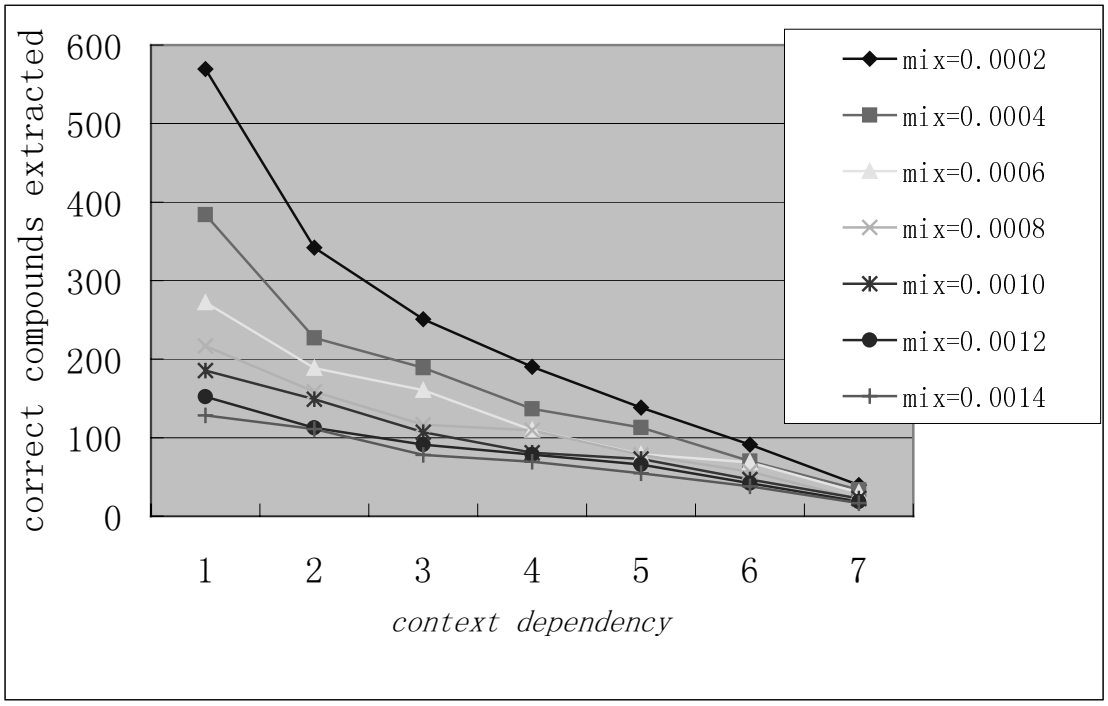

Figure 3 Impact of Parameter *Context Dependency*

The precision of extraction is estimated in the following way. We extract a set of compounds based on a series of pre-defined parameter set. For each set of compounds, we randomly select 200 compounds. Then we merge those selected compounds to a new file for manually check. This file consists of about 9,800 new compounds because there are 49 compounds lists. One person will group these 'compounds' into two sets, say set A and set B. Set A contains the items that are considered to be correct, and set B contains incorrect ones. Then for each original group of about 200 compounds we select in the first step, we check how many items that also appear in set A and how many items in set B. Suppose these two values are *a1* and *b1*, then we estimate the precision as *a1/(a1+b1)*.

So, there are two important points in our evaluation process. First, it is difficult to give a definition of the term "compound" to be accepted popularly. Different people may have different judgement. Only one person takes part in the evaluation in our experiment. This can eliminate the effect of divergence among different persons. Second, we merge those items together. This can eliminate the effect of different time period. One may feel tired after checked too many items. If he checks those 49 files one by one, the latter results are incomparable with the previous one.

The precisions estimated by the above method are not exactly correct. However, as described above, the precisions of different parameter settings are comparable. In this experiment, what we want to show is how the parameter settings affect the results.

Both *MI* and *CD* can affect number of extracted compounds, as shown in table 7. Compared with *MI*, *CD* has stronger effect in this aspect. For each row in table 7, numbers of extracted compounds finally decrease to 10% of that showed in the first column. For each column, while *MI* changes from 0.0002 to 0.0014, the number is decreased of about 20%. This may be explained by the fact that it is difficult for candidate to fulfill all four restrictions in *CD* simultaneously. Many disqualified candidates are cut off. Table 7 lists the precisions of extracted results. It shows that there is no clear increasing/decreasing pattern in each row. That is to say, *CD* doesn't strongly affect the precision. When we check each column, we can see that precision is in a growing progress. As we defined above, *MI* and *CD* are two different measurements. What role they play in our extraction procedure? Our conclusion is that *mutual information mainly affects the precision while context dependency mainly affects the count of extracted items*. This conclusion is also confirmed by Fig2 and Fig3. That is, the curves in Fig2 are more flat than corresponding curves in Fig3.

3.3 Testing the Extracted Compounds in Information Retrieval

In this experiment, we apply our method to improve information retrieval results. We use SMART system (Buckley 1985) for our experiments. SMART is a robust, efficient and flexible information retrieval system. The corpus used in this experiment is TREC Chinese corpus (Harman and Voorhees, 1996). The corpus contains about 160,000 articles, including articles published in the People's Daily from 1991 to 1993, and a part of the news released by the Xinhua News Agency in 1994 and 1995. A set of 54 queries has been set up and evaluated by people in NIST(National Institute of Standards and Technology).

We first use an initial lexicon consisting of 65,502 entries to segment the corpus. When running SMART on the segmented corpus, we obtain an average precision of 42.90%.

Then we extract new compounds from the segmented corpus, and add them into the initial lexicon. With the new lexicon, the TREC Chinese corpus is re-segmented. When running SMART on this re-segmented corpus, we obtain an average precision of 43.42%, which shows a slight improvement of 1.2%.

Further analysis shows that the new lexicon brings positive effect to 10 queries and negative effect to 4 queries. For other 40 queries, there is no obvious effect. Some improved queries are listed in table 8 as well as new compounds being contained.

As an example, we give the segmentation results with the two lexicons for query 23 in table 9.

Table 8: Improved Query Samples

Query ID	Base line precision	New precision	Improvement	Extracted compounds
9	0.3648	0.4173	14.4%	毒品买卖(drugs sale),中国毒品问题(Drug Problems in China)
23	0.3940	0.5154	30.8%	联合国安理会(the UN Security Council),和平建议(peace proposal)
30	0.3457	0.3639	5.3%	
46	0.3483	0.4192	20.4%	中越(China and Vietnam)
47	0.5369	0.5847	8.9%	皮纳图博火山(Mount Minatubo),臭氧层(ozone layer), 苏比克(Subic)

Table 9: Segmented Corpus with the Two Lexicons for Query 23

Query 23 segment with small lexicon 相关 文件 应 提及 苏联 在 海湾战争 中 如何 担任 调停 的 角色 , 包括 与 伊拉克 之间 的 沟通 , 苏联 在 [联合国] [安理会] 中 提出 的 停火协议 , 以及 要求 多国部队 从 伊拉克 撤出 的 [和平] [建议]
Query 23 segment with new lexicon 相关 文件 应 提及 苏联 在 海湾战争 中 如何 担任 调停 的 角色 , 包括 与 伊拉克 之间 的 沟通 , 苏联 在 **[**联合国安理会**]** 中 提出 的 停火协议 , 以及 要求 多国部队 从 伊拉克 撤出 的 **[**和平建议**]**

Another interesting example is query 30. There is no new compound extracted from that query. Its result is also improved significantly because its relevant documents are segmented better than before.

Because the compounds extracted from the corpus are not exactly correct, the new lexicon will bring negative effect to some queries, such as query 10. The retrieval precision changes from 0.3086 to 0.1359. The main reason is that "中国新疆"(Chinese XinJiang) is taken as a new compound in the query.

4 Related works

Several methods have been proposed for extracting compounds from corpus by statistical approaches. In this section, we will briefly describe some of them.

(Lee-Feng Chien 1997) proposed an approach based on PAT-Tree to automatically extracting domain specific terms from online text collections. Our method is primary derived from (Lee-Feng Chien 1997), and use the similar statistical features, i.e. *mutual information* and *context dependency*. The difference is that we use n-gram instead of PAT-Tree, due to the efficiency issue. Another difference lies in the experiments. In Chien's work, only domain specific terms are extracted from domain specific corpus, and the size of the corpus is relatively small, namely 1,872 political news abstracts.

(Cheng-Huang Tung and His-Jian Lee 1994) also presented an efficient method for identifying unknown words from a large corpus. The statistical features used consist of string (character sequence) frequency and entropy of left/right neighboring characters (similar to left/right context dependency). The corpus consists of 178,027 sentences, representing a total of more than 2 million Chinese characters. 8327 unknown words were identified and 5366 items of them were confirmed manually.

(Ming-Wen Wu and Keh-Yih Su 1993) presented a method using mutual information and relative frequency. 9,124 compounds are extracted from the corpus consists of 74,404 words, with the precision of 47.43%. In this method, the compound extraction problem is formulated as classification problem. Each bi-gram (tri-gram) is assigned to one of those two clusters. It also needs a training corpus to estimate parameters for classification model. In our method, we didn't

make use of any training corpus. Another difference is that they use the method for English compounds extraction while we extract Chinese compounds in our experiments .

(Pascale Fung 1998) presented two simple systems for Chinese compound extraction—CXtract. CXtract uses predominantly statistical lexical information to find term boundaries in large text. Evaluations on the corpus consisting of 2 million characters show that the average precision is 54.09%.

We should note that since the experiment setup and evaluation systems of the methods mentioned above are not identical, the results are not comparable. However, by showing our experimental results on much larger and heterogenous corpus, we can say that our method is an efficient and robust one.

5 Conclusion

In this paper, we investigate a statistical approach to Chinese compounds extraction from very large corpora using *mutual information* and *context dependency*.

We explained how the performance can be influenced by different parameter settings, corpus size, and corpus heterogeneousness. We also refine the lexicon with information retrieval system by adding compounds obtained by our methods, and achieve 1.2% improvements on precision of IR.

Through our experiments, we conclude that statistical method based on *mutual information* and *context dependency* is efficient and robust for Chinese compounds extraction. And, *mutual information* mainly affects the precision while *context dependency* mainly affects the count of extracted items.

Reference

Lee-Feng Chien, (1997) “PAT-tree-based keyword extraction for Chinese Information retrieval”, *ACM SIGIR’97*, Phliadelphia, USA, 50-58

WU, Dekai and Xuanyin XIA. (1995). "Large-scale automatic extraction of an English-Chinese lexicon". *Machine Translation* 9(3-4), pp.285-313.

Ming-Wen Wu and Keh-Yih Su. (1993). "Corpus-based Automatic Compound Extraction with Mutual Information and Relative Frequency Count," *Proceedings of R. O. C. Computational Linguistics Conference* VI . Nantou, Taiwan, R. O. C., pp.207-216.

Pascale Fung. (1998). "Extracting Key Terms from Chinese and Japanese texts ". *The International Journal on Computer Processing of Oriental Language, Special Issue on Information Retrieval on Oriental Languages*, pp.99-121.

Cheng-Huang Tung and His-Jian Lee. (1994). “Identification of Unknown Words From a Corpus”. *Compouter Processing of Chinese and Oriental Languages* Vol.8, pp.131-145.

Buckley, C. (1985). *Implementation of the SMART information retrieval system*, Technical report, #85-686, Cornell University.

Harman, D. K. and Voorhees, E. M., Eds. (1996). *Information Technology: The Fifth Text Retrieval Conference(TREC5)*, NIST SP 500-238. Gaithersburg, National Institute fo standards and Technology.