Automatic Extraction of Chinese-English Synonyms Based on a Three-Phase Approach

Yih-Jeng Lin(林義証)
Department of Information Management
Chien- Kuo Technology University
Changhua, 500 Taiwan
yclin@ckit.edu.tw

Fong-Long Huang (黃豐隆)
Department of Computer Science
and Information Engineering
National Nuited University
1 Lien-Da, MiaoLi, Taiwan 360
flhuang@nuu.edu.tw

Abstract

In the paper, we propose a novel method to extract automatically the Chinese-English Synonyms based on a three-phase approach (TPA) without the usage of dictionary. The extraction processes are divided into two categories: synonyms which occurred 1) just once, and 2) occurred more than once in the Chinese corpus. The procedure is based on the three-phase system, in which the people's writing custom and statistical approaches are used to find out the correct synonyms. Our methods also can extract the unknown words in documents. The empirical results demonstrate the precision rates of outside tests reach 89.1%, which proves that the proposed methods of three-phase approach in the paper is feasible and effective.

Keywords: *Chinese-English Synonyms, Unknown Word, and Translation System*.

1. Introduction

There are many synonyms (同義詞) occurring in the papers or documents on Internet. Synonym means the abbreviation or acronym of several words, for example, AIDS stands for "愛滋病 (Acquired Immune Deficiency Syndrome)", and IP stands for 網繚網路協定 (Internet Protocol) or 智慧財 (Intellectual Property). The synonyms, which simplify the readings and writing, can be easily found in newspaper and documents on Internet. The Chinese-English synonyms can be used widely in many applications, such as language translator, Text-To Speech (TTS) system, and so on.

In the paper, we propose the three-phase approach (TPA) to extract automatically the Chinese-English synonyms from Chinese corpus without using any dictionary.

1.1 Previous works of Synonyms Extraction

Previous works related with the extraction of synonyms can be found in [3, 4, 5]. Such synonyms may be some unknown words [1, 2]. Author of [7] proposed a method to extract the synonyms form documents of Internet, in which they only processed the synonyms occurring more than once. The final precision rate was 88%. Shih [5] made change on Shen [4], but they employed the dictionary to process word segmentation(斷詞). Precision rate was 83%. Authors of [3] proposed a method on the database of national dissertation and thesis papers for extracting abstracts, keywords and titles. They reached 94% of precision rate. However, the volume of database is relatively small. Its performance will degrade while their methods are used on other domains or diverse materials, such documents on Internet.

1.2 Data Collection

The extraction processes of synonyms are divided into two categories: synonyms which occurred 1) just once, and 2) occurred more than once (>=1) in the Chinese corpora. Based on our observation, the most synonyms always occur in the international, science and information technology (IT). Therefore, we use the China news electrical papers as our material, in which large documents are divided into several domains, such as science and information technology, finance and so on.

The translation items containing foreign characters can be found in many documents. The sentences with the translation items in paragraph will be extracted from documents. Example (E1) is one of sentences in the collected material:

(E1) 我們認爲惠普(HP)是世界知名印表機公司之一。

We consider HP is one of most famous printer company.

In Example (E1), the synonym is "惠普"-"HP" of Chinese-English pair. Some other items are 佳能（CANON）, 三菱(MITSUBISHI), 超微(AMD)， and so on； in which the English characters with both left and right parentheses follow the Chinese item (or string). Some other synonyms are presented in Table 1; 莫特查(MITUL MOTECHA）is a person name, and 混合信號(MIXED SIGNAL) a jargon; terminology. Note that MITSUBUSHI is a wrong spelling word for "MITSUBISHI"; which occur once in the corpus. All the material downloaded from Internet contains many noises, for example, html tag, advertise-used word, script and CSS code, which should be removed in advance. Note that same material should be avoided collecting more than twice.

Table 1. Some sentences with candidate of Chinese-English synonyms.

與三菱(MITSUBISHI)合作關係生變。
三菱(MITSUBISHI)的發言人指出；
力晶維持八年合作關係的三菱(MITSUBUSHI)
銀負責外匯研究部門的莫特查(MITUL MOTECHA)
持續增加混合信號(MIXED SIGNAL)
與混合信號(MIXED SIGNAL)

The paper is organized as follows. In Section 2, our proposed three-phase approach for the extraction of synonyms with English string occurring more than once will be expressed in detail. Section 3 focuses on the extraction of synonyms with English string occurring only once. The empirical results are presented and analyzed in Section 4. The conclusions and future studies are shown in the final section.

2. Extracting the Synonyms with English Strings Occurring More Than Once

In the section, we focus on expressing the automatic extraction of synonyms with English strings occurring more than once in the corpus. Our method is based on the three-phase approach (TPA) which is composed of three phases to extract the synonyms form documents in our corpus.

2.1 Algorithm to Extract Candidate of Synonyms

Basically, a meaning string (Mandarin word or frequent used string) will occur not only in a document, but also occur in other documents more than once. On the other hand, the meanings string may be used by some people and then appear in different papers. It leads to the occurrence in different documents more than once. Because of absence of white space between two meaning Mandarin strings, we need an algorithm to calculate the frequency of meaning string occurring in corpus. All the extracted string can be regarded as the candidate of synonyms, which should be processed further to decide whether it is a synonym or not. The algorithm is used by phase 2 in our TPA.

For instance, as shown in Example (E2), a meaning string "科技大學" occurs twice, while string "科技" is just part of other longer string "科技大學". Therefore, the frequency *NFw* of string "科技" is zero, calculated by the algorithm shown in Fig. 1.

(E2) 建國科技大學是一所在彰化的科技大學

Chien-Kuo Technology University is a technology university at Changhua.

```
Input: a corpus $C$ and given string $w$
Output: the count of $w$ occurring in $C$
Definitions of symbols:
  $w, v, u, s$ : Chinese strings.
  $L(w)$:Number of Chinese characters of string
        $w$.
  $Fw$ : Number of appearance of $w$ in $C$.
  Last($w,n$): Last $n$ characters of $w$.
  First($w,n$): First $n$ characters of $w$.
Algorithm:
  Step 1. calculate frequency $Fw$ of string
          $w$ in corpus $C$
  Step 2. Collect all Chinese string $v$ in set
          $B$, where $v$ staisfies:
              1. $Fv>1$,
              2. $L(v)=L(w)-1$,
              3. Last($v$, $L(w)$)=$w$.
              Let $TF_B$ be the total frequency of
              each element in $B$.
  Step 3. Collect all Chinese string $v$ in set
          $D$, where $v$ satisfies:
                  1. $Fv>1$,
                  2. $L(v)=L (w)-1$,
                  3. First($v$, $L(w)$)=$w$.
              Let $TF_C$ be the total frequency of
              each element in $D$.
  Step 4. For every $v$ in $B$ and every $u$ in $D$, let
          $L(s)=L(v)+1$ and $s$ satisfies:
                  1. First($s$, $L(v)$)= $v$,
                  2. Last($s$, $L(u)$)= $u$.
          Let $TF$ be the total number of such $s$.
  Step 5. The Net Frequency $NFw$ of $w$ is
defined
          by  $NFw= Fw – (TF_B + TF_C) + TF$.
```

Fig. 1: Algorithm for calculating the frequency *NFw* of string *w* occurring in corpus *C*.

2.2 The Overview of TPA

The extraction of Chinese-English synonyms is based on our proposed method: the three-phase approach. As shown in Fig. 2[1], the three-phase approach contains three phases and each phase will play a function to extract the synonyms. Note that each phase will employee linguistic features or statistical methods to extract the synonyms in documents. We will explain each phase by examples. The processes of TPA are presented as shown in Fig. 2. The sentence with candidate synonyms is read in. If

[1] Fig. 2 is in page 6.

first phase can choose the synonym, it will output and the process finishes. Otherwise, the sentence will be passed into second phase, and the algorithm in Fig. 1 will be used to extract the candidate synonyms. If the phase can choose one synonym from candidate, the result will be output. Otherwise, the third phase will be triggered and extract the synonym.

2.3 The First Phase of TPA

In the first phase of TPA, we observe that some punctuation marks, such as 「 」, 《 》 and 『 』, the string appear within these punctuation marks in front of bracket "(" and its length is equal or less than 2 (<=2) will be chosen as the synonym. All the candidate synonyms are collected and decide furthermore the synonym. As shown in Table 2 and 3, the string "美國電話電報公司" and "超微"are chosen by the process of phase 1 of TPA.

Table 2. Sentences with special punctuation marks

教唆他更改對美國電話電報公司(AT&T)
當「美國電話電報公司」(AT&T)[2]
美國電話電報公司(AT&T)

Table 3. Sentences with less characters (<=2)

一反八月時超微(AMD)
超微(AMD)[3]
逐步把上述下單模式擴增到超微(AMD)

2.4 The Second Phase of TPA

The goal of second phase of TPA is to extract the synonym from candidate's English strings which occur more than once. In the phase, the algorithm of Fig. 1 will be employed to calculate the frequency of Chinese strings. As shown in Table 4, all the strings with (ACER) in our corpus are listed. The frequency of these possible candidate strings can be calculated by our algorithm, which outputs are shown in Table 5, in which three possible strings are collected. The string "宏碁" with max frequency 6 is our final choice. In the example, our TPA extract automatically a Chinese-English synonym "宏碁(ACER)".

Table 4. All the substring related with ACER in our corpus.

宏碁(ACER)
並將替宏碁(ACER)
本土品牌宏碁(ACER)
並將替宏碁(ACER)
如宏碁(ACER)
原屬宏碁(ACER)
因此將取代以往宏碁(ACER)
而宏碁(ACER)
宏碁(ACER)
未來將取代以往宏碁(ACER)

Table 5. The possible synonyms of ACER

宏碁*[4]	6
並將替宏碁	2
將取代以往宏碁	2

2.5 The Third Phase of TPA

The possible synonyms from the candidate string with frequency more than once can be extracted by using the second phase. However, some strings occurring only once don't match such situation and then we can't obtain the appropriate output. Based on our observation, the string will appear again usually in its original document. That is; we will calculate the frequency of possible string in original document based on the algorithm in Fig. 1. For example, as shown in Table 6, only two possible string are related with "AIDS", in which the string "愛滋病" occur 9 times and other string "後天免疫不全症候群" appears 2 times in its original document, respectively. Therefore, string "愛滋病" is chosen as the synonym.

Table 6. The string related with AIDS

顯然與愛滋病(AIDS)
後天免疫不全症候群(AIDS)

3. Extracting Synonyms with English Strings Occurring Once

Because of the less frequent strings, there are many synonyms with English strings occurring only once. As shown in Example (E3), the English string "Nokia" appear once in our corpus. Therefore, the possible synonym "諾基亞-Nokia" is hard to extract. It is hard to extract automatically these synonyms employing the methods in the Section 2. In the past, few previous papers addressed the issues. Our approach uses the statistical method based on the linguistic feature. Like Fig. 2, the procedure contains also three phases; each phase will be expressed in the following section.

(E3) 在台灣的最大客戶則爲諾基亞(Nokia),

The biggest client Taiwan is Nokia

3.1 The First Phase

Some strings within some special punctuation marks, such as「 」,《 》and『 』, cab be extract directly, like explain in Sec. 2.3.

3.2 The Second Phase

In the phase, we use statistical method to

[2] The string "美國電話電報公司" within punctuation marks 「 」 is chosen as the synonym.

[3] The length of string "超微" at left bracket "(" is equal or less than 2 (<=2).

[4] The string "宏碁" with maximum net frequency is chosen as synonym.

find the characters in front of Chinese string of synonyms. As shown in Example (E5), the character “師” in front of “薩都賓”is found. The string “薩都賓” in Example (E5) between stop character “師” and left bracket “（“ will be extracted.

(E5) 研究中心的機 師 薩都賓（DEREK SADUBIN）已抵達台北。

The engineer DEREK SADUBIN of research center arrived at Taipei.

We call it first set of stop characters. There are 26 Mandarin characters are found based on the statistical method as follow:

”的長師家是與爲人裁及在以和如商授司於管員理有者席用廠”

Furthermore, we find second set of 23 stop characters as follow:

“出括國任輯個由片得了從據到行牌種而主業事入年之”

Note that two sets of stop characters are used to extract the Chinese String of synonyms, respectively. In our demonstration, it is helpful for us to extract the synonyms. Results will display in next section.

3.3 The Third Phase

The average number of characters of correctly extracted synonyms is 5.2 Chinese characters in average. Therefore, if the length of Chinese string of a possible synonym is less than 6 (<=6), it will be extracted directly. As shown in Example (E6), the length of string “德意志銀行” is less than 6 (<=6), this string will be extracted. Therefore, synonym”德意志銀行-DEUTSCHEB ANK” is obtained.

(E6) 德意志銀行(DEUTSCHE BANK)下個月即將開張。

DEUTSCHE BANK will open next month.

4. Empirical Results

4.1 Data Set

Our data sets are collected from China news electric papers, from 2002/03~2003/10. Upon the text preprocess, there are 1,655,400 Chinese characters, in which 5,481 types of possible English strings of synonyms with English string occurring more than once (1407 types) or only once (4074 types) are found.

4.2 Results

In the following, we will present the empirical results; which are divided into two categories: possible synonyms with English string occurring more than once (seeing Table 7) and just once (seeing Table 8).

Table 7. Results for English string occurring more than once (>=1)

Phase	Extracted types/total types processed by the phase	Correct number of that type	Precision (%)
1	403 / 1407	399	99
2	1,004 / 1,004	897	89.34
3	53 / 53	13	24.52
Total	1,407	1,309	93.03

As shown in Table 7, there are 403 types of possible 1407 synonyms to be extracted in the first phase. Among them, 309 for 403 types can be extracted correctly. The reminding 1004 (1407-403=1004) types are processed furthermore by phase 2. 897 for 1004 types are correct. Finally, all the reminding 53 types of possible synonyms are not found by phase 2, and then processed by phase 3. The empirical results are shown in Table 7 also. Precision rates of phase 2 and 3 reaches 89.34% and 24.52%, respectively. Final precision rate is 93.03%. Compared with [4] and [5], which only reached 88% of precision rate, our method outperforms.

Table 8. Results for synonyms with English string occurring once (=1)

Phase	Extracted types/total types processed by the phase	Correct number of types	Precision (%)
1	794 / 4,074	787	99.11
2	3000/ 3,280 (4074-794)	2615	87.1 7
3	280/280 (3280-280)	168	60
Total	4,074	3,570	87.63

In our empirical demonstration, we analyze the performance of stop characters of phase 2 in Section 3.2 for the possible synonyms with English occurring only once. As presented in Fig. 3, first set of 26 stop characters are tested. It is apparent that in the first 26 stop characters, the more stop characters the higher the precision rate. However, on the 27th stop character degrade the performance. Therefore, only 26 stop characters are used. The second set of stop character is analyzed in the same way and then 23 characters are chosen, as shown in Section 3.2. The papers of [4] and [5] can’t extract this kind of synonyms with English string occurring only once.

In our observation, some wrong extraction of synonyms is caused by inappropriate sentences, such as absence of right parentheses or misspelling. Note that among the data sets; most possible synonyms are with English string occurring only once (4074/5481=74.3%). How to extract the synonyms with English only once is a more important task. In average, Final precision rate reaches 89.1% based on the three-phase approach.

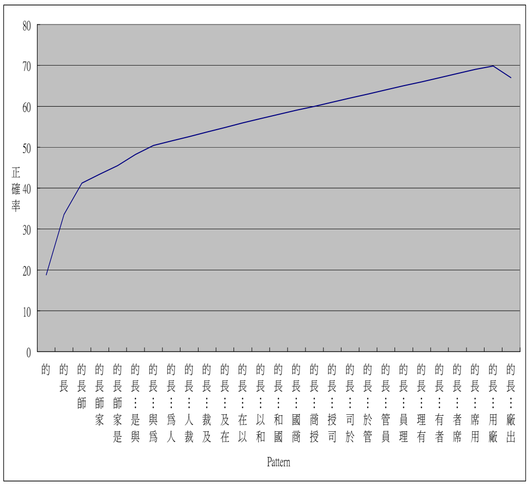

Fig.3. Precision rates for first set of top 26 stop characters.

5. Conclusions and Future Works

In the paper, we propose a three-phase approach (TPA) to extract automatically Chinese-English Synonyms based on without using any dictionary. Most possible synonyms are with English string only once while few precious papers discuss the issues in detail for extracting. According to the empirical results, the proposed methods can extract automatically the synonyms. Final precision rate reaches 89.1%.

The papers of [4] and [5] just extracted the synonyms occurring more than once. Our proposed method outperform for extracting the synonyms with string occurring more once. Furthermore, we can extract the synonyms only occurring once.

In the future, we will expand our method to extract synonyms in multi-lingual documents.

References:

[1] J. S. Chang, ”Automatic Lexicon Acquisition and Precision-Recall Maximization for Untagged Text Corpora,” Ph.D. Thesis, Dept. of Electrical Engineering, National Tsing-hua University,1997

[2] 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.

[3] C. Y. Huang, ” A Study of Theses Clustering Method Based on Synonymous Chinese and English Keyword Sets,” Master thesis, Department of Computer Science and Information Engineering, TungHai University, Taiwan, 2004.

[4] Y. B. Shen, “Thesaurus Extraction From Web, ” Master thesis, Department of Computer Science and Information Engineering, National Chung-Cheng University, Taiwan, 2001.

[5] Y. M. Shih, “Thesaurus Extraction From the World Wide Web, ” Ph.D. thesis, Department of Computer Science and Information Engineering, National Chung-Cheng University, Taiwan, 2002.

[6] T. Y. Wang, ” Q & A in Web Search Engine”， Master thesis, Department of Computer Science and Information Engineering, National Chung-Cheng University, Taiwan, 2003.

[7] M. F. Wang, Michael F., and Ross W., “Using Clustering and Classification Approaches in Interactive,” Information Processing and Management, 37(3), pp.459-484, 2001.

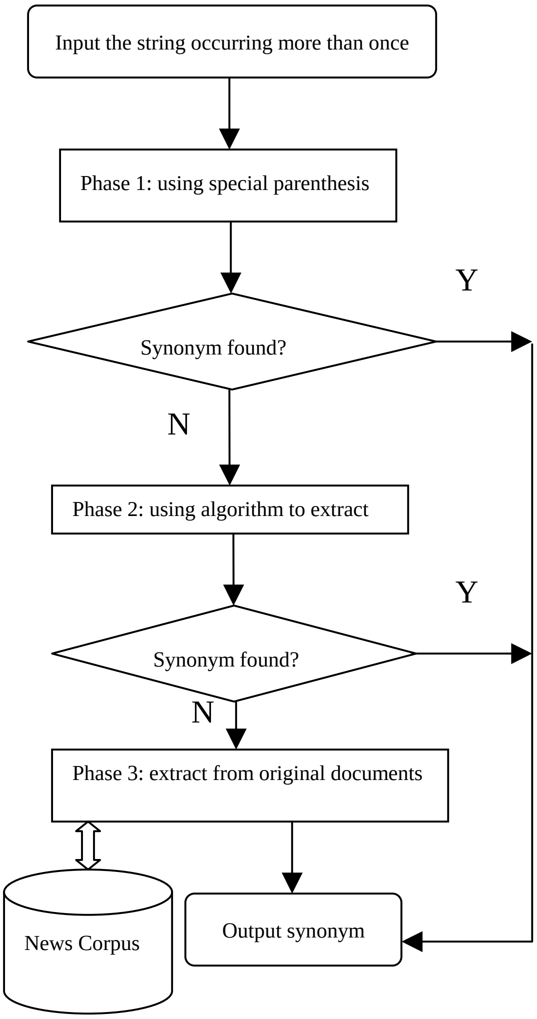
Input the string occurring more than once
Phase 1: using special parenthesis
Synonym found?
Y
N
Phase 2: using algorithm to extract
Synonym found?
Y
N
Phase 3: extract from original documents
News Corpus
Output synonym

Fig. 2: Three-phase approach for extract synonyms occurring more than once.