A Way to Extract Unknown Words Without Dictionary from Chinese Corpus and Its Applications

Yih-Jeng Lin, Ming-Shing Yu, Shyh-Yang Hwang and Ming-Jer Wu
(林義証) (余明興) (黃世陽) (吳明哲)
Department of Applied Mathematics
National Chung-Hsing University, Taichung, Taiwan

Abstract

We propose a way to detect the unknown words from the corpus. We call such unknown words Chinese frequent strings(CFS). The strings could be the combinations of some common Chinese words that are defined in a traditional dictionary. Such Chinese frequent strings appear more than once in some Chinese texts. The method we proposed can automatically detect such strings without using any lexicon, and no word segmentation is needed.

We retrieve 55,518 Chinese frequent strings (reached for 13-gram in character) from a corpus consisting of 536,171 characters. To show that the strings we got are useful, we use these strings in Chinese phoneme-to-character and character-to-phoneme tasks. The test corpus contains manually-tagged phonetic symbols for each character. The correctness of the phoneme-to-character test is 96.5% and the correctness of the character-to-phoneme test is 99.7%. We make an MOS test about the determination of prosodic segments. The MOS score is 4.66 relative to the prosodic segments in spontaneous speech. This shows that the strings we retrieved are helpful in this aspect.

Keywords: *unknown words, phoneme-to-character, character-to-phoneme, prosodic segment*

1. Introduction

An intact Chinese electric dictionary is required in the processing of Chinese natural

language (Chang 1997). Such dictionary plays an important role in machine translation, text-to-speech system, speech recognition system, and intelligent Chinese input methods. Yet there are many unknown words or new words coming into being in the world. The unknown words are various and diversified. Such as nominal compounds, verbal compounds, personal names, organization names and their abbreviations (Chang 1997, Chen 1997). There are many works focus on this problem recently (Chang 1994, Chang 1997, Chang etc. 1994, Chen and Bai 1997, Chen and Bian 1997, Chen 1994, Chien 1995, Sun 1994, Wu 1994).

We will propose a method to extract unknown words from corpus in Sect. 2. In Sect. 2, we also try to find some unknown words in a Chinese corpus with phonetic symbol for each character by our proposed method. And we applied the words we extracted to three applications in Sect. 3 to show that the words we get are reasonable and useful. We will discuss about our method in Sect. 4. We make some conclusions in Sect. 5.

2. The Proposed Method

Let's give definition to a Chinese unknown word at the beginning of this section. A Chinese unknown word is a Chinese string that is used frequently by the people. There are about 13000 Chinese characters. The number of combinations of several characters is very large. But there could be very few combinations meaningful. Such meaningful combinations are Chinese words in the lexicon or unknown words if they do not appear in the lexicon. It is more appropriate to use Chinese frequent strings (CFS) instead of Chinese unknown words. For example, "不得不去讀書" (can not help but study) is a Chinese frequent string since we find such Chinese pattern in many Chinese texts. It is not a word but a combination of some words that presents some meaningful idea.

2.1 Extracting Chinese Frequent Strings from Corpus

If a combination of some characters is a meaningful pattern, such combination will very

likely appear more than once in a large corpus. We make a Chinese string pattern be a Chinese frequent string if it appears twice or more in the corpus. Since not all the strings are meaningful. It is very important to let a computer make a decision that which pattern is meaningful.

Our method is divided into two steps. The first step is searching for all characters in the corpus to determine which patterns appear more than one time. Such patterns will be gathered into a database that is so called "may be a word" database. The entries in the "may be a word" database are the strings and their number of occurrences. The second step is to find the raw frequency for each entry in the database mentioned just now. The raw frequency of a entry is the frequency of self appearance of the entry. And we can decide which patterns are meaningful patterns according to the raw frequency.

2.2 Constructing the Database of "may be a word"

The first step is to construct the database "may be a word" from corpus. In this step we collect each pattern that the number of characters of the pattern is less than or equal to 15. The number of characters of every entry in the database is not greater than 15. This is because that a breath group contains no more than 13 characters according to the experiments in our group (Pan 1998, Jen 1997). We need to find the strings of length two or more than 13. The reason will be explained later.

The frequency of each entry in the database is greater than or equal to two. Yet not all the entries are the patterns we want. Some of them are nonsense. For instance, consider the following fragment:

"從二次大戰以後，自然科學的重要，幾乎沒有人不知道了。戰爭固然需要自然科學，和平更需要自然科學。「原子能和平用途」的目標，差不多是全世界文明人類的希望所寄託的。要達到這個目標，自然科學的研究是唯一的途徑。但我覺得現在世俗有一種誤解，就是許多人都以爲自然科學的用處只在增進物質上的文明，而和人類的精神文明沒有關係。要討論青年的求學問題，似乎應該先在這點上做匡謬正俗的功夫。所謂精神文明，當指人類道德的觀念和社會的組織各端而言。譬如仁義的敷教，法律的制定，都是精神

文明範圍以內的事。自然科學愈進步，則根據這些科學而制定的法律必更適合於文明人群的需要，至於「博愛之謂仁」，「行而宜之之謂義」，也不是每個人照著自己的意思可以定得好的。”

There are 36 patterns that appear more then once. They are listed in the first column of Table 1. There are some patterns that we may not wish to treat them as the unknown word. Such as “然科”, “然科學”, and “然科學的”, etc. However, The computer has no idea to get rid of such items. They can work well by computing. The rest task of our method for extracting unknown words is to identify the patterns automatically using a computer. This the second step in our proposal.

2.3 Extracting Unknown Words form Database “may be a word”

We have constructed a database that each entry may be a new word in the previous subsection. The main idea of this subsection is to make sure which pattern we indeed need. We will exclude a pattern whose appearance is due to the appearance of a longer pattern. For instance, all the occurrences of “然科” are caused by the occurrences of “自然科”. So “然科” will not be an unknown word. Again since “自然科學” are brought in by “自然科學的” and “需要自然科學”. The frequency of “自然科學” appear by itself is only 1. The frequency that a possible word appear by itself is shown in the third column of Table 1.

We could extract the entry which frequency of self appearance is more than once to be as our unknown words according to the third column of Table 1.

2.4 Implementation

We use a fraction of the Academia Sinica Balance Corpus as our training data (Chen 1996). The corpus we used is raw text. There is no information of segmentation. The total number of characters is 536,171. And the number of entries in “may be a word” database is 115,140. We applied the method mentioned in subsection 2.3 to decide which entry in the database is a real word that appears twice or more truly in the training corpus. We get 55,518

words. Appendix shows some high frequency words we extracted which can not be found in an ordinary dictionary.

Table 1. The listing of string patterns.

String Pattern	Occurrences	Frequency of self appearance
人類	3	1
人類的	2	2
之謂	2	2
文明	6	1
文明人	2	2
目標	2	2
自然	6	0
自然科	6	0
自然科學	6	1
自然科學的	3	3
沒有	2	2
制定	2	2
和平	2	2
明人	2	0
法律	2	2
科學	7	1
科學的	3	0
要自	2	0
要自然	2	0
要自然科	2	0
要自然科學	2	0
神文	3	0
神文明	3	0
然科	6	0
然科學	6	0
然科學的	3	0
精神	3	0
精神文	3	0
精神文明	3	3
需要	3	1
需要自	2	0
需要自然	2	0
需要自然科	2	0
需要自然科學	2	2
學的	3	0
類的	2	0

3. System Applications

In order to make sure that the words we retrieved are useful, we applied the words we retrieved to three tasks, namely Chinese character-to-phoneme(CTP) conversion, Chinese phoneme-to-character(PTC) conversion, and the determination of breath groups in an input

Chinese sentence.

3.1 The Chinese Character-to-Phoneme Task

The Mandarin text-to-speech system needs a character-to-phoneme system to get the correct syllables (Lin 1998, Ouh-Young 1985, Pan 1998). To have high performance of Chinese character-to-phoneme, we applied the words we retrieved in this aspect. The lexicon we used is the combination of Academia Sinica dictionary and the words we retrieved from the training corpus.

There are four principles in our CTP task. They are (1) the number of words should be minimized. (2) The number of characters in a word should be as many as possible. (3) The number of mono-character word should be minimized. (4) The probability of the combination of words should be high.

We performed an outside test for the Chinese character-to-phoneme task. The test corpus consists 82,610 characters with corresponding phonetic symbols for each character. To ensure the correctness, we check the phonetic symbols manually. The correctness of outside test is 99.7%.

3.2 The Chinese Phoneme-to-Character Task

The second work is the reverse of the work mention in subsection 3.1. That is Chinese phoneme-to-character task. This task is more difficult. There are some studies in this problem (Hsu 1995, Ho 1997). As we know, the best performance is 96% that is done by Hsu in Academic Sinica (Hsu 1995). Two possible applications of Chinese phoneme-to-character are Chinese speech-recognition system and Chinese input system. There are many methods to solve the problem. Three approaches have been proposed for this problem, (1) is the statistical approach, (2) is the grammatical approach, and (3) is the semantic template.

We applied the lexicon mentioned in subsection 3.1 to finish the task. We also use the same four principles in CTP to the task of PTC.

The test of Chinese phoneme-to-character is inside test. The test corpus contains 536,171 characters with corresponding phonetic symbols. The correctness of this test is 96.5%. Some errors like "它", "他", "她", "牠" are not identified by our system. Such an identification is generally considered an difficult problem. However, the proper names we have trained could be identified by our strategies.

3.3 The Determination of Prosodic Segments

The third task we do is the determination of prosodic segments or breath groups (Lopez-Gonzalo 1997, Pan 1998, Jen 1997) in a sentence. The prosodic segments are important for a Mandarin text-to-speech system. To show the words we extracted are useful in this aspect, we try to decide the prosodic segments in input Chinese sentence. We have analyzed the real speech to get the prosodic segments. We found that a prosodic segment is about 7 to 10 characters. And a prosodic segment contains one or more words. No word will be divided to belong to different prosodic segments.

The evaluation we take is objective MOS evaluation (Wei 1997). The evaluation data are two paragraphs, say paragraph A and paragraph B. Two versions of the prosodic segments of these two paragraphs are provided. One version is obtained by analyzing the real speech and the other is got by our system. There are forty-two undergraduate students in the Department of Chinese Lecture in our university making the evaluation. The students are divided into two groups. One group takes the evaluation in paragraph A with prosodic segments determined according to real speech and paragraph B with prosodic segments determined by our system. The other group takes the evaluation in paragraph A with prosodic segments determined by our system and paragraph B with prosodic segments determined according to real speech. There are totally 50 sentences for every student.

The results are 3.834 for the prosodic segments of real speech and 3.572 for the prosodic segments of our system. The relative score is 3.572/3.834*5=4.66. This means that the words

we retrieved make good help in determining the prosodic segments. Since some meaningful patterns or proper names can be identified by our system. The characters in such a pattern should belong to the same prosodic segment.

4. Discussion

The features of our extraction of unknown words can be summarized in the followings

(1) Our method can extract any unknowns for corpus. Such unknowns could be people names, origination names, verbal compounds, foreign translated nouns, and so no.

(2) We need no dictionary while extracting unknown words. And no Chinese word segmentation is needed in preprocessing.

(3) We could find long word patterns in one pass.

(4) There is not any rule applied in our system. The generation of rules very often requires many human works.

(5) The computation of frequency for each word is very accurate in our system. We have proposed a method to compute the frequency of occurrence for each word.

(6) There is no part-of-speech information needed. The use of part-of-speech information generally requires a large amount of human involvement.

5. Conclusions

We have proposed a robust way to extract unknown words form corpus. We also show that the words we retrieved are very useful. Such words make good help in Chinese phoneme-to-character, Chinese character-to-phoneme, and the generation of prosodic segments in a Mandarin TTS system.

Some future works we want to do are in the following three aspects.

(1) To extract more unknown words in a large corpus.

(2) Developing a robust system for Chinese phoneme-to-character task.

(3) Developing a high performance Mandarin text-to-speech system.

References

[1] C. H. Chang, "A Pilot Study on Automatic Chinese Spelling Error Correction," Communication of COLIPS, Vol.4, No.2, pp.143-149, 1994.

[2] J. S. Chang, "Automatic Lexicon Acquisition and Precision-Recall Maximization for Untagged Text Corpora," Ph.D. Thesis, National Ching-Hua University, 1997.

[3] J. S. Chang, S. D. Chen, S. J. Ker, Y. Chen, and J. Liu, "A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts," Computer Processing of Chinese and Oriental Languages, Vol. 8, No.1, pp. 75-85,1994.

[4] K. J. Chen and M. H. Bai, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," Proceeding of ROCLING X, pp.159-174, 1997.

[5] H. H. Chen and G. W. Bian, "Proper Name Extraction from Web Pages for Finding People in Internet," Proceeding of ROCLING X, pp.143-158, 1997.

[6] K. J. Chen, C. R. Hunag, L. P. Chang, and H. L. Hsu, "SINICA CORPUS: Design Methodology for Balanced Corpora," Proceeding of PACLIC 11th Conference, pp.167~176, 1996.

[7] H. H. Chen and J. C. Lee, "The Identification of Organization Names in Chinese Texts," Communication of COLIPS, Vol.4, No.2, pp. 131-142, 1994.

[8] L. F. Chien, "尋易(Csmart)—A High-Performance Chinese Document Retrieval System," Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages, ICCPOL'95, pp. 176-183, 1995.

[9] Eduardo Lopez-Gonzalo, Jose M. Rodriguez-garcia, Luis Hernandez-Gomez, and Juan M. Villar, "Automatic Prosodic Modeling for Speaker and Task Adaptation in Text-to-Speech," Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 927-930, 1997.

[10] W. L. Hsu, "Chinese Parsing in a Phoneme-to-Character Conversion System Based on Semantic Pattern Matching," International Journal on Computer Processing of Chinese and Oriental Languages 40, pp.227-236, 1995.

[11] T. H. Ho, K. C. Yang, J. S. Lin, and L. S. Lee, "Integrating Long-Distance Language Modeling to Phoneme-to-Character Conversion," Proceeding of ROCLING X, pp.287-229, 1997.

[12] Y. J. Lin and M. S. Yu, "An Efficient Mandarin Text-to-Speech System on Time Domain," to appear in IEICE Transactions on Information and Systems in June 1998.

[13] M. Ouh-Young, "A Chinese Text-to-Speech System," Master thesis, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, June 1985.

[14] N. H. Pan, "Prosody Model for Syllable Energy and Intonation in a Mandarin Text-to-Speech System," Master thesis, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan, June 1998.

[15] W. T Jen, "Prediction Models for Syllable Duration in a Mandarin Text-to-Speech System," Master thesis, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan, June 1997.

[16] M. S. Sun, C. N. Huang, H. Y. Gao, and Jie Fang, "Identifying Chinese Names in Unrestricted Texts," Communication of COLIPS, Vol.4, No.2, pp. 113-122, 1994

[17] H. N. Wei, "An Approach to the Measurement of Fondness and Similarity on Speech," Master thesis, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan, June 1997.

[18] Zimin Wu and Gwyneth Tseng, "Chinese Text Segmentation for Text Retrieval: Achievements and Problems," JASIS, 44(9), pp. 532-542, 1994.

Appendix: Some unknown words we extracted from corpus

Unknown Words	Frequency	Unknown Words	Frequency
的人	114	在日常生活中	17
他說	74	中央研究院的	8
也是	71	五十而知天命	8
他的	70	生活與工作的	7
這是	65	人活在世界上	6
這個	62	民族學研究所	6
這種	59	每個人都可以	6
也有	53	近代史研究所	6
爲了	49	己欲立而立人	4
的時候	79	人力資源的開發	8
的問題	68	自我的重新探索	6
事實上	53	金車文教基金會	5
這就是	51	做爲一個現代人	5
的工作	48	本院數理組院士	4
的生活	44	在生活與工作中	4
這樣的	42	研討會議程如下	4
有一位	39	學者社區的形成	4
的方法	39	七十而從心所欲	3
的關係	39	人與人之間的關係	5
一九九二	24	山豬窟垃圾掩埋場	5
我們可以	20	數學研究所研究員	5
山西商人	19	人多的地方不要去	4
我們應該	18	天下無不是的父母	4
的年輕人	18	不是一件容易的事	3
我們必須	17	本院歷史語言研究所	4
一個人的	15	不可或缺的靈魂人物	3
千萬不要	15	民族學研究所研究員	3
有很大的	15	一個完全不同的人格	2
這是一個	15	一個發出菩提心的人	2
學者的社區	13	歷史語言研究所研究員	4
人與人之間	12	一場沒有終止線的競爭	2
的觀點來看	11	人生最大的本錢是尊嚴	2
我們可以說	9	不要擔心別人不了解我	2