The Identification and Classification of Unknown Words in Chinese

An N-Grams-Based Approach

Mei-Chu Wang, Chu-Ren Huang, and Keh-jiann Chen

CKIP, Academia Sinica

{kchen, churen}@iis.sinica.edu.tw

1. Introduction

In this paper, we propose a new approach to identify unknown words in Chinese. This approach adopts an n-grams program to sort out the collocating word / character sequences which are possible words and phrases in Chinese. In addition to proposing the criteria for identifying Chinese new words, was also classify these new words according to their structural and semantic characteristics.

The corpus-based approach in identifying Chinese disyllabic words based on mutual information was first studied by Sproat and Shih [1]. The attempt here is to identify Chinese unknown words by collocations. Collocations are sequences of words that tend to appear together. In this paper we describe a set of techniques based on statistical methods for retrieving and identifying unknown words from a Chinese corpus. Here unknown words refer to words that are not included in the 90,000 entries CKIP Electronic Dictionary developed at Institute of Information Science, Academia Sinica. The results retrieved by the n-grams program will be crucial information for updating dictionaries.

The n-grams program locates words in context and makes statistical observations to identify collocations. It produces a wide range of collocations which can be further sub-classified as abbreviational words, derived words, proper names, new words, ambiguous words, and collocating strings. The effectiveness of the n-grams program as a retrieval tool for unknown words is measured and evaluated.

2. What Are Unknown Words in Chinese?

Burchfield [2] observes that novel compounds are often used by the media, the computer industry, political circles, as well as by scientists and technologists. Since our data is retrieved from newspaper reportage, it is very likely that there are words which were newly-coined, or borrowed from other languages.

Generally speaking, these words are at first unknown to readers. After the ideas have been spread and accepted by the community, the unknown words become lexicalized words. Downing [3] proposes that the newly-coined word is semantically highly transparent; but once it has been accepted by the community as a conventionalized lexical item, it may come to be as arbitrary as monomorphemic lexemes.

Here unknown words refer to words that are not registered in the CKIP lexicon. The CKIP lexicon is a general purpose lexicon which provides the lexical

information for identifying Chinese words and parsing Chinese sentences. It is necessary to patch unknown words into the CKIP lexicon to maintain its usability. In addition, as a general purpose lexicon, CKIP lexicon does not often in-depth coverage of special domains. Thus acquisition of unknown words will be essential when this lexicon is applied to special domains.

3. Six Categories of Unknown Words

We classify the potential unknown words identified in the study into six categories depending on whether they are generated by a productive morphological rule or not.

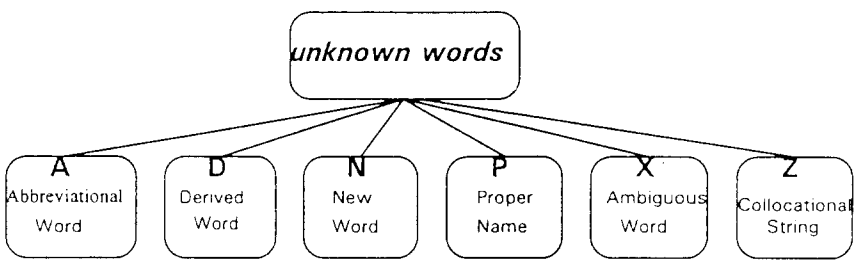

Figure 3.1 The sub-classifications of unknown words in Chinese

As shown in figure 3.1, we proposed that Chinese unknown words can be subcategorized into six classes. These are A: Abbreviational Word; D: Derived Word; N: New Word; P: Proper Name; X: Ambiguous Word, and Z: Collocational String.

3.1 Abbreviational Word

According to Huang et al. [4], Suoxie "abbreviation" is a specific rule in Chinese that derives new lexical items from longer corresponding forms. They assume that such processes are part of the mental lexicon. In general, words formed by abbreviation take one character (syllable) from each word in a compound/phrase, though not necessarily the first character in each word. However, the complexity is that any native speaker knows how to form new words by such a method even though these rules do not seem to follow the form and nature of traditional morphological rules. Abbreviation is a very productive new word formation process in Chinese. From the retrieved results it is clearly shown that most of the abbreviations come from the names of institutions, companies and organizations. For example, 工技學院(工業技術學院) (Industry Technology College) "National Taiwan Institute of Technology", 市銀 (台北市銀行) (Taipei City Bank) "Taipei City Bank", 消基會 (消費者文教基金會) (Consumer Cultural Foundation) "Consumer Foundation of ROC". Apart from the above point, it is worth noting that there are more abbreviational nouns than verbs. Abbreviational verbs are few, such as 申購 (申請購買) (apply and purchase)，聯檢〔聯合檢查〕(joint investigation)。 Common noun abbreviation is also rare. Examples are 片商〔影片商人〕(film merchant)，職棒〔職業棒球〕(professional baseball)。

Center), 立陶宛 (Lithuania) . It should also be noted that by their nature proper names cannot be listed exhaustively in the lexicon and that in Chinese there are no direct heuristic devices, such as capitalization, by which to identify them in written texts.

3.5 Ambiguous Word

"Ambiguous words" are words whose lexical reference value cannot be determined precisely, either because of the possibility of semantic transfer (eg.民主之聲 "voice of democracy" or a broadcast station of that name) , or because of alternative structure analyses (eg.新建工程局 to rebuild the bureau of engineering, or the bureau of new engineering constructions). The reason we have established this category is to make sure that the information conveyed by the words is totally collected by the annotators. We wish to avoid imposing the subjective points of view of the annotators. The purpose of this category is to provide more imaginative space to annotators.

3.6 Collocating Strings

There are a lot of collocating 2- and 3-grams which do not belong to the above categories. The name "collocating string" was given to them because entities identified in this way are neither words nor phrases, yet can be regarded as some kind of grammatical unit. Apparently, these items play a connecting or tone-shifting role within a sentence. Knowledge of such sequences must be stipulated in a generative framework for language parsing and speech recognition and will be very useful in automatic segmentation. Examples: 一定 是 (must be) , 乃 於 (as for) 。

4. Algorithm and Methodology

Except in a very few cases, such as Sproat and Shish's [1] research using mutual information to identify di-syllabic known words and Chang et. al.'s [9] study on proper names, previous work on Chinese word identification has been conducted by manual investigation. Here we propose a method by investigating collocations in a large corpus to find the unknown words in Chinese. The algorithm, called the n-grams program, automatically segments words and phrases; then retrieves high-frequency collocations. The selected collocations are then classified by manual investigation.

The procedure undertaken in the n-grams program is a combination of computer retrieval and manual post-editing.

According to Smadja [10], a collocation can simply be considered as a sequence of words (or n-grams) that frequently appear together among millions of other possible sequences. The function of an n-grams program is to locate words in context and make statistical observations so as to identify collocations. Before locating collocates in contexts, an automatic segmentation routine is employed to find words and phrases in a sentence. Since words are not conventionally marked and the collocating strings in our corpus are formed by either known words or single characters, the collocating strings identified can be either an unknown word

3.2 Derived Word

Words that are derived by morphological rules (compounding or derivations) are taken as "derived words." According to Tang [5], derivative words are made up of a stem and one or more than one affix. In addition, these affixes can clearly specify their categories or change categories from one to another. Affixes can be further sub-classified as prefix, infix and suffix.

Hong et al. [6], compiled a comprehensive list of common Chinese affixes: including the prefixes 可 (-able), 第 (-th), 初 (first), 老 (old); the infix 得 (-able, as in 跑得快，看得遠); and the suffixes, which form the largest affix class in Chinese: 化 (-ize), 性 (-ness), 度 (-ity), 員 (-er), 師 (-er; professional), 者 (-er). However, the division between compounds and derivative words is not clear-cut. Words with 機 (machine) and 器 (utensil) can be considered either as compounds or as derivative words, e.g. 洗衣機 (washing machine)，遙控器 (remote control device). For this reason, our criterion for "derived word" identification is extended to the more general notion of word formation. Derived words in the sense of this report are words formed either by the process of compounding or by morphological affixation rules.

Verb resultative compounds are one type of derived word that always appear in bi-gram. According to Lin [7, 8], VR compounds are also very productive in Chinese. In these compounds the second component is the result caused by the first component, for instance, 沖 毀 (flood and destroy)，拆 遷 (dismantle and move)，挖 斷 (dig and break)。

3.3 New Word

As the name implies, "new words" are words that are newly-coined. The reason for such coinages can be simply attributed to the appearance of new concepts or products. New words can be classified into two types, native neologism and loan words. Examples for native neologisms are 八 點 檔 (a TV program shown at 8:00 p.m., a prime-time TV program), 電 聯 車 (a transportation car). Loan words can be sub-categorized as translation and transliteration. However, in actual use, a single word often involve both strategies. For instance, 寶 特 瓶 (PET bottle) and 登 革 熱 (Dengue fever) both translate the head noun while transliterate the pre-modifier.

3.4 Proper Name

"Proper name" is a category, in which new words are increasing particularly rapidly. This is because proper names can not be exhaustively listed in the lexicon and because new objects, especially places, organizations and people, are constantly coming into existence or brought to the attention of the public. Examples are 尤 清 (Yo Chin, Magistrate of Taipei), 李 鵬 (Li Peng, Prime Minister of Mainland China), 中國 生產力 中心 (China Productivity

distribution of collocating 2- to 8-grams is given in Table 5.4. It needs to be clarified that the units in any collocating sequences may be either characters or words. Based on the CKIP lexicon word identification is carried out by an automatic segmentation program [11]. It is estimated that the automatic segmentation program identifies more than 97% of the words in texts. The remaining one-character elements could be either one-character words or part of the 3% unknown words. For instance, 大陸 客 (mainlander), 生 鮮 超 市 (fresh food supermarket) are recognized and identified as 2- and 3-grams respectively, because 大_陸 and 超 市 are listed in the lexicon while 大_陸 客, 生 鮮 and 生 鮮 超市 are not.

5.1 Bi-collocation

As its name implies, "bi-collocation" refers to textual entities with *two* collocating units. It is explained above that the units are characters or words when automatic dictionary lookup is successful. A bi-collocation might be composed of two independent characters, 拒 收 (refuse to accept), one word and one character, 大陸 客 (mainlander), or two words, 活動 中心 (activity center). Above examples are collocations retrieved from bi-grams. Table 5.1 shows the distribution of bi-collocations over the categories defined above.

category	a	d	n	p	x	z
number	154	330	494	160	46	1819

Table 5.1 The statistical analysis of bi-collocations by categories

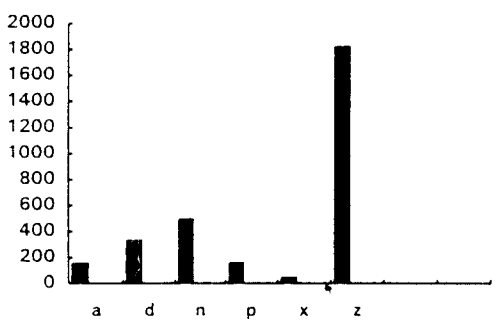

Figure 5.1 The statistical analysis of bi-collocations by categories

According to Figure 5.1, non-word collocational strings (category Z) predominate in bi-collocation. In other words, binary grammatical linked pairs are both most common and most useful. In addition, the number of "new words" (category N) is the second highest among the six categories. This implies that new word formation is very productive in bi-grams.

or a phrase. More precisely, a collocation (n-grams) is selected as a possible word provided that both the frequency of occurrence and the value of the association with the adjacent words are above a threshold number. The procedure undertaken by the n-grams program was as follows.

First, for each word x, the corpus was examined to find its potentially lexical collocates x_i. Secondly, the distribution of the frequencies $freq_i$ of its collocates x_i was analyzed, and its average frequency f and standard deviation σ around f were computed according to (4.1) & (4.2). Then, the association strength k_i of the word pair x and x_i was derived according to (4.3). A collocate was selected if the association strength of the word pair x and x_i was greater than a threshold value c.

$$f = \frac{\sum_{i=1}^{n} freq^{i}}{n} \text{----------(4.1)}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (freq^{i} - f)^{2}}{n}} \text{----------(4.2)}$$

$$k_i =_{def} \frac{freq^{i} - f}{\sigma} \text{----------(4.3)}$$

The above process was iterated to find bi-collocations, tri-collocations,...and up to eight-collocations. Although the statistics-based automatic process makes efficient word identification possible, the result is not 100% accurate. There are roughly 3% false recalls that had to be sorted out by manual post-editing. There is no denying manual post-editing plays a very important role in an n-gram-based approach. It can be said that the n-grams program find out the possible words and phrases, while manual post-editing picks up the real words and phrases. This back-up procedure is necessary to make the list of unknown words correct and complete.

The remaining collocations after manual post-editing were classified into the six categories discussed above or discarded. The collocations which were discarded were all clearly either sentential phrases or parts of phrases, for example: 技術 研究(technique research), 有關 單位 (related organization).

5. Analysis of Results

The results were obtained from a corpus of 4,024,370 characters composed of 2,644,183 known words and forming 351, 026 sentences. The

5.2 Tri-collocation

category	a	d	n	p	x	z
number	69	21	74	697	25	26

Table 5.2 The statistical analysis of tri-collocations by categories

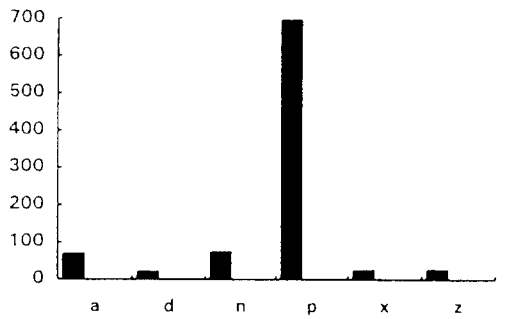

Figure 5.2 The statistical analysis of tri-collocations by categories

Table 5.2 and Figure 5.2 show the unequivocal preponderance of proper names (category P) among tri-collocations. This is because most Chinese personal names are composed of three characters. The numbers of new words and abbreviational words are the second and third highest in 3-grams. It indicates new word formation is relatively productive in 3-grams.

5.3 Quad-collocation

category	a	d	n	p	x	z
number	8	0	26	108	3	0

Table 5.3 The statistical analysis of quad-collocations by categories

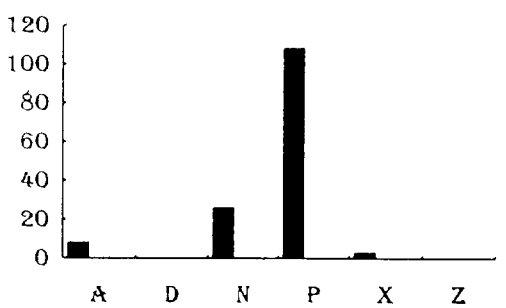

Figure 5.3 The statistical analysis of quad-collocations by categories

These results show that except for a number of new words (category N) accounted for by new inventions or products, most of the quad-collocations can be classified as idiomatic phrases. In addition, there are virtually no derived

words or collocational strings. This is explained by the fact that four characters are the typical length of idiomatic phrases and proverbs in Chinese. Meanwhile, the number of the proper names (category P) is the highest among all of the categories. As for the abbreviational words (category A), all of them are the shortened forms of the names of institutes, companies and organizations.

5.4 Five, Six, Seven and Eight Collocations

All of the examples amongst these four longer collocations types belong to the category of proper names that are names of people, places, order, and organizations.

5.5 Discussion

Collocation	Two	Three	Four	Five	Six	Seven	Eight	Total
number	13,857	3,208	1,181	549	281	160	85	19,321

Table 5.4 The distribution of collocations by length

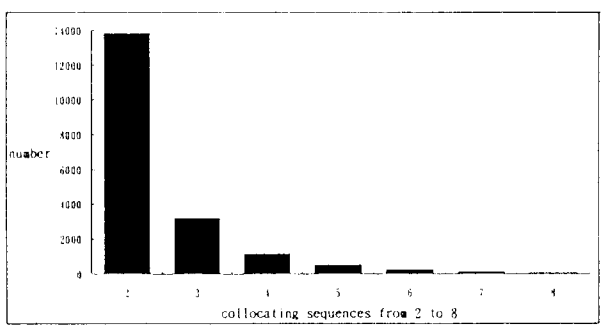

Figure 5.4 The distribution of collocations by length

The numbers in Table 5.4 represent collocations of 2- to 8-grams that are (possible words or phrases in Chinese) retrieved by the n-grams program. While the numbers in Table 5.5 are the numbers of real words and phrases in Chinese that are obtained after manual post-editing and examination.

	a: abbreviation	d: derivation	n: new word	p: proper name	x: uncertain	z: collocating strings	discarded strings	selected/ total	real words/ phrases %
2	154	330	494	160	46	1,819	10,854	3,003/13,857	21.67
3	69	21	74	697	25	26	2,296	912/3,208	28.43
4	8	0	26	108	3	0	1,036	145/1,181	12.28
5	0	0	0	25	0	0	524	25/549	4.55
6	0	0	0	6	0	0	275	6/281	2.14
7	0	0	0	4	0	0	156	4/160	2.50
8	0	0	0	5	0	0	80	5/85	5.88

Table 5.5 The statistics of unknown words in Chinese by category

Figure 5.4 shows that collocations come mostly in the length of 2 characters and decrease as length increases. Furthermore, although only 21.22%

(4,100/19,321) of collocations are useful (including unknown words) and more than 78% are discarded, the searching space for discovering new words is reduced from a multi-million words corpus to 19,321 collocations.

The "collocational strings" (category Z) that appear in 2-grams are grammatical discourse linkage entities that occur in context. This suggests that it is feasible to accumulate such strings in a file to facilitate new word selection. That is, before we undergo the procedure for new word selection these strings can be always discarded first. As far as the "new word" (category N) is concerned, the zero number among 5- to 8-grams indicates that new word search is only needed in 2- to 4-grams. As for 3- and 4-grams, the results show that "proper names" constitute 3/4 of the "unknown" word types. That is one of the salient features of Chinese "unknown" words. Above all, 3-grams excluding the "proper names" has the highest percentage of "new words."

6. Related Work

The n-grams-based approach to Chinese unknown word identification is to locate collocations that are possible words and phrases. Based on the criteria proposed for the classification of Chinese unknown words, manual post-editing is applied to the retrieved collocates.

Although the n-grams-based approach to new word selection is conclusive, it is not faultless. The n-grams program cannot find words with low frequencies. A means to compensate for this disadvantage is the tagtool program. [11]

The tagtool program is a tagging tool for tagging a Chinese corpus. It segments Chinese sentences into words sequences and provides their syntactic categories. Editing functions are also provided by the tagtool to revise possible errors caused by automatic segmentation and tag assignment. A profile of revised results including new words will be reported after manual editing. The reasons for wrong segmentation come from ambiguous segmentation or the lack of an entry in our lexicon. [11] If the reason for wrong segmentation is the later one, then a new word is identified by the manual error checking procedure and added in our lexicon. In this a way the n-grams program provides a way of finding most frequent new words, thereby drastically reducing possible segmentation errors, while low frequency unknown words are discovered through the tagging process. Examples retrieved by the tagtool program are 高階 (high level as in "high level language"), and 先斬後奏 (kill before report).

7. Conclusion

The n-grams method for Chinese unknown word identification looks at 2- to 8-grams, and sorts out possible words and phrases in Chinese. Its advantages lie in its effectiveness for new word finding. New words can be regarded as a type of collocation with fixed components in a fixed order and of relatively high frequency. The n-grams-based approach applies this algorithm and finds out not only the Chinese unknown words but also the new words and phrases in Chinese.

The findings through the n-grams program tell that the collocational strings (category Z) in bi-grams can be always omitted before we start new word selection. In addition, new words often appear in 2- to 4-grams. This implies that it is not necessary to use 5- to 8-grams for new word search. The distribution of proper names among 3- and 4- grams is one of the salient features for the identification of Chinese unknown words. The performance of the n-gram program comes to its optimum when applied with 3-grams. If efficiency is a main concern, our study also show that we could limit an unknown word search to 2- to 4-grams. This greatly reduces the computing time while still covers 99.25% (4,060/4,100) of possible unknown words, and even has a higher recall rate of 22.25% (4,060/18,246).

The tagtool program functions as an extra-fine filter for new word that escapes the collocational method. Its report on words and phrases that do not exist in our lexicon can be compiled to supplement automatic new word identification. Since the purpose of tagging is to clearly and correctly mark every word in the text, we can be sure that each and every new word will be identified. New words found in this way could be words and phrases that cannot be obtained by the n-grams program because of low frequency in the corpus. Thus, the use of the n-grams and tagtool programs together ensures that the findings identification of new words is more accurate and more complete.

To sum up, the n-grams-based approaches are shown to be versatile in Chinese language processing. In addition to the difficult unknown word identification and classification problem discussed in this paper, similar approaches have been adopted to finding collocation without segmentation [12], identification of proper names [9], and extracting new terms [13].

8. References

[1] Sproat, Richard, and Chilin Shih. 1990. 'A Statistical Method for Finding Word Boundaries in Chinese Text'. *Computer Processing of Chinese and Oriental Languages. Vol. 4*, N. 4336-351.

[2] Burchfield, R. 1992. *Points of View: Aspects of Present-Day English.* Oxford: Oxford University Press.

[3] Downing, Pamela 1977. 'On the Creation and Use of English Compound Nouns'. in *Language 53. 810-41.* Baltimore: the Linguistic Society of America.

[4] Huang, Chu-Ren, Kathleen Ahrens, and Keh-jiann Chen. 1993. 'A Data - Driven Approach to Psychological Reality of the Mental Lexicon: Two Studies on Chinese Corpus Linguistics, in *Proceedings of International Conference on the Biological Basis of Language*,.pp. 53-68. Chung Cheng University Press.

[5] Tang, Ting-Chi. 1989. *Studies on Chinese Morphology and Syntax: 2* (in Chinese). Taipei: Student Book Co.

[6] Hong, Wei-Mei, Chu-Ren Huang, Chih-Chen Tang, and Keh-Jiann Chen. 1991. 'Towards Morphological Rules of Mandarin Derived Words'. Working Papers in the Third Chinese Teaching Symposium, Taipei.

[7] Lin, Fu-Wen. 1992. 'On V-N Compounding Nouns' (in Chinese) in *Technical Report no. 92-02.* Taipei: Academia Sinica.

[8] Huang, Chu-Ren, Fu-Wen Lin. 1992. "Composite Event Structures and Complex Predicates: A Template-based Approach to Argument Selection." Paper presented at the Third Annual Meeting of the Formal Linguistic Society of Midamerica (FLSM III), May 15-17. Northwestern University.

[9] Chang, J. S. 1994. "A Multi-Corpus Approach to Recognition of Proper Names in Chinese Texts." *Computer Processing of Chinese & Oriental Languages.* Chinese Language Computer Society.

[10] Smadja, Frank. 1993. 'Retrieving Collocations from Text: Xtract'. *Computational Linguistics,* Vol 19. No. 1. 143-177.

[11] Chen, K. J. et al. 1994. 'A Practical Tagger for Chinese Corpora'. To be presented at R.O.C. Computational Linguistics Conference VII (1994). Hsinchu, Taiwan, August 12-13, 1994.'

[12] Huang, C. R., K. J. Chen, and Y. Y. Yang. 1994. "Computer-based Collocation for Mandarin Chinese." in *Proceedings of the 15th International Conference on Computational Linguistics* (Section on Computational Linguistics), pp. 540-543. Kyoto, Japan.

[13] Fung, Pascale, and Dekai Wu. 1994. "Statistical Augmentation of a Chinese Machine-Readable Dictionary." *Second Annual Workshop on Very Large Corpora (WVL II).* Kyoto, Japan.

Acknowledgments

The authors of this paper would like to thank Gao Zhao-Ming, for his thorough analysis of all bi-collocations. A special word of thanks in this connection is due to Lin Sung-Chien, who worked so diligently and meticulously to turn the available n-grams programs into something we could use. Detailed comments from Paul Thompson on an earlier version has helped us to clarify some issues as well as to correct some oversights. Any remaining errors are the responsibility of the authors.