Chinese Abbreviations and Expansion
中文縮寫與還原

Chuen-min Huang 黃純敏、Chuan-Pu Yang 楊顓溥
Department of Information Management, National Yunlin University of Science & Technology
國立雲林科技大學　資訊管理系
huangcm@yuntech.edu.tw; g9223727@yuntech.edu.tw

Abstract

The form of abbreviation is commonly used in the Chinese text. For instance, we often transform “台灣鐵路局” into”台鐵局”. This kind of transformation is timesaving and convenient. However, this merit also brings some challenges in Chinese text processing. In keyword-based information retrieval system, using the abbreviated form and the original form as search entry respectively, usually return different results even though both carry the same meaning. In addition, the influences of abbreviation on Chinese word segmentation, automatic documents clustering and weight of terms are obvious.

To solve the semantic ambiguity problem in Chinese text processing, we propose an approach to bridge these two forms and construct an abbreviation list automatically without consulting any dictionary.

Two major tasks are included in our experiments— finding the best candidate of original forms or abbreviations, and recognizing potential abbreviations and original forms in documents. Besides, there are two kinds of procedures including detecting abbreviation of original forms, and executing expansion of abbreviation. Considering feature selection, it is suggested to combine both nouns and POS rather than only process either of them respectively.

Our study shows that the performance of precision with documents from single news category, especially the category of finance, is the best. Thus, our method may be suitable for corpus with static types of words. It is worth further exploring that if the performance could be improved if noise information is removed from contextual information.

Keyword: Chinese abbreviation, Longest Common Subsequence, Feature Selection, Maximum Entropy Principle

摘要

在中文文件中，字詞常以縮寫型態出現，例如:「台灣鐵路局」縮寫成「台鐵局」。這種高度「可縮寫性」之用法雖然為現代人爭取了時效及便利性，但也為中文字詞處理帶來了一些挑戰。像是對於以關鍵字為基礎的資訊檢索系統而言，使用者在進行檢索時所下之關鍵字為「縮寫詞」或是「原形詞」對搜尋引擎而言是兩個不同的詞，造成回傳結果遺失許多資訊。抑或在進行中文斷詞、文件自動分群及字詞權重計算等處理時，亦會對系統效能產生影響。

基於這些問題，本研究提出一個中文縮寫詞或原形詞的對應機制。藉由這樣的機制能夠將文件中的縮寫詞和與之對應原形詞連結起來，而不需依賴任何固定辭典，如同利用語料庫以建立一個動態縮寫詞對照表，也很容易應用到其它語言上。

本研究中，我們以電子新聞文件為語料庫進行了幾項實驗。主要分為選取最佳候選詞、擷取正確縮寫詞及原形詞兩大部分，每部分又分為原形詞對應縮寫詞、縮寫詞對應原形詞兩組流程。實驗結果發現，用詞性配合名詞內容為特徵所訓練出的模型之選取候選詞的精確率最高。實驗結果也發現，單一類別新聞文件之績效比混合類別新聞之實驗結果為佳;而在單一類別新聞文件之中，又以財經類新聞表現最好。因此，本對應機制可能仍限制在品質較穩定之文件中，才能有較好的實驗績效。鑒於新聞文件用詞的多樣性，未來研究必須考慮加強前處理之步驟，試圖將文件集的雜訊降至最低，以更進一步提昇精確率。

關鍵字:中文縮寫詞、最長共同子序列、特徵選取、最大熵

1. Introduction

Abbreviation is often used in the Chinese documents. For example, we used to substitute “台鐵局” for “台灣鐵路局”. Due to their interchangeable feature, the abbreviated form and original form increase the difficulty of NLP applications. In keyword-based information retrieval system, the abbreviated form and original form often return

different results even though they are equivalent in meaning. To solve this problem, we propose a mechanism utilizing sequence similarity and Maximum Entropy Model to construct a linkage between these two forms. The maximum entropy model offers 'feature templates' that incorporate various contextual information. By scanning the training data with feature templates, the feature sets are automatically generated which help us to predict the best candidate and make an abbreviation list without consulting any dictionary.

2. Literature Review

2.1 Chinese word segmentation

Chinese words segmentation is a difficult task because there is no word boundary between words and most of the Chinese words are composed of several characters. Three known methods applied in Chinese word segmentation include dictionary-based approaches, N-Gram approaches and a combination of dictionary and N-gram.

2.2 Longest Common Subsequence

Longest Common Subsequence (LCS) is a dynamic programming algorithm used for finding the longest common subsequence of two given sequences (Taghva & Gilbreth, 1999). A subsequence is obtained by removing zero or more elements from a given sequence. For two sequences X and Y, we say that a sequence Z is a common subsequence of X and Y if Z is a subsequence of both X and Y.

Given two sequences X and Y,

X=<a,c,b,c,e,a,c> , Y=<c,e,b,a,c,a>

Then, <a>, <c>, <c,b>, <c,b,c>, <c,b,a> <c,b,c,a>… are part of the common subsequences of X and Y. Take <c,b,c> as an example, it's a common subsequence of X and Y of length 3. Observe that <c,b,c,a> and <c,e,a,c> are also common subsequences of X and Y (length 4), and there is no common subsequence of length greater than 4 . That is to say, <c,b,c,a> and <c,e,a,c> are the longest common subsequences of string X and Y. We also indicate that LCS <c,b,c,a> can be generated from X by index <2,3,4,6> and from Y by index <1,3,5,6>. Another LCS <c,e,a,c> is generated from X by indices <2,5,6,7> or <4,5,6,7> and from Y by index. <1,2,4,5>

2.3 Expansion of Abbreviations

In related works of expansion of English abbreviations, context information is always taken as an important clue. Toole (Toole, 2000) adopted binary decision tree to distinguish abbreviations from other types of unknown words such as names and misspellings. To identify possible original forms, Toole processed two steps: (1) identifying letters that appear both in the abbreviation and the original form in the same order, and (2) making sure all letters are contained in original form. Next, Word Net was used to find synonyms and definition of possible expansions. Finally, the frequency of these related content words was taken into consideration to choose expansion candidates.

Leah et al. (Leah, Ogilvie, Price, & Tamilio, 2000) proposed four extraction algorithms— finding and evaluating acronyms — including contextual, canonical, canonical/contextual, and simple canonical. The contextual algorithm finds expansions based on context information of abbreviations. The simple canonical algorithm only deals with the abbreviation which appears beside original forms, such as 'Wall Street Journal (WSJ)', 'Wall Street Journal, WSJ', ' (WSJ) Wall Street Journal'. The other two algorithms process the ways in between. Experimental results indicate that the average precision of these four algorithms is higher than 90%. Though the precision of canonical/context algorithm is slightly lower than the others, its recall outperforms them all.

Park and Byrd (Park & Byrd, 2001) proposed a method to find potential abbreviations based on pre-defined rules, and then to discover possible original forms within a pre-defined range. They generated 45 rules after applying their method on 4,500 pairs of abbreviations with their original forms in computer science domain. Afterwards, they verified the rules in the domain of automotive engineering, pharmaceutics, and aviation. The experimental result indicates that the precision and recall are both above 90%.

Akira and Takenobu proposed an automatic dis-abbreviation method (Akira & Takenobu, 2001). They used two aviation-related corpora – one is abbreviation-poor and the other is abbreviation-rich to identify the expansion candidates based on the similarity comparison between the context of a target abbreviation and that of its expansion candidates. The highest score of extracted abbreviation candidates will be automatically expanded to its original form.

2.4 Principle of Maximum Entropy

The concept of maximum entropy was first proposed by Edwin Thompson Jaynes in 1957 (Jaynes, 1957). The purpose of this method is to find the correct distribution which maximizes entropy, or "uncertainty", subject to the constraints. Maximum entropy principle is frequently applied in the domains of Frequency Spectrum Analysis, Images Analysis, Seismology, and Hydrologic Frequency Analysis. In

recent decades, many foreign researchers point out that maximum entropy modeling is an ideal solution for solving many problems in NLP field, such as Part of Speech (Armbrecht Jr et al.) (Lin & Yuan, 2002; Ratnaparkhi, 1996), Recognition of Sentence Boundaries (Reynar & Ratnaparkhi, 1997), Machine Translation (Berger, Pietra, & Pietra, 1996), Co-reference (Kehler, 1997), Named Entity recognition (Borthwick, Sterling, Agichtein, & Grishman, 1998), and Prepositional Phrase Attachment (Ratnaparkhi, JeffReynar, & Roulos, 1994).

2.4.1 Maximum Entropy for NLP

Most of natural language processing involves with estimation of parameter, such as the distribution of observed phenomenon. For instance, if we want to estimate the probability of "class" *a* occurring with "context" *b*, or P(*a*, *b*), here "class" *a* is POS and "context" *b* means the surroundings of the target term. Of course, the definition of *a* and of *b* is different rely on varied problems (Ratnaparkhi, 1997). Improved Iterative Scaling (Venkatraman, Tanriverdi, & stokke) algorithm (Pietra, Pietra, & Lafferty, 1995) is usually adopted to optimize parameters , and to find the best model—Maximum Entropy Model.

According to maximum entropy principle, known information will be taken as constraints. Probability distribution *P* maximizes entropy must conform to these constrains. In other words, *P* must:

1. Maximize entropy.
2. Conforms to the known constraints.

In 1996, IBM Research Center developed a French-to-English machine translation system based on maximum entropy modeling called 'Candide'(Berger et al., 1996). Given a French sentence, the system outputs a corresponding English sentence. Candide take contextual information into consideration and transform it into features functions as follows:

$$f_1(x, y) = \begin{cases} 1, y=en \,\&\, April \text{ follows } in \\ 0, otherwise \end{cases}$$

x stands for contextual information and *y* is some particular French word. In the case the feature function means when *April* follows *in* and *en* is the translation of *in*, *f*(x, y) = 1; otherwise *f*(x, y) = 0. By using maximum entropy model trained by these features functions, it has a great improvement in solving the ambiguity of words.

A state-of-the-art POS tagger (Ratnaparkhi, 1996) was implemented based on a statistical model which is trained from a corpus annotated with part of speech. Maximum entropy model is good at using contextual information, and does not need distributional assumptions on the training data. Besides, maximum entropy model can be applied to solve the problem with sparsely appearance in the corpus that would be difficult to detect in traditional method. The results of POS experiments achieved higher than 95% precision.

Reynar and Ratnaparkhi (Reynar & Ratnaparkhi, 1997) collected 40,000 sentences of Wall Street Journal articles as test data for identifying sentence boundaries via maximum entropy approach. These features are automatically produced by scanning the training data with contextual templates. As a result, no hand-crafted rules or lists are required by the system and it can be easily applied in other languages or text genres. Their results show that the system has good performance even with a small corpus.

Though the performance of using n-gram model or hidden Markov model in Chinese part of speech tagging is well, the obvious drawback is that a lot of useful contextual information could be ignored. Moreover, it results in the failure of recognizing words with low term frequency. To extend previous studies, Lin and Yuan (Lin & Yuan, 2002) applied maximum entropy method to implement Chinese speech tagging. Their results showed that maximum entropy method is flexible in selecting and using contextual information for improving precision of POS tagging.

3. Experiments

Figure 1 illustrates an overview of our experiments. In the process of dis-abbreviation, we extract possible abbreviations from corpus then expand their original forms, then reverse the process order. The first step is to apply word segmentation and POS tagging for obtaining possible abbreviations and originals forms. Next, a target abbreviation and its original form candidates are determined by sequence similarity measurement between strings and predefined rules. Finally, we choose the best candidate determined by maximum entropy method.

3.1 Part-of-Speech Tag and Chinese Word Segmentation

The POS tagging system and the Chinese Corpus developed by the CKIP (Chinese Knowledge and Information Processing) group at Institute of Information Science of Academia Sinica are adopted in our research. Basically, Chinese word segmentation is regarded as character-to-word assignment. In Chinese text, word length is proportioned to its meaning of content. A long phrase composed of several sub-phrases is more meaningful than its sub-phrases. Thus, to extract more meaningful terms, the word length is always taken into consideration. In this study, we conduct

word-segmentation and POS tasks to extract meaningful terms and detect unknown words.

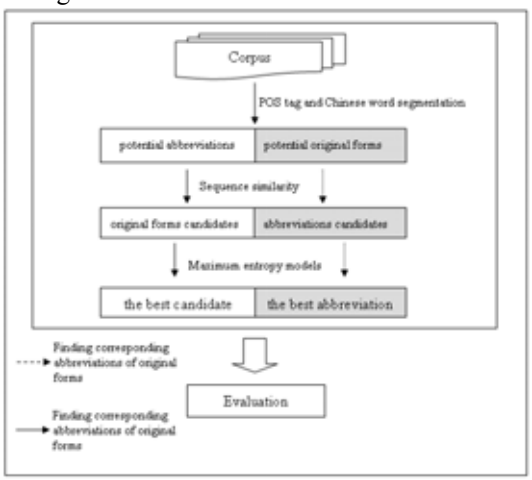

Figure 1 Overview of our experiments

3.2 Possible Abbreviations and Original Forms

Though the abbreviation is only part of its original word, most of the abbreviations are recognizable and representative, such as the word "台鐵局" is the abbreviated form of "台灣鐵路局". With the merit of timesaving and convenient, the abbreviation is gaining popularity and used as a substitute for its original form.

It is discovered that the majority of Chinese abbreviations are composed of two or three characters. (e.g., '稅捐處', '融券'). Only a few cases are four-characters (e.g., '關貿總協'). As for the abbreviations composed of more than five characters are rare. However, we encounter the difficulty of distinguishing abbreviations from original forms when words are four or five characters long. These words could be abbreviations or original forms. For instance, "關貿總協" is the abbreviation of "關稅暨貿易總協定", but "台灣大學" is the original forms of "台大".

The following rules describe the preconditions we consider to extract possible abbreviations and original forms from the corpus:

1. A word, composed of two or three characters, is regarded as a potential abbreviation.
2. A word composed of more than six characters is regarded as a potential original form.
3. A word composed of four or five characters, needs more complex processes.

After conducting the stated preconditions, we examine whether all characters of a word are completely included in any potential original forms by using LCS algorithm. If that is the case, we classify it as the potential abbreviation. If not, the word will be regarded as a potential original form. For instance, "關貿總協" is a potential abbreviation because all of its characters are included in "關稅暨貿易總協定" . On the other hand, characters of "台灣大學" are not included in any of other potential original forms; hence, it is regarded as a potential original form.

3.3 Candidates of the Abbreviation and the Corresponding Original Form

We observe that most of the characters in abbreviations coming from their corresponding original forms. Thus, LCS algorithm is adopted to compute the longest common subsequence between potential abbreviations and original forms. If the length of LCS equals to that of the abbreviation, this pair of words are taken as candidates of abbreviation and the corresponding original form. Take "商檢", the abbreviation as an example, the length of LCS for "商檢" and "商品檢驗" equals to that of "商檢" and "商品檢驗局". Thus, "商品檢驗" and "商品檢驗局" are both candidates of original forms of "商檢". If there does not exist any potential candidate of original forms, the abbreviation will be removed from further analysis.

3.4 The Best Candidate of the Abbreviation and Original Form

Based on the result of section 3.3, the next step is to determine the best candidate of the abbreviation and original form. Regarding the merit in handling contextual information and the flexibility of feature selections, we adopt maximum entropy method to conduct this task. The following section describes details of choosing the best candidate.

3.4.1 Procedure

The first step is to generate candidate features from processing contextual information. The feature selection algorithm is used to conduct this task. Improved Iterative Scaling algorithm is adopted to optimize parameters. The procedures are shown in figure 2.

3.4.2 Feature Template

When contextual information is transformed into "features", we apply the pre-defined "feature template" to generate candidate features. As for the previous discussion in section 3.4, a binary feature function is $f(x,y) \rightarrow \{0,1\}$ in which x is the contextual information within a defined range and y represents candidates of abbreviations and original forms. Feature function is introduced as follows:

$$f(x,y)=\begin{cases}1, \text{if a certain word appears within three words around a given word \& y=\"a given word\"}\\ 0, \text{otherwise}\end{cases}$$

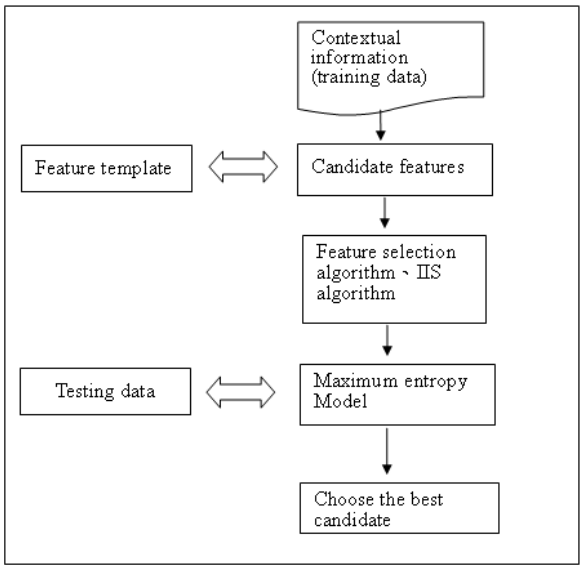

Figure 2 Processes of Choosing the Best Candidate

Feature templates are the pattern of features extracted. Table 1 shows if defined range is six terms:

Table 1 Feature Template

ID	f(x,y)=1 if…
1	y=◇ & □□□▲■□□
2	y=◇ & □□□▲■■■
3	y=◇ & □■■▲□□□

In Table 1, ◇ and ▲represent candidates while ■ means term's location. If now we find an abbreviation '台大', the feature may be generated by template 1 as follows:

$$f(x,y)=\begin{cases}1, \text{if \"教育\" appears following the word \"台大\" \& y=\"台大\"}\\ 0, \text{otherwise}\end{cases}$$

And via template 2:

$$f(x,y)=\begin{cases}1, \text{if \"陳維昭\" appears within three words following the word \"台大\" \& y=\"台大\"}\\ 0, \text{otherwise}\end{cases}$$

Feature selection plays a key role in maximum entropy principle. The collection of all feature candidates is called *F*. Only representative features that qualify the principle of formula 3 are kept in the following process. One algorithm of feature selection is adopted to choose candidate features.

The feature selection algorithm is an incremental procedure that only adds one candidate feature per time into *S* which is a subset of *F*. *S* is called the set of active features and the initial feature set is called *C(S)*. When a feature *f* is adjoined to *S*, we obtain the new set *C(S*∪*f)*.

3.4.3 Extract Contextual Information

Contextual information of candidates is retained. "Search Window (SW)" defines the number of words located ahead or after the target word. Words with short distance to the target word are extracted as features for being the candidate of best abbreviation or original form.

Table 2 The Feature Template for a Part of an Article

Feature template	f(x,y)=1 if…
	y＝中研院 & ■■■■■▲■■■■■

For example, if a potential abbreviation "中研院" is found, and nouns within 5 search window, then "部會", "疫苗", "經濟", "研究所", "研究員", and "謝啟瑞" will be extracted. The result of feature template is shown in Table 3.

Figure 3 Scope of Search

Table 3 An Example of Part of Features

$f(x,y)=\begin{cases}1, \text{if \"部會\" appears within the next three words following the word \"中研院\"\& y=\"中研院\"}\\ 0, \text{otherwise}\end{cases}$
$f(x,y)=\begin{cases}1, \text{if \"疫苗\" appears within the next three words following the word \"中研院\"\& y=\"中研院\"}\\ 0, \text{otherwise}\end{cases}$
$f(x,y)=\begin{cases}1, \text{if \"研究所\" appears within the next three words following the word \"中研院\" \& y=\"中研院\"}\\ 0, \text{otherwise}\end{cases}$

4. Experiment Design and Results

4.1 Resource

Our experiment corpus collected from news website "Yahoo!"【http://tw.news.yahoo.com】dates from 2004/11/23 to 2004/11/30. The number of news documents totals 8,500. The categories include political, financial, Strait (Cross-Strait) and society.

4.2 Results and Analysis

Two tasks are included in our experiment — choosing the best candidate and finding corresponding abbreviations or original forms.

4.2.1 The Best Candidate

The purpose of this experiment is to select the best candidate from potential abbreviations or original forms via executing maximum entropy method. Only those with more than one potential

corresponding abbreviations or original forms after processing are retained for further testing. Here is an example as follows:

<中研院　中央研究院, 中華經濟研究院>

It is obvious from the above example that the potential original forms of "中研院" are "中央研究院" and "中華經濟研究院", therefore, this case will be retained for further analysis.

News documents are divided into two parts shown in Table 4. Eighty percentages of the collection are treated as training data set, and the rest are taken as testing data set.

Table 4 Training and Testing Data Set

	Politics	Finance	Strait	Society
Training data	2,165	1,776	1,439	1,421
Testing data	541	443	360	355
Total numbers of documents	2,706	2,219	1,799	1,776

Once the contextual information of training data sets are turned into features, these features are taken care of and optimized to produce maximal entropy model. The best candidate of words is singled out based on the maximal entropy model from processing of training data. Context information of testing data set will also be transformed.

To examine contextual implication of target word in a sentence, we adopt search window (SW) concept to limit the range of context information. SW means the number of words ahead or after the target word. To evaluate the performance of our estimation, we also compute precision of each experiment. Precision is defined as the percentage of correct results in relation to the number of results retrieved.

4.2.1.1 The Best Candidate of Original Forms

Experiment 1: Nouns and SW=5

We limit the contextual features as noun within five SW of candidates of original forms for the first experiment. For example, the word "中研院" has two candidates of original forms "中央研究院" and "中華經濟研究院". The maximum entropy model is applied to single out the best candidate from "中央研究院" and "中華經濟研究院" based on contextual features of "中研院" within five SW. The precision results are shown in Table 5.

Table 5 Original Forms: Nouns and SW=5

	Politics	Finance	Strait	Society
Precision	60% (6/10)	77% (7/9)	60% (4/6)	60% (3/5)

Experiment 2: Nouns and SW=10

During the previous experiment, we observe that the performance would be raised if more significant nouns included. To justify this observation, we enlarge the scope of SW as ten and the result of precision is improved. Table 6 shows the results.

Table 6 Original Forms: Nouns and SW=10

	Politics	Finance	Strait	Society
Precision	80% (8/10)	89% (8/9)	83% (5/6)	60% (3/5)

Experiment 3: POS and SW=5

In addition to handling nouns, we consider to analyze POS of contextual information within five SW as potential features for the training model. However, the result of precision is not convincing when SW is five.

Table 7 Original Forms: POS and SW=5

	Politics	Finance	Strait	Society
Precision	30% (3/10)	44% (4/9)	50% (3/6)	40% (2/5)

Experiment 4: POS, and SW=10

As the scope of SW enlarges to ten, the result is only fair shown in Table 8.

Table 8 Original Forms: POS and SW=10

	Politics	Finance	Strait	Society
Precision	50% (5/10)	56% (5/9)	60% (4/6)	60% (3/5)

Experiment 5: Nouns, POS, and SW=5

In this experiment, we combine both nouns and POS to train the maximum entropy model and set the size of SW as five. The result is very promising shown as Table 9.

Table 9 Original Forms: Nouns, POS and SW=5

	Politics	Finance	Strait	Society
Precision	70% (7/10)	77% (7/9)	83% (5/6)	80% (4/5)

Experiment 6: Nouns, POS, and SW=10

In this experiment, we combine both nouns and POS as features but the scope of SW expands to ten. The result is also encouraging as shown in Table 10.

Table 10 Original Forms: Nouns, POS and SW=10

	Politics	Finance	Strait	Society
Precision	90% (9/10)	89% (8/9)	83% (5/6)	80%(4/5)

4.2.1.2 The Best Candidate of Abbreviations

In this experiment, the procedure is conducted in reverse order to choose the best candidate of abbreviations.

Experiment 7: Nouns and SW=5

For example, the word, ”中華經濟研究院” has two candidates of abbreviation “中經院” and “中研院”. To train the maximum entropy model, only nouns within five SW surrounding “中經院” and “中研院”are extracted. Then, we input the contextual features of ‘中華經濟研究院’ extracted from testing data to the maximum entropy model to predict which is the best candidate of abbreviation. Table 11 shows that precision of all categories exceeds 70%.

Table 11 Abbreviations: Nouns and SW=5

	Politics	Finance	Strait	Society
Precision	73% (11/15)	77% (10/13)	70% (7/10)	70% (4/6)

Experiment 8: Nouns and SW=10

As we enlarge the scope of SW as ten, the precision of all categories except “society” is higher than that of experiment 7. Table 12 shows the result.

Table 12 Abbreviations: Nouns and SW=10

	Politics	Finance	Strait	Society
Precision	80% (12/15)	84% (11/13)	80% (8/10)	70% (4/6)

Experiment 9: POS and SW=5

This experiment tests POS of contextual information within five SW as potential features for the training model. The result shown in Table 13 is also not convincing as experiment 3 either.

Table 13 Abbreviations: POS and SW=5

	Politics	Finance	Strait	Society
Precision	40% (6/15)	46% (6/13)	50% (5/10)	30% (2/6)

Experiment 10: POS and SW=10

As the scope of SW enlarges to ten, the result shown in Table 14 is only fair as experiment 4.

Table 14 Abbreviations: POS and SW=10

	Politics	Finance	Strait	Society
Precision	53% (8/15)	54% (7/13)	60% (6/10)	50% (3/6)

Experiment 11: POS, Noun, and SW=5

In this experiment, we combine both nouns and POS within five SW to train the maximum entropy model. The result is very promising shown as Table 15. The precision of category “finance” even reaches 92%.

Table 15 Abbreviations: POS, Noun and SW=5

	Politics	Finance	Strait	Society
Precision	86% (13/15)	92% (12/13)	80% (8/10)	83% (5/6)

Experiment 12: POS, Noun, and SW=10

In this experiment, we combine both nouns and other POS within ten SW to train the maximum entropy model. The result shown in Table 16 is also very encouraging as the previous experiment 6.

Table 16 Abbreviations: POS, Noun and SW=10

	Politics	Finance	Strait	Society
Precision	86% (13/15)	92% (12/13)	90% (9/10)	83% (5/6)

4.2.2 Expansion of Abbreviation

In view of the promising performance of experiment 6, we select its features as the test base for processing expansion of abbreviation to its original form.

Experiment 13:

1,000 news documents of finance category are randomly selected to process expansion of abbreviation, however, only 45 out of 162 pairs are correct. The precision is only 28%. To explore the reasons of low precision, we analyze it by observing the data. We discover that some potential abbreviations such as “銀行 (bank) ”, ”公司 (company)” and “委員會(committee)” are not abbreviations themselves. Those words are the tail part of name of general organizations.

To avoid the error indicated above, we confine the abbreviation with consecutive characters included in original form to be ignored. For instance, the abbreviation “銀行” are consecutively included in “中國商業銀行” would be discarded. Consequently, we obtain a significant improvement on precision shown in Table 17.

Table 17 Results of Experiment 13

	Pre-Correction	Correction
Precision	28% (45/162)	56% (38/68)

Experiment 14:

Another kind of error case exists when the difference of word length between a potential abbreviation and its corresponding original forms is only one character. Since the difference of word length between abbreviation and its corresponding original forms is always more than one. Thus the wrong case like the abbreviation "水公司" for original form "水泥公司" and the abbreviation "中美洲" for original form "中南美洲" will not exist. After considering this restriction, the result of precision increases to 61% as Table 18.

Table 18 Result of Experiment 14

Experiment 14	
Precision	61% (38/62)

Experiment 15:

Next, we observe a phenomenon that the abbreviation of five-character words composed of a two-character noun and a three-character noun is always the latter noun. For example, the abbreviation of "中央研究院" composed of the two-character noun "中央" (Nc) and the three-character noun "研究院" (Nc) is "研究院" (Nc). This assumption can be applied to other similar cases, like "證券交易所". Thus, the case like <農會,農業委員會> would be discarded from our experiment. Furthermore, we apply this rule to seven-character and nine-character words, and discover that their abbreviations have at least three-character long.

Table 19 Result of Experiment 15

Experiment 15	
Precision	67% (38/56)

As shown in Table 19, result of precision is slightly increased to 67% after adopting previous rule to remove wrong abbreviations of words with five, seven, and nine characters long.

Experiment 16:

The transliterations of foreign names of persons, organizations, and locations frequently appear in news documents (E.g., "巴斯克, 哈巴羅夫斯克"," 塔爾, 萊溫塔爾"). These words belong to the category "Unknown Word" of CKIP. Abbreviations of these words are rare, and it decreases the performance of our system. Thus, we decide to exclude "Unknown Word" from our experiment. The result indicates that precision is improved to 71% after adding this condition.

Table 20 Result of Experiment 16

Experiment 16	
Precision	71% (36/51)

Experiment 17:

In this experiment, 1,000 news documents of all categories are randomly selected to process the steps of experiments 13-16 for expansion of abbreviation. The restrictions stated above are taken into consideration. The result shows that the precision of all categories is above 60% as Table 21.

Table 21 Result of Experiment 17

	Politics	Finance	Strait	Society
Precision	61% (27/44)	71% (36/51)	63% (26/41)	81% (17/21)

Experiment 18:

When adding more documents to the test base, the precision of all but finance categories is slightly lower than that of the result of experiment 16. It is assumed that the specific words used in finance news are quite stable than those of other categories. Therefore, the variety of words add complexity and uncertainty to this process and decrease the performance of precision and recall.

Table 22 Result of Experiment 18

	Politics	Finance	Strait	Society
No. of Documents	2,706	2,219	1,799	1,776
Precision	65% (57/88)	72% (63/87)	66% (38/58)	75% (33/44)

Experiment 19:

In this experiment, 5,000 documents are randomly selected from all categories and divide into two parts, training data and test data. Eighty percentages of documents are taken as training data, and the rest of documents are regarded as testing data. The result of the precision shown as Table 23 is inferior to those with single category of news documents.

Table 23 Result of Experiment 20

Experiment 20	
Precision	63% (95/150)

4.2.3 Abbreviation of Original Forms

In view of the promising performance of experiment 12, we select its features as the test base for generating abbreviation of original forms.

Experiment 21:

In this experiment, 1,000 news documents of finance category are randomly selected to process the steps of experiments 13-16 for transformation of abbreviations from its original forms. The best result of precision reaches 80% as Table 24.

Table 24 Result of Experiment 21

	Exp 13	Exp 14	Exp 15	Exp 16
Precision	62% (39/62)	68% (39/57)	74% (39/53)	80% (39/49)

The transformation result shows that the precision of the original form corresponding to its abbreviation is about 5%-10% higher than the opposite direction. During the filtering procedure, it is found that the performance of transformation results from the original form to its abbreviation is better than the opposite direction. That brings a convenient way to filter out incorrect correspondences.

Experiment 22:

Table 25 shows the result of experiments with randomly selected 1,000 documents from each category. Repeating the steps described in experiment 13-17, we obtain the precision ranging 67% to 80%.

Table 25 Result of Experiment 22

	Politics	Finance	Strait	Society
Precision	67% (29/43)	80% (39/49)	68% (27/40)	77% (17/22)

Experiment 23:

In this experiment, we add more news documents of all categories to process the steps of experiments 13-16 for generating abbreviations. However, the precision shown as Table 26 is slightly lower than that of experiment 22.

Table 26 Result of Experiment 23

	Politics	Finance	Strait	Society
No. of Documents	2706	2219	1799	1776
Precision	64% (55/86)	78% (63/81)	62% (36/58)	74% (30/41)

Experiment 24:

In this experiment, we randomly select 5,000 documents from news documents to process the steps of experiments 13-16 for generating abbreviations. To retrain the maximum entropy models for choosing the best candidate, we split the 5,000 documents into two parts, including eighty percentages as training data and the rest of documents as testing data. The result is shown in Table 27. The result of the precision shown as Table 2 is inferior to those with single category of news documents.

Table 27 Result of Experiment 24

Experiment 24	
Precision	65% (90/138)

5. Conclusions and Future Improvements

To solve the semantic ambiguity problem in Chinese text processing, we proposed a corpus-based approach by using concepts of sequence similarity and maximum entropy principle to generate abbreviations and the corresponding original forms. Our method only relies on contextual information of target words to train maximum entropy model. No dictionary is needed.

In this research, we perform several experiments with a variety of feature measurements. In viewing of SW, it is clear from the experiments that the performance reaches the best when the scope of SW is ten. However, when we increase the scope of SW above 10, the performance decreases.

Two major tasks are included in our experiments— finding the best candidate of original forms or abbreviations, and recognizing potential abbreviations and original forms in documents. Besides, there are two kinds of procedures including detecting abbreviation of original forms, and executing expansion of abbreviation. Considering feature selection, it is suggested to combine both nouns and POS rather than only process either of them respectively.

In the experiments of using maximum entropy models to choose the best candidate, the result indicates that precision of choosing the best candidate is 80 to 90 percentages. On the other hand, the precision of mapping abbreviations and original forms is 70 to 80 percentages. Performance could be enhanced by implementing the following concerns.

1. Since feature selection is essential for training maximum entropy models, it is suggested that other characteristics of words like locations and frequency could be also adopted in future study.

2. Even though the best results are obtained when extracting features within the ten SW, the value of SW could be changed if the number

of documents is increased.

Our study shows that the performance of precision with documents from single news category, especially the category of finance, is the best. Thus, our method may be suitable for corpus with static types of words. It is worth further exploring that if the performance could be improved if we remove noise information from contextual information.

Our proposed mechanism is helpful for applications of natural language processing such as Chinese tokenizer, search engine, translation system and question answering system. Moreover, it is expected to adopt our efforts in text summarization.

Reference

[1] Akira, T., & Takenobu, T. (2001). Automatic disabbreviation by using context information. *in Proceedings of the sixth natural language processing pacific rim symposium workshop on automatic paraphrasing:theories and applications*, 21-28.

[2] Armbrecht Jr, F. M. R., Chapas, R. B., Chappelow, C. C., Farris, G. F., Friga, P. N., Hartz, C. A., et al. (2001). Knowledge Management in Rd - F. *Research technology management, 44*(4), 1.

[3] Berger, A., Pietra, S. D., & Pietra, V. D. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, 22*(1), 39-71.

[4] Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). *Exploiting diverse knowledge sources via maximum entropy in named entity recognition.* Paper presented at the The Sixth Workshop on Very Large Corpora.

[5] Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review, 106*(4), 620-630.

[6] Kehler, A. (1997). *Probabilistic Coreference in Information Extraction.* Paper presented at the Second Conference on Empirical Methods in Natural Language Processing.

[7] Leah, L., Ogilvie, P., Price, A., & Tamilio, B. (2000). Acrophile:An Automated Acronym Extractor and Server. *In Proceedings of the ACM digital libraries conference*, 205-214.

[8] Lin, H., & Yuan, C. F. (2002). *Chinese Part Of Speech Tagging Based on Maximum Entropy Method.* Paper presented at the First International Conference on Machine learning and Cybernetics,Beijing, Beijing.

[9] Park, Y., & Byrd, R. J. (2001). Hybrid text mining for finding abbreviations and their definitions. *In Proceedings of EMNP2001*.

[10] Pietra, V. D., Pietra, S. D., & Lafferty, J. (1995). Inducing features of random fields: Technical Report CMU-CS95-144,School of Computer Science,Carnegie-Mellon University.

[11] Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. *In conference on Empirical Methods in Natural Language Processing*.

[12] Ratnaparkhi, A. (1997). A Simple Introduction to Maximum Entropy Models for Natural Language processing. *Technical Report 97-08,Institute for Research in Cognitive Science,University of Pennsylvania*.

[13] Ratnaparkhi, A., JeffReynar, & Roulos, S. (1994). A Maximum Entropy Model for Prepositional Phrase Attachment. *In Proceedings of the Human Language Technology Workshop(ARP,1994)*, 250-255.

[14] Reynar, J. C., & Ratnaparkhi, A. (1997). A Maximum Entropy Approach to Identifying Sentence Boundaries. *In Fifth Conference on Applied Natural Language Processing*, 16-19.

[15] Taghva, K., & Gilbreth, J. (1999). Recognizing acronyms and their definitions. *International journal on document analysis and recognition(IJDAR)*, 191-198.

[16] Toole, J. (2000). A hybrid Approach to the Identification and Expansion of Abbreviations. *In Proceedings of RIAO'2000, 1*, 725-736.

[17] Venkatraman, N., Tanriverdi, H., & stokke, P. (1999). Working from Home - Is it Working? *European management journal, 17*(5), 19.