Chinese Term Extraction from Web Pages Based on Compound word Productivity

Hiroshi Nakagawa
Information Technology Center, The University of Tokyo
7-3-1 Hongou, Bunkyo
Tokyo, JAPAN, 113-0033
nakagawa@dl.itc.u-tokyo.ac.jp

Hiroyuki Kojima*, Akira Maeda†
Faculty of Economics, The University of Tokyo
7-3-1 Hongou, Bunkyo
Tokyo, JAPAN, 113-0033
* kojima@e.u-tokyo.ac.jp,
† maeda@lib.u-tokyo.ac.jp

Abstract

In this paper, we propose an automatic term recognition system for Chinese. Our idea is based on the relation between a compound word and its constituents that are simple words or individual Chinese character. More precisely, we basically focus on how many words/characters adjoin the word/character in question to form compound words. We also take into account the frequency of term. We evaluated word based method and character based method with several Chinese Web pages, resulting in precision of 75% for top ten candidate terms.

1 Introduction

Automatic term recognition, ATR in short, aims at extracting domain specific terms from a corpus or Web pages. Domain specific terms are terms that expresses the concept specifically defined in the given domain. They are required to have a unique meaning in order for efficient communication about the topic of the domain. It is, however, difficult to decide whether they are unique automatically. So we put this issue aside. In terms of feasibility, their grammatical status is important, for instance part of speeches. Although they are not necessarily confined to nouns, the majority of them are actually simple or compound words, where "simple word" means a word which could not be further divided into shorter and more basic words. Thus, we here focus only on simple and compound words.

In terms of text length, even one Web page which is not long gives us a number of domain specific vocabulary like "national library", "library policy" if the Web page is about libraries. If we expand domain specific terms to this extent, the big portion of domain specific terms are compound words. Obviously, the majority of compound words consist of relatively small number of distinct simple words. In this situation, it is natural to pay attention to the relation among compound words and their constituent simple words.

(Kageura & Umino 96) proposed an important feature of domain specific terms called *termhood* which refers to the degree that a linguistic unit is related to a domain-specific concept. Presumably, it is necessary to develop an ATR method that calculates termhood of each term candidate extracted from a domain corpus that usually consists of a number of documents. Many works of ATR use statistics of term candidate distribution in a corpus such as term frequency to calculate the termhood of each term candidate.

This frequency based methods, however, heavily depend on the size of corpus. Thus we could not expect good result if we extract domain specific terms from one or a few Web pages. If we shift our focus from a corpus based statistics like frequency to **term space** that consists of all term candidates, we expect better result of extracted terms even from one Web page because of the following reason: A set of term candidates has its own structure like relations between compound words and their constituent simple words as stated before. The statistical information about these relations comes from more microscopic structure than term frequency. Thus, if we utilize more information from term space, it is reasonable in extracting from a small text like one Web page. Without this kind of information, we will be suffering from the shortage of information for ATR.

Now look at frequency based information and information inherent with term space more closely. Even though several kinds of statistics about actual use in a corpus such as term frequency give a good approximation of termhood. They are not necessarily meanings in a writer's mind. On the contrary, the statistics of term space can reflect the meaning in a writer's mind because it is up to a writer's decision how to make a compound word term to express a complicated concept using simple word terms as its components. More precisely, if a certain simple word, say *N,*

expresses the basic concept of a domain that the document treats, the writer of the document, we expect, uses *N* not only many times but in various ways. One of typical way of this kind is that he/she composes quite a few compound words using *N* and uses these compound words in documents he/she writes. For this reason, we have to focus on the relation among simple words and compound words when pursuing new ATR methods.

One of the attempts to make use of this relation has been done by Nakagawa and Mori (2003). Their method is based on the number of distinct simple words that come left or right of a simple word term to make up compound word terms. In this paper, we apply their method to deal with Web pages written in Chinese.

In this paper, section 2 gives the background of ATR methods. In section 3 we introduce ATR method developed by Nakagawa and Mori(2003). Section 4, 5 and 6 are for how to apply their method to Chinese language and evaluation of two proposed method: 1) Word based method using Chinese morphological analyzer ICTCLAS(Zhang, Yu, Xiong and Liu. 2003), 2) Stop character based method.

2 Background

2.1 Typical Procedures of Automatic Term Recognition

An ATR procedure consists of two procedures in general. The first one is a procedure of extracting term candidates from a corpus. The second procedure is to assign each term candidate a score that indicates how likely the term candidate is a term to be recognized. Then all candidates are ranked according to their scores. In the remaining part of this section, we describe the background of a candidate extraction procedure and a scoring procedure respectively.

2.2 Candidates Extraction

In term candidates extraction from the given text corpus, we mainly focus on compound words as well as simple words. To extract compound words which are promising term candidates and at the same time to exclude undesirable strings such as "*is a"* or "*of the"*, the frequently used method is to filter out the words that are the member of a *stop-word-list*.

The structure of complex term is another important factor for automatic term candidate extraction. A syntactic structure that is the result of parsing is focused on in many works. Since we focus on these complex structures, the first task in extracting term candidates is a morphological analysis including part of speech (POS) tagging. There are no explicit word boundary marker in Chinese, we first have to do morphological analysis which segments out words from a sentence and does POS tagging at the same time.

After POS tagging, the complex structures mentioned above are extracted as term candidates. Previous studies have proposed many promising ways for this purpose, for instance, Smadja and McKeown (1990), and Frantzi and Ananiadou (1996) tried to treat more general structures like collocations.

2.3 Scoring

The next step of ATR is to assign each term candidate its score in order to rank them in descending order of termhood. Many researchers have sought the definition of term candidate's score which approximates termhood. In fact, many of those proposals make use of statistics of actual use in a corpus such as term frequency which is so powerful and simple that many researchers directly or indirectly have used it. The combination of term frequency and inverse document frequency is also well studied i.e. (Uchimoto et al 2000), (Fukushige and Noguchi 2000). On the other hand, several scoring methods that are neither directly nor heavily based on frequency of term candidates have been proposed. Among those, Ananiadou et al. proposed C-value (Frantzi and Ananiadou 1996) which counts how independently the given compound word is used in the given corpus. Hisamitsu (2000) proposes a way to measure termhood which estimates how far the document containing given term is different from the distribution of documents not containing the given term. However, the method proposed in (Nakagawa and Mori 2003) outperforms these methods in terms of NTCIR1 TMREC task(Kageura, et al, 1999).

2.4 Chinese Term Extraction

As for Chinese language NLP, very many works about word segmentation were published i.e. (Ma and Xia 2003). Nevertheless the term "Term extraction" has not yet been used for Chinese NLP, key words extraction have been a target for a long time. For instance, key words extraction from news articles (Li. et al. 2003) is the recent result which uses frequency and length of character string for scoring. Max-duplicated string based method (Yang and Li. 2002) is also promising. In spite of previous research efforts, there have been no attempt so far to apply the relation between simple and compound word to Chinese term extraction, and that is exactly what we propose in this paper.

3 Scoring methods with Simple word Bigrams

3.1 Simple word Bigrams

The relation between a simple word and complex words that include the simple word is very important in terms of term space structure. Nevertheless, to my knowledge, this relation has not been paid enough attention so far except for. (Nakagawa and Mori 2003). In this paper, taking over their works, we focus on compound words among the various types of complex terms. In technical documents, the majority of domain-specific terms are noun phrases or compound words consisting of small size vocabulary of simple words. This observation leads to a new scoring methods that measures how many distinct compound words contain the simple word in question as their part in a given document or corpus. Here, suppose the situation where simple word: N occurs with other simple words as a part of many compound words shown in Figure 1 where [N M] means bigram of noun N and M.

[LN1 N] (#L1)	[N RN1](#R1)
[LN2 N] (#L2)	[N RN2](#R2)
:	:
[LNn N](#Ln)	[N RNm](#Rm)

Figure 1. Noun Bigram and their Frequency

In Figure 1, [LNi N] (i=1,..,n) and [N RNj] (j=1,...,m) are simple word bigrams which make (a part of) compound words. #Li and #Rj (i=1,..,n and j=1,..,m) mean the frequency of the bigram [LNi N] and [N RNj] in the corpus respectively. Note that since we depict only bigrams, compound words like [LNi N RNj] which contains [LNi N] and/or [N RNj] as their parts might actually occur in a corpus. Note that this noun trigram might be a part of longer compound words. We show an example of a set of noun bigrams. Suppose that we extract compound words including "trigram" as term candidates from a corpus as shown in the following example.

Example 1.
trigram statistics, word trigram, class trigram, word trigram, trigram acquisition, word trigram statistics, character trigram

Then, noun bigrams consisting of a simple word "trigram" are shown in Figure 2 where the number between (and) shows the frequency in the corpus.

word trigram (3)	trigram statistics (2)
class trigram (1)	trigram acquisition (1)
character trigram(1)	

Figure 2. An example of noun bigram

Now we focus on and utilize simple word bigrams to define the scoring function. Note that we are only concerned with simple word bigrams and not with a simple word per se because, as stated before, we are concerned with the relation between a compound word and its component simple words.

3.2 Scoring Function

3.2.1 Score of simple word

Since there are infinite number of scoring functions based on [LNi N] or [N RNj], we here consider the following simple but representative scoring functions.

#LDN(N) and **#RDN(N) :** These are the number of distinct simple words which directly precede or succeed **N**. These coincide with "n" and "m" in Figure 1 respectively. For instance, in an example shown in Figure 2, #LDN(trigram)=3, #RDN(trigram)=2.

Using #LDN and #RDN we define **LN(N)** and **RN(N):** These are based on the number of occurrence of each noun bigram, and defined for [LNi N] and [N RNj] as follows respectively.

$$LN(N) = \sum_{i=1}^{\#LDN(N)} (\#Li) \qquad (1)$$

$$RN(N) = \sum_{j=1}^{\#RDN(N)} (\#Rj) \qquad (2)$$

LN(N) and **RN(N)** are the frequencies of nouns that directly precede or succeed **N**. For instance, in an example shown in Figure 2, LN(trigram)=5, and RN(trigram)=3.

Let's think about the background of these scoring functions. #LDN(N) and #RDN(N), where we do not take into account the frequency of each noun bigram but take into account the number of distinct nouns that adjoin to N to make compound words. That indicates how linguistically and domain dependently productive the noun:N is in a given corpus. That means that if N presents a key and/or basic concept of the domain treated by the corpus, writers in that domain work out many distinct compound words with N to express more complicated concepts. On the other hand, as for **LN(N)** and **RN(N)**, we also focus on frequency of each noun bigram as well. In other words, statistic

bias in actual use of noun:N is, this time, one of our main concern. For example, in Figure 2, LN(trigram,2)=11, and RN(trigram,2)=5. In conclusion, since **LN(N)** and **RN(N)** convey more information than #LDN(N) and #RDN(N), we adopt **LN(N)** and **RN(N)** in this research.

3.2.2 Score of compound words

The next thing to do is expanding those scoring functions for simple word to the scoring functions for compound words. We adopt a geometric mean for this purpose. Now think of a compound word : $CN = N_1 N_2 \ldots N_{L,}$ where N_i (i= 1,.., L) is a simple word. Then a geometric mean: **LR** of CN is defined as follows.

$$LR(CN)=\left(\prod_{i=1}^{L}(LN(N_i)+1)(RN(N_i)+1)\right)^{1/2L} \quad (3)$$

For instance, if we use LN(N) and RN(N) in example 1, LR(trigram) = $\sqrt{(3+1)\times(5+1)}$ = 4.90. LR does not depend on the length of CN where "length" means the number of simple words that consist of CN. This is because since we have not yet had any idea about the relation between the importance of a compound word and a length of the compound word, it is fair to treat all compound words, including simple words, equally no matter how long or short each compound word is.

3.2.3 Combining LR and Frequency of Nouns

We still have not fully utilized the information about statistics of actual use in a corpus in the bigram based methods described in 3.2.1 and 3.2.2. Among various kinds of information about actual use, the important and basic one is the frequency of single-and compound words that occur independently. The term "independently" means that the left and right adjacent words are not nouns. For instance, "word patterns" occurs independently in "we use the word patterns which occur in this sentence." Since the scoring functions proposed in 3.2.1 is noun bigram statistics, the number of this kind of independent occurrences of nouns themselves have not been used so far. If we take this information into account, the better results are expected. Thus, if a simple word or a compound word occurs independently, the score of the noun is simply multiplied by the number of its independent occurrences. We call this new scoring function as **FLR(CN)** which is defined as follows.

if N occurs independently
then $FLR(CN) = LR(CN) \times f(CN)$
where f(CN) means the number of independent occurrences of noun CN (4)

4 Term Extraction for Chinese based on Morphological Analysis

If we try to apply the scoring method proposed in section 3 directly to a Chinese text, every word should be POS tagged because we extract multi-word unit of several types of POS tag sequences as candidates of domain specific terms. For this we need a Chinese morphological analyzer because Chinese is an agglutinative language. Actually, we use Chinese morphological analyzer: ICTCLAS(Zhang and Liu 2004). As term candidates, we extract compound word: MWU having the following POS tag sequence expressed in (5). A multi-word-unit: MWU is defined by the following CFG rules where the right hand sides are expressed as a regular expression.

MWU ← [ag a]* [ng n nr ns nt nz nx vn an i j]+"
MWU ← MWU?b [ng n nr ns nt nz nx vn an i j]+
MWU ← [ag a]+ [u k] MWU
MWU ← MWU (u|k|*he-2*|*yu-3*) MWU (5)

where ag,a,n,…,u are all tags used in ICTCLAS.

Roughly speaking (5) means an adjective followed by the repetition of [adjective noun particle] followed by a noun. The problem is the ambiguity of POS tagging because the same word is very often used verb as well as noun. In addition, unknown words like newly appeared proper names also impairs the accuracy. Due to this problem caused by morphological analyzer, the accuracy is degraded.

Once we segment out word sequences conforming the above POS tag sequences, we calculate LN and RN of each component word. In calculation of LN and RN, a word whose POS is c, u or k is omitted. In other words, if a word sequence "w1 w2 w3" where POS of w2 is c u or k, then we calculate RN of w1 and LN of w3 by regarding the word sequence as "w1 w3."

Then we combine LN and RN of each word to calculate FLR by definition of (3) and (4) to sort all extracted candidates in descending order of FLR.

We apply the proposed methods to 30 Web pages from People's Daily news. The areas are social, international and IT related news. The average length is 592.6 characters. Firstly, we extract relevant terms by hand from each news article and use it as the gold standard. The average number of gold standard terms per one news particle is 15.9 words. Secondly, we extract terms from each news article and sort them in descending

order by the proposed method and evaluate them by a precision of top N terms defined as follows.

$CT(K)$= 1 if Kth term is one of the gold standard terms.
0 otherwise

$$precision(K) = \frac{\sum_{i=1}^{K} CT(i)}{K} \quad (6)$$

where N is the number of the gold standard terms, and in our experiment, N=20. *Precision(K)*, where K=1,..,20, are shown in Figure 3 as "Strict."

We also use another precision rate *precision'* which is not strict and defined as follows.

$CTpart(K)$= 1 if one of gold standard terms. is a part of Kth term
0 otherwise

$$precision'(K) = \frac{\sum_{i=1}^{K} CTpart(i)}{K} \quad (7)$$

These are also shown in Figure 3 as "Partly."

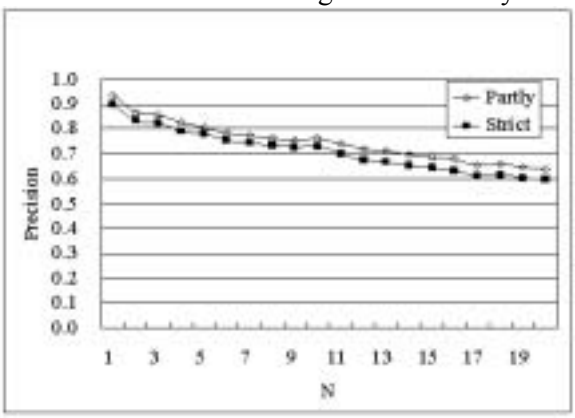


Figure 3. Strict and partly precision of word based extraction method.

From Figure 3, we see that If we pick up the ten highest ranked terms, about 75% of them meet the gold standard. The case we loosen the definition of precision shows better than the strict case of (6) but the difference is not so large. That means that the proposed word based ranking method works very well to extract important Chinese terms from news articles.

5 Character based Term Extraction

There are several reasons why we would like to develop a term extraction system without morphological analyzer.

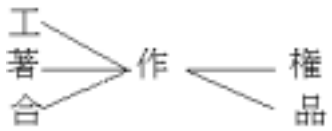
The first reason is that the accuracy of morphological analyzer is, in spite of the great advancement of these years, still around 95% (GuoDong and Jian 2003).

The second reason is that there possibly exist terminologies with unexpected POS sequences. If we deal only with academic papers or technical documents, we expect POS sequences of terminologies with high accuracy. However, if we consider terminology extraction from Web pages, the possibility of unexpected POS sequence may rise.

The third reason is language independency. Currently proposed and/or used morphological analyzers heavily depend either upon the sophisticated linguistic knowledge about the target language or upon a big size corpus of the target language if machine learning is employed. These linguistic resources, however, are not always available.

Due to these reasons, we also developed term candidate extraction system which does not use a morphological analyzer. Instead of morphological analyzer, we try to employ a stop word list. In Chinese, as stop words, we find many character unigrams and bigrams because one Chinese character conveys larger amount of information than a character of Latin alphabet. They are partly shown in Appendix A.

As term candidates, we simply extract character strings between two stop words that are nearest each other within a sentence. Obviously, the character strings thus extracted are not necessarily meaningful compound words. Therefore we cannot directly use these strings as words to calculate LN and RN function. Back to the idea that Chinese characters are ideograms, we come up to the idea that we calculate LN and RN of each character appearing within every character strings extracted as candidates. An example is shown in Figure 4.



LN(*zuo-4*)=3 RN(*zuo-4*)=2

Figure 4. LN and RN of Chinese character *zuo-4*

In calculation of LN and RN, we neglect characters whose POS are c ,u or k as same as we did in morphological analyzer based method.

Once we calculate LN and RN of each character, FLR of every character string is calculated as defined by (3) and (4) to sort them in descending order of FLR.

Actually this idea is very similar with left and right entropy used to extract two character Chinese words from a corpus (Luo and Sun. 2003). However what we would like to extract is a set of longer compound words or even phrases used in a Web page. Moreover we only use the Web page and do not use any other language resources such as a big corpus at all due to the reason described above in this section.

We evaluate the proposed character based extraction method against the same Web pages from People's Daily news used in Morphological Analysis based method described in Section 4. We also use the same gold standard terms described in Section 4 for evaluation. The strict and partly precision defined by (6) and (7) are used. The result is shown in Figure 5.

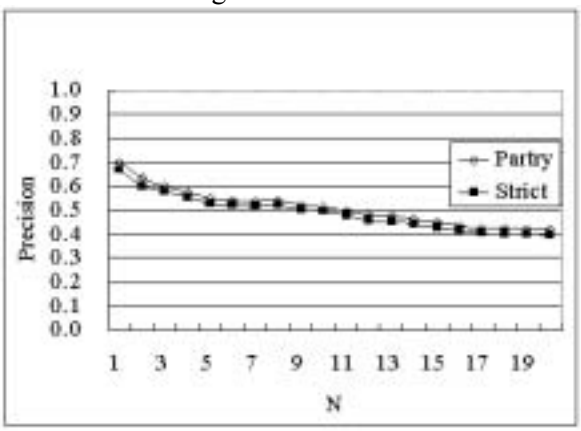


Figure 5. Strict and partly precision of character based extraction method.

Comparing Figure 3 with Figure 5, apparently the result of extracted terms of word based method is better than that of character based method. However, it does not necessarily mean that the character based term extraction is inferior.

If you take a glance at the stop word list of Appendix A, it seems that several of the stop words are selected mainly from words in auxiliary verbs, pronouns, adverbs, particles, prepositions, conjunctions, exclamations, onomatopoeic words and punctuation marks. However, in reality, our selection is based rather on meaning, usage and generally frequency of use than parts of speech. Thus some of them are not function words but content words in order to exclude non-domain-specific words. Actually, the stop words are not only character unigram but character bigram. Because Chinese character is ideograph and each character may have plural meanings, it is difficult only to use character unigram as a stop word in Chinese.

Our method based on these viewpoints resulted in getting an interesting consequence. We show an example of news article and extracted terms from it by this method in Appendix B and Appendix C. This news article is entitled "The Culture of Tibetan Web Site is formally created." Let's take a look at an underlined sentence in this short article and underlined terms extracted from there. This sentence says: According to the introduction, The Culture of Tibetan Web Site is a site of special pure culture for the purpose of "investigating the essence of Tibetan culture, showing the scale of Tibetan culture and raising the spirit of Tibetan culture". In the case of method based on stop word list, we can extract compound term of "investigating the essence of Tibetan culture (盘点藏族文化精英)", "showing the scale of Tibetan culture (展示藏族文化规模)", "raising the spirit of Tibetan culture (弘扬藏族文化精神)" and so on from this sentence. On the contrary, by the term extraction method based on morphological analysis, gerund , for example, "showing(展示)" and "raising (弘扬), can not be extracted.

We said that the majority of domain specific terms are noun phrases or compound words consisting of small size vocabulary of simple words as stated in section 3. So we especially have paid attention to relation among nouns. However most of Chinese nouns can also be used as verbs. Moreover inflection of Chinese verbs can hardly be recognized visually. It is difficult to distinguish verb from noun by morphological analysis. Certainly ICTCLAS classifies the character that has meaning of both verb and noun into the category of vn (verb and noun). But gerunds and verbal noun infinitives are not contained in vn. For instance, "写信" means not only "write a letter" but "writing letter." Thus we have to pay attention to verbs in Chinese too. Only by tuning up stop word list, we can take gerunds and verbal noun infinitives into account to some extent. Appendix C shows one of the evidence of this observation.

6 Conclusion

In this paper, we apply automatic term recognition system based on FLR (Nakagawa and Mori 2003) to Chinese Web pages. We proposed two methods: word based and character based extraction and ranking. Since the accuracies of term recognition are around 60% for top 1,000 term candidates in NTCIR TMREC task(Kageura et al 1999), the result of 75% accuracy of top ten candidates is a good start because the term extraction from small text like one Web page is the future oriented topic.

References

Ananiadou, S. 1994. "A Methodology for Automatic Term Recognition". In *Proceedings of 14th International Conference on Computational Linguistics* :1034 - 1038.

GuoDong Zhou and Jian Su. 2003. A Chinese Efficient Analyser Integrating Word Segmentation, Part-Of-Speech Tagging, Partial Parsing and Full Parsing, *In Proceedings of The Second SIGHAN Workshop, on Chinese Language Processing .ACL2003* :78-83

Frantzi, T.K. and Ananiadou, S. 1996. “Extracting nested collocations”. In *Proceedings of 16th International Conference on Computational Linguistics* :41-46.

Fukushige, Y. and N. Nogichi. 2000. “Statistical and linguistic approaches to automatic term recognition* NTCIR experiments at Matsushita”. Terminology 6(2) :257-286

Hua-PingZhang, Hong-KuYu, De-Yi Xiong and Qun Liu. 2003. “HHMM-based Chinese Lexical Analyzer ICTCLAS”. In *Proceedings of The Second SIGHAN Workshop, on Chinese Language Processing .ACL2003* :184-187

Hisamitsu, T, 2000. “A Method of Measuring Term Representativeness”. In *Proceedings of 18th International Conference on Computational Linguistics*, :320-326.

Kageura, K. and Umino, B. 1996. “Methods of automatic term recognition: a review”. *Terminology* 3(2) :259-289.

Kageura, K. et al, 1999. TMREC Task: Overview and Evaluation. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition* :411-440.

Shengfen Luo and Maosong Sun. 2003. “Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures”. In *Proceedings of The Second SIGHAN Workshop, on Chinese Language Processing .ACL2003* :24-30

Qing Ma and Fei Xia. 2003. Proceedings of The Second SIGHAN Workshop, on Chinese Language Processing .ACL2003 Sappro

Sujian Li,Houfeng Wang,Shiwen Yu and Chengsheng Xin. 2003. “News-Oriented Automatic Chinese Keyword Indexing”. In *Proceedings of The Second SIGHAN Workshop, on Chinese Language Processing .ACL2003* :92-97

Nakagawa, H. and Tatsunori Mori. 2003. Automaic Term Recognition based on Statistics of Compound words and their Components", Terminology 9(2) :201-219

Smadja, F.A. and Mckeown, K.R. 1990. “Automatically extracting and representing collocations for language generation”. In *Proceedings of the 28th Annual Meetings of the Association for Computational Linguistics* :252-259.

Uchimoto, K., S.Sekine, M. Murata, H.Ozaku and H. Isahara. 2000. “Term recognition using corpora from different fields”. Terminology 6(2) :233-256

Wenfeng Yang and Xing Li. 2002. “Chinese keyword extraction based on max-duplicated strings of the documents”. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and develop-ment in information retrieval*, pp. 439-440.

Kevin Zhang and Qun Liu. 2004. ICTCLAS. *http://www.nlplab.cn/zhangle/morphix-nlp /manual /node12.html*

Appendix A: A part of stop word list

本月 乒乓 扑通 比较 毕竟 必定 必然 嘻嘻
也不 别看 别说 何必 哎呀 我国 起来 来着
啊 按 吧 把 被 比 彼 必 边 便 别 并 不
到 得 等 点 顶 都 对 吨 多 俄 而 耳 尔

Appendix B: An example of news article

藏人文化网站正式启动

本报兰州电 记者阿旦增报道：由甘肃雪域藏人文化传播有限责任公司创办的藏人文化网站（www.tibetcul.com），经过近一年的筹建，日前在兰州正式启动。

据介绍，以“盘点藏族文化精英，整合藏族文化队伍，展示藏族文化规模，弘扬藏族文化精神”为宗旨的藏人文化网是一个纯文化专业网站。网站设有唐古拉风、藏人之友等 9 个频道。网站总监旺秀才丹表示，藏人文化网将倾力关注当代藏族人的文化和生活，反映当代藏族人的精神世界以及他们在 21 世纪社会转型和文化重建时期崭新的精神、文化、生活风貌。

Appendix C: Terms extracted from Appendix B.

Word Based (Top 10 terms with score of equation (4))

文化(12.49), 藏人文化网站(12.36), 藏人文化网(5.47), 藏族人的文化和生活(4.94), 藏族文化精神(4.33), 甘肃雪域藏人文化(4.29), 宗旨的藏人文化网(4.17), 藏人之友(3.88), 藏族文化精英(3.60), 藏族文化规模(3.60)

Character Based(Top 11 terms with score equation (4))

文化(13.49), 藏人文化网(10.00), 藏人(9.58), 藏族文化精英(6.78), 宗旨的藏人文化网(6.01), 代藏族人的文化和生活(5.76), 藏人文化网站正式启动(5.45), 弘扬藏族文化精神(5.22), 展示藏族文化规模(4.38), 纯文化专业网站(4.37), 整合藏族文化队伍(4.27)