Word-Transliteration Alignment

Tracy Lin
Dep. of Communication Engineering
National Chiao Tung University,
1001, Ta Hsueh Road,
Hsinchu, 300, Taiwan
tracylin@cm.nctu.edu.tw

Chien-Cheng Wu
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road,
Hsinchu, 300, Taiwan
g904374@oz.nthu.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road,
Hsinchu, 300, Taiwan
jschang@cs.nthu.edu.tw

Abstract

The named-entity phrases in free text represent a formidable challenge to text analysis. Translating a named-entity is important for the task of Cross Language Information Retrieval and Question Answering. However, both tasks are not easy to handle because named-entities found in free text are often not listed in a monolingual or bilingual dictionary. Although it is possible to identify and translate named-entities on the fly without a list of proper names and transliterations, an extensive list certainly will ensure the high accuracy rate of text analysis. We use a list of proper names and transliterations to train a *Machine Transliteration Model*. With the model it is possible to extract proper names and their transliterations in a bilingual corpus with high average precision and recall rates.

1. Introduction

Multilingual named entity identification and (back) transliteration has been increasingly recognized as an important research area for many applications, including machine translation (MT), cross language information retrieval (CLIR), and question answering (QA). These transliterated words are often domain-specific and many of them are not found in existing bilingual dictionaries. Thus, it is difficult to handle transliteration only via simple dictionary lookup. For CLIR, the accuracy of transliteration highly affects the performance of retrieval.

Transliteration of proper names tends to be varied from translator to translator. Consensus on transliteration of celebrated place and person names emerges over a short period of inconsistency and stays unique and unchanged thereafter. But for less known persons and unfamiliar places, the transliterations of names may vary a great deal. That is exacerbated by different systems used for Ramanizing Chinese or Japanese person and place names. For back transliteration task of converting many transliterations back to the unique original name, there is one and only solution. So back transliteration is considered more difficult than transliteration. Knight and Graehl (1998) pioneered the study of machine transliteration and proposed a statistical transliteration model from English to Japanese to experiment on back transliteration of Japanese named entities. Most previous approaches to machine transliteration (Al-Onaizan and Knight, 2002; Chen et al., 1998; Lin and Chen, 2002); English/Japanese (Knight and Graehl, 1998; Lee and Choi, 1997; Oh and Choi, 2002) focused on the tasks of transliteration and back-transliteration. Very little has been touched upon for the issue of aligning and acquiring words and transliterations in a parallel corpus.

The alternative to on-the-fly (back) machine transliteration is simple lookup in an extensive list automatically acquired from parallel corpora. Most instances of (back) transliteration of proper names can often be found in a parallel corpus of substantial size and relevant to the task. For instance, fifty topics of the CLIR task in the NTCIR 3 evaluation conference contain many named entities (NEs) that require (back) transliteration. The CLIR task involves document retrieval from a collection of late 1990s news articles published in Taiwan. Most of those NEs and transliterations can be found in the articles from the Sinorama Corpus of parallel Chinese-English articles dated from 1990 to 2001, including “Bill Clinton,” “Chernobyl,” “Chiayi,” “Han dynasty,” “James Soong,” “Kosovo,” “Mount Ali,” “Nobel Prize,” “Oscar,” “Titanic,” and “Zhu Rong Ji.” Therefore it is important for CLIR research that we align and extract words and transliterations in a parallel corpus.

In this paper, we propose a new machine transliteration method based on a statistical model trained automatically on a bilingual proper name list via unsupervised learning. We also describe how the parameters in the model can be estimated and smoothed for best results. Moreover, we show how the model can be applied to align and extract words and their transliterations in a parallel corpus.

The remainder of the paper is organized as follows: Section 2 lays out the model and describes how to apply the model to align word and transliteration. Section 3 describes how the model is trained on a set of proper names and transliterations. Section 4 describes experiments and evaluation. Section 5 contains discussion and we conclude in Section 6.

2. Machine Transliteration Model

We will first illustrate our approach with examples. A formal treatment of the approach will follow in Section 2.2.

2.1 Examples

Consider the case where one is to convert a word in English into another language, says Chinese, based on its phonemes rather than meaning. For instance, consider transliteration of the word "Stanford," into Chinese. The most common transliteration of "Stanford" is "史丹福." (Ramanization: [shi-dan-fo]). We assume that transliteration is a piecemeal, statistical process, converting one to six letters at a time to a Chinese character. For instance, to transliterate "Stanford," the word is broken into "s," "tan," "for," and "d," which are converted into zero to two Chinese characters independently. Those fragments of the word in question are called transliteration units (TUs). In this case, the TU "s" is converted to the Chinese character "史," "tan" to "丹," "for" to "佛," and "d" to the empty string λ. In other words, we model the transliteration process based on independence of conversion of TUs. Therefore, we have the *transliteration probability* of getting the transliteration "史丹福" given "Stanford," P(史丹佛 | Stanford),

P(史丹佛 | Stanford) = P(史 | s) P(丹 | tan) P(佛 | for) P(λ | d)

There are several ways such a machine transliteration model (MTM) can be applied, including (1) *transliteration* of proper names (2) *back transliteration* to the original proper name (3) *word-transliteration alignment* in a parallel corpus. We formulate those three problems based on the probabilistic function under MTM:

Transliteration problem (TP)

Given a word *w* (usually a proper noun) in a language (L1), produce automatically the transliteration *t* in another language (L2). For instance, the transliterations in (2) are the results of solving the TP for four given words in (1).

(1) Berg, Stanford, Nobel, 清華
(2) 伯格, 史丹佛, 諾貝爾, Tsing Hua

Back transliteration Problem (BTP)

Given a transliteration *t* in a language (L2), produce automatically the original word *w* in (L1) that gives rise to *t*. For instance, the words in (4) are the results of solving the BTP for two given transliterations in (3).

(3) 米開朗基羅, Lin Ku-fang
(4) Michelangelo, 林谷芳

Word Transliteration Alignment Problem (WTAP)

Given a pair of sentence and translation counterpart, align the words and transliterations therein. For instance, given (5a) and (5b), the alignment results are the three word-transliteration pairs in (6), while the two pairs of word and back transliteration in (8) are the results of solving WTAP for (7a) and (7b)

(5a) Paul Berg, professor emeritus of biology at Stanford University and a Nobel laureate, …
(5b) 史丹佛大學生物系的榮譽教授，諾貝爾獎得主伯格[1],

(6) (Stanford, 史丹福), (Nobel, 諾貝爾), (Berg, 伯格)

(7a) PRC premier Zhu Rongji's saber-rattling speech on the eve of the election is also seen as having aroused resentment among Taiwan's electorate, and thus given Chen Shui-bian a last-minute boost.

(7b) 而中共總理朱鎔基選前威脅台灣選民的談話，也被認為是造成選民反感，轉而支持陳水扁的臨門一腳。[2]

(8) (Zhu Rongji, 朱鎔基), (Chen Shui-bian, 陳水扁)

Both transliteration and back transliteration are important for machine translation and cross language information retrieval. For instance, the person and place names are likely not listed in a dictionary, therefore should be mapped to the target language via run-time transliteration. Similarly, a large percentage of

[1] Scientific American, US and Taiwan editions. What Clones? Were claims of the first human embryo premature? Gary Stix and 潘震澤(Trans.) December 24, 2001.

keywords in a cross language query are person and place names. It is important for an information system to produce appropriate counterpart names in the language of documents being searched. Those counterparts can be obtained via direct transliteration based on the machine transliteration and language models (of proper names in the target language).

The memory-based alternative is to find those word-transliteration in the aligned sentences in a parallel corpus (Chuang, You, and Chang 2002). Word-transliteration alignment problem certainly can be dealt with based on lexical statistics (Gale and Church 1992; Melamed 2000). However, lexical statistics is known to be very ineffective for low-frequency words (Dunning 1993). We propose to attack WTAP at the sub-lexical, phoneme level.

2.2 The Model

We propose a new way for modeling transliteration of an English word *w* into Chinese *t* via a Machine Transliteration Model. We assume that transliteration is carried out by decomposing *w* into *k* translation units (TUs), $\omega_1, \omega_2, \ldots, \omega_k$ which are subsequently converted independently into $\tau_1, \tau_2, \ldots, \tau_k$ respectively. Finally, $\tau_1, \tau_2, \ldots, \tau_k$ are put together, forming *t* as output. Therefore, the probability of converting *w* into *t* can be expressed as P(*t* / *w*) = $\max_{k, \omega_1 \ldots \omega_k, \tau_1 \ldots \tau_k} \prod_{i=1,k} P(\tau_i \mid \omega_i)$, where $w = \omega_1 \omega_2 \ldots \omega_k$, $t = \tau_1 \tau_2 \ldots \tau_k$, $|t| \le k \le |t|+|w|$, $\tau_i \ \omega_i \neq \lambda$. See Equation (1) in Figure 1 for more details.

Based on MTM, we can formulate the solution to the Transliteration Problem by optimizing P(*t* / *w*) for the given *w*. On the other hand, we can formulate the solution to the Back Transliteration Problem by optimizing P(*t* / *w*) P(*w*) for the given *t*. See Equations (2) through (4) in Figure 1 for more details.

[2] Sinorama Chinese-English Magazine, A New Leader for the New Century--Chen Elected President, April 2000, p. 13.

The word-transliteration alignment process may be handled by first finding the proper names in English and matching up with the transliteration for each proper name. For instance, consider the following sentences in the Sinorama Corpus:

(9c) 「當你完全了解了太陽、大氣層以及地球的運轉，你仍會錯過了落日的霞輝，」西洋哲學家懷海德說。

(9e) "When you understand all about the sun and all about the atmosphere and all about the rotation of the earth, you may still miss the radiance of the sunset." So wrote English philosopher Alfred North Whitehead.

It is not difficult to build part of speech tagger or named entity recognizer for finding the following proper names (PN):

(10a) Alfred, (10b) North, (10c) Whitehead.

We use Equation (5) in Figure 1 to model the alignment of a word *w* and its transliteration *t* in *s* based on the *alignment probability* P(*s*, *w*) which is the product of transliteration probability $P(\sigma \mid \omega)$ and a trigram match probability, $P(m_i \mid m_{i-2}, m_{i-1})$, where m_i is the type of the *i*-th match in the alignment path. We define three match types based on lengths *a* and *b*, $a = |\tau|$, $b = |\omega|$: match(*a*, *b*) = *H* if *a* = 0, match(*a*, *b*) = *V* if *b* = 0, and match(*a*, *b*) = *D* if *a* > 0 and *b* > 0. The *D*-match represents a non-empty TU ω matching a transliteration character τ, while the *V*-match represents the English letters omitted in the transliteration process.

Machine Transliteration Model: The probability of transliteration *t* of the word *w*

$$\mathrm{P}(t\ /\ w) = \max_{k,\omega_1\ldots\omega_k,\tau_1\ldots\tau_k} \prod_{i=1,k} P(\tau_i \mid \omega_i), \tag{1}$$

where $w = \omega_1\omega_2\ldots\ \omega_k$,

$t = \tau_1\tau_2\ldots\tau_k$,

$|\,t\,| \ \le\ k\ \le\ |\,t\,| + |\,w\,|$,

$|\ \tau_i\ \omega_i\,| \ge 1$.

Transliteration: Produce the phonetic translation equivalent *t* for the given word *w*

$$t = \arg\max_{t} \mathrm{P}(t\ /\ w) \tag{2}$$

Back Transliteration: Produce the original word *w* for the given transliteration *t*

$$\mathrm{P}(w\ /\ t) = \frac{\mathrm{P}(t \mid w)\,\mathrm{P}(w)}{\mathrm{P}(t)} \tag{3}$$

$$w = \arg\max_{t} \frac{\mathrm{P}(t \mid w)\,\mathrm{P}(w)}{\mathrm{P}(t)} = \arg\max_{t} \mathrm{P}(t \mid w)\,\mathrm{P}(w) \tag{4}$$

Word-Transliteration Alignment: Align a word *w* with its transliteration *t* in a sentence *s*

$$\mathrm{P}(s,\, w) = \max_{k,\omega_1\ldots\omega_k,\sigma_1\ldots\sigma_k} \prod_{i=1,k} \mathrm{P}(\sigma_i\ /\ \omega_i)\ \mathrm{P}(m_i \mid m_{i-2},\, m_{i-1}), \tag{5}$$

where $w = \omega_1\omega_2\ldots\omega_\kappa$,

$s = \sigma_1\sigma_2\ldots\sigma_\kappa$, (both ω_i and σ_i can be empty)

$|\,s\,|\ \le\ k\ \le\ |\,w\,| + |\,s\,|$, $|\omega_i\ \sigma_i| \ge 1$,

m_i is the type of the $(\omega_i\ ,\ \sigma_i)$ match, $m_i = \mathrm{match}\ (|\omega_i\,|, |\ \sigma_i\,|\ )$,

match(a, b) = H, if $b = 0$,

match(a, b) = V, if $a = 0$,

match(a, b) = D, if $a > 0$ and $b > 0$,

$\mathrm{P}(m_i \mid m_{i-2},\, m_{i-1})$ is trigram Markov model probabiltiy of match types.

$$\alpha(i, j\,) = \mathrm{P}(s_{1:i-1},\, w_{1:j-1}). \tag{6}$$

$$\alpha(1,\, 1) = 1,\ \mu(1,\, 1) = (H,\, H). \tag{7}$$

$$\alpha(i, j\,) = \max_{a=0,1,\ b=0,6} \alpha(i\text{-}a,\, j\text{-}b)\ \mathrm{P}(s_{j-a:j-1} \mid w_{i-b:i-1})\ \mathrm{P}(\ \mathrm{match}(a,\, b) \mid \mu(i\text{-}a,\, j\text{-}b)\ ). \tag{8}$$

$$\mu(i, j) = (m,\ \mathrm{match}(a^*,\, b^*)),\ \text{where}\ \mu(i\text{-}a^*,\, j\text{-}b^*) = (x,\, m), \tag{9}$$

where $(a^*, b^*) = \arg\max_{a=0,1,\ b=0,6} \alpha(i\text{-}a,\, j\text{-}b)\ \mathrm{P}(s_{j-a:j-1} \mid w_{i-b:i-1})\ \mathrm{P}(\ \mathrm{match}(a,\, b) \mid \mu(i\text{-}a,\, j\text{-}b)\ )$.

Figure 1. The equations for finding the Viterbi path of matching a proper name and its translation in a sentence

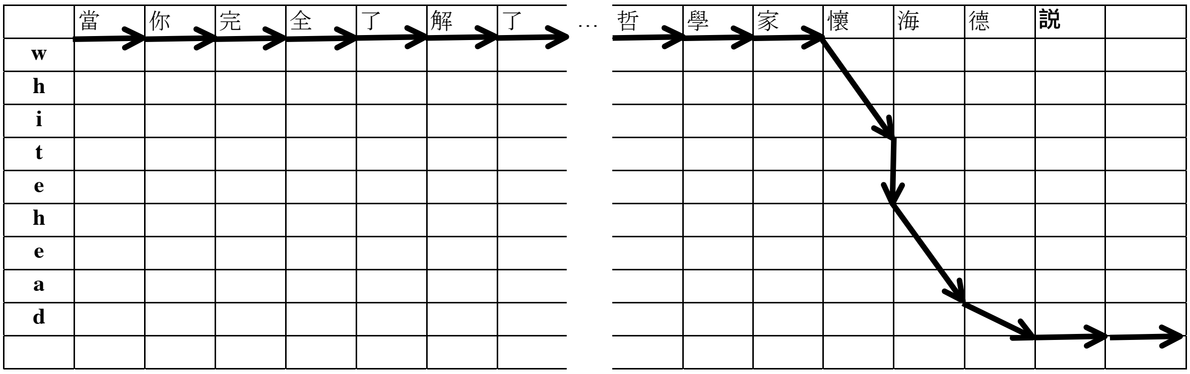


Figure 2. The Viterbi alignment path for Example (9c) and the proper name “Whitehead” (10c) in the sentence (9e), consisting of one *V*-match (te-λ), three *D*-matches (whi–懷, hea–海, d–德), and many *H*-matches.

To compute the alignment probability efficiently, we need to define and calculate the *forward probability* $\alpha(i, j)$ of $P(s, w)$ via dynamic programming (Manning and Schutze 1999), $\alpha(i, j)$ denotes the probability of aligning the first i Chinese characters of s and the first j English letters of w. For the match type trigram in Equation (5) and (8), we need also compute $\mu(i, j)$, the types of the last two matches in the Viterbi alignment path. See Equations (5) through (9) in Figure 1 for more details.

For instance, given w = "Whitehead" and s = "「當你完全了解了太陽、大氣層以及地球的運轉, 你仍會錯過了落日的霞輝，」西洋哲學家懷海德**説**。," the best Viterbi path indicates a decomposition of word "Whitehead" into four TUs, "whi," "te," "hea," and "d" matching "懷," λ, "海," "德" respectively. By extracting the sequence of *D*- and *V*-matches, we generate the result of word-transliteration alignment. For instance, we will have (懷海德, Whitehead) as the output. See Figure 2 for more details.

3. Estimation of Model Parameters

In the training phase, we estimate the transliteration probability function $P(\tau | \omega)$, for any given TU ω and transliteration character τ, based on a given list of word-transliterations. Based on the Expectation Maximization (EM) algorithm (Dempster et al., 1977) with Viterbi decoding (Forney, 1973), the iterative parameter estimation procedure on a training data of word-transliteration list, (E_k, C_k), $k = 1$ to n is described as follows:

Initialization Step:

Initially, we have a simple model $P_0(\tau | \omega)$

$$\begin{aligned} P_0(\tau | \omega) &= \mathrm{sim}(R(\tau) \mid \omega) \\ &= \mathrm{dice}(t_1 t_2 \ldots t_a, w_1 w_2 \ldots w_b) \qquad (8) \\ &= \frac{2c}{a+b} \end{aligned}$$

where $R(\tau)$ = Romanization of Chinese character τ

$R(\tau) = t_1 t_2 \ldots t_a$

$\omega = w_1 w_2 \ldots w_b$

c = # of common letters between $R(\tau)$ and ω

For instance, given *w* = '*Nayyar*' and *t* = '納雅,' we have and R(τ_1) = 'na' and R(τ_2) = 'ya' under Yanyu Pinyin Romanization System. Therefore, breaking up *w* into two TUs, ω_1 = 'nay' ω_2 = 'yar' is most probable, since that maximizes $P_0(\tau_1 \mid \omega_1) \times P_0(\tau_2 \mid \omega_2)$

$P_0(\tau_1 \mid \omega_1)$= sim(na | *nay*) = 2 × 2 / (2+3) = 0.8
$P_0(\tau_2 \mid \omega_2)$= sim(ya | *yar*) = 2 × 2 / (2+3) = 0.8

Expectation Step:

In the Expectation Step, we find the best way to describe how a word get transliterated via decomposition into TUs which amounts to finding the best Viterbi path aligning TUs in E_k and characters in C_k for all pairs (E_k, C_k), *k* = 1 to *n*, in the training set. This can be done using Equations (5) through (9). In the training phase, we have slightly different situation of *s* = *t*.

Table 1. The results of using $P_0(\tau|\omega)$ to align TUs and transliteration characters

w	*s=t*	ω-τ match on Viterbi path
Spagna	斯帕尼亞	s-斯 pag-帕 n-尼 a-亞
Kohn	孔恩	koh-孔 n-恩
Nayyar	納雅	nay-納 yar-雅
Alivisatos	阿利維撒托斯	a-阿 li-利 vi-維 sa-撒 to-托 s-斯
Rivard	里瓦德	ri-里 var-瓦 d-德
Hall	霍爾	ha-霍 ll-爾
Kalam	卡藍	ka-卡 lam藍
Salam	薩萊姆	sa-薩 la-萊 m姆
Adam	亞當	a-亞 dam當
Gamoran	蓋莫藍	ga-蓋 mo-莫 ran-藍
Heller	赫勒	hel-赫 ler-勒
Adelaide	阿得雷德	a-阿 de-得 lai-雷 de-德
Nusser	努瑟	nu-努 sser-瑟
Nechayev	納卡耶夫	ne-納 cha-卡 ye-耶 v-夫
Htler	希特勒	hi-希 t-特 ler-勒
Hunt	杭特	hun-杭 t-特
Germain	杰曼	ger-杰 main-曼
Massoud	馬蘇德	ma-馬 ssou-蘇 d-德
Malong	瑪隆	ma-瑪 long-隆
Gore	高爾	go-高 re-爾
Teich	泰許	tei-泰 ch-許
Laxson	拉克森	la-拉 x-克 son-森

The Viterbi path can be found via a dynamic programming process of calculating the forward probability function α(*i*, *j*) of the transliteration alignment probability P(E_k, C_k) for $0 < i < | C_k |$ and $0 < j < | E_k |$. After calculating P(C_k, E_k) via dynamic programming, we also obtain the TU matches (τ, ω) on the

Viterbi path. After all pairs are processed and TUs and translation characters are found, we then re-estimate the transliteration probability P(τ | ω) in the Maximization Step

Maximization Step:
Based on all the TU alignment pairs obtained in the Expectation Step, we update the maximum likelihood estimates (MLE) of model parameters using Equation (9).

$$\mathrm{P}_{MLE}(\tau \mid \omega) = \frac{\sum_{i=1}^{n} \sum_{\tau \text{ matches } \omega \text{ in } (E_i, C_i)} \mathrm{count}(\tau, \omega)}{\sum_{i=1}^{n} \sum_{\tau' \text{ matches } \omega \text{ in } (E_i, C_i)} \mathrm{count}(\omega)} \qquad (9)$$

The Viterbi EM algorithm iterates between the Expectation Step and Maximization Step, until a stopping criterion is reached or after a predefined number of iterations. Re-estimation of P(τ | ω) leads to convergence under the Viterbi EM algorithm.

3.1 Parameter Smoothing

The maximum likelihood estimate is generally *not* suitable for statistical inference of parameters in the proposed machine transliteration model due to data sparseness (even if we use a longer list of names for training, the problem still exists). MLE is not capturing the fact that there are other transliteration possibilities that we may have not encountered. For instance, consider the task of aligning the word “Michelangelo” and the transliteration “米開朗基羅” in Example (11):

(11) (Michelangelo, 米開朗基羅)

It turns out in the model trained on some word-transliteration data provides the MLE parameters in the MTM in Table 2. Understandably, the MLE-based model assigns 0 probability to a lot of cases not seen in the training data and that could lead to problems in word-transliteration alignment. For instance, relevant parameters for Example (11) such as P(開 | che) and P(朗 | lan) are given 0 probability. Good Turing estimation is one of the most commonly used approaches to deal with the problems caused by data sparseness and zero probability. However, GTE assigns identical probabilistic values to all unseen events, which might lead to problem in our case.

Table 2. $\mathrm{P}_{MLE}(t \mid n)$ value relevant to Example (11)

English TU ω	**Transliteration τ**	$\mathrm{P}_{MLE}(\tau \mid \omega)$
mi	米	**0.00394**
mi	密	0.00360
mi	明	0.00034
mi	麥	0.00034
mi	邁	0.00017
che	傑	0.00034
che	切	0.00017
che	其	0.00017
che	奇	0.00017
che	契	0.00017
che	科	0.00017
che	**開**	**0**
lan	蘭	0.00394
lan	藍	0.00051
lan	倫	0.00017
lan	**朗**	**0**
ge	格	0.00102
ge	奇	0.00085
ge	吉	0.00068
ge	**基**	**0.00017**
ge	蓋	0.00017
lo	洛	0.00342
lo	**羅**	**0.00171**
lo	拉	0.00017

We observed that although there is great variation in Chinese transliteration characters for any given English word, the initial, mostly consonants, tend to be consistent. See Table 3 for more details. Based on that observation, we use the linear interpolation of the Good-Turing estimation of TU-to-TU and the class-based initial-to-initial function to approximate the parameters in MTM. Therefore, we have

$$\mathrm{P}_{li}(c \mid e) = 0.5\,\mathrm{P}_{GT}(c \mid e) + 0.5\,\mathrm{P}_{MLE}(\mathrm{init}(c) \mid \mathrm{init}(e))$$

4 Experiments and evaluation

We have carried out rigorous evaluation on an implementation of the method proposed in this paper. Close examination of the experimental results reveal that the machine transliteration is general effective in aligning and extracting proper names and their transliterations from a parallel corpus.

The parameters of the transliteration model were trained on some 1,700 proper names and transliterations from Scientific American Magazine. We place 10 *H*-matches before and after the Viterbi alignment path to simulate the word-transliteration situation and trained the trigram match type probability. Table 4 shows the estimates of the trigram model.

Table 3. The initial to initial correpsondence of ω amd R(τ)

ω	**τ**	**R(τ)**	**Init(ω)**	**Init(R(τ))**
mi	米	mi	m	m
mi	密	mi	m	m
mi	明	min	m	m
mi	麥	mai	m	m
mi	邁	mai	m	m
che	傑	jei	ch	j
che	切	chei	ch	ch
che	其	chi	ch	ch
che	奇	chi	ch	ch
che	契	chi	ch	ch
che	科	ke	ch	k
che	開	kai	ch	k
lan	蘭	lan	l	l
lan	藍	lan	l	l
lan	倫	lun	l	l
lan	朗	lang	l	l
ge	格	ge	g	g
ge	奇	chi	g	ch
ge	吉	ji	g	j
ge	基	ji	g	j
ge	蓋	gai	g	g
lo	洛	lo	l	l
lo	羅	Lo	l	l
lo	拉	La	l	l

Table 4. The stastical estimates of trigram match types

Match Type Trigram $m_1\ m_2\ m_3$	Count	$P(\ m_3 \mid m_1\ m_2\ )$
DDD	1886	0.51
DDH	1627	0.44
DDV	174	0.05
DHD	0	0.00
DHH	1702	1.00
DHV	0	0.00
DVD	115	0.48
DVH	113	0.47
DVV	12	0.05
HDD	1742	0.96
HDH	7	0.01
HDV	58	0.03
HHD	1807	0.06
HHH	29152	0.94
HHV	15	0.00
HVD	15	1.00
HVH	0	0.00

The model was then tested on three sets of test data:

(1) 200 bilingual examples in Longman Dictionary of Comtemporary Dictionary, English-Chinese Edition.
(2) 200 aligned sentences from Scientific American, US and Taiwan Editions.
(3) 200 aligned sentences from the Sinorama Corpus.

Table 5 shows that on the average the precision rate of exact match is between 75-90%, while the precision rate for character level partial match is from 90-95%. The average recall rates are about the same as the precision rates.

Table 5. The experimental results of word-transliteration alignement

Test Data	# of words (# of characters)	# of matches (# of characters)	Word precision (Characters)
LODCE	200	179	89.5%
	(496)	(470)	(94.8%)
Sinorama	200	151	75.5%
	(512)	(457)	(89.3%)
Sci. Am.	200	180	90.0%
	(602)	(580)	(96.3%)

5. Discussion

The success of the proposed method for the most part has to do with the capability to balance the conflicting needs of capturing lexical preference of transliteration and smoothing to cope with data sparseness and generality. Although we experimented with a model trained on English to Chinese transliteration, the model seemed to perform reasonably well even with situations in the opposite direction, Chinese to English transliteration. This indicates that the model with the parameter estimation method is very general in terms of dealing with unseen events and bi-directionality.

We have restricted our discussion and experiments to transliteration of proper names. While it is commonplace for Japanese to have transliteration of common nouns, transliteration of Chinese common nouns into English is rare. It seems that is so only when the term is culture-specific and there is no counterparts in the West. For instance, most instances “旗袍” and “瘦金體” found in the Sinorama corpus are mapped into lower case transliterations as shown in Example (11) and (12):

(11a) 中國國服——旗袍真的沒落了嗎？
(11b) Are ch'i-p'aos--the national dress of China--really out of fashion?

(12a) 一幅瘦金體書法複製品
(12b) a scroll of shou chin ti calligraphy

Without capitalized transliterations, it remains to be seen how word-transliteration alignment related to common nouns should be handled.

6. Conclusion

In this paper, we propose a new statistical machine transliteration model and describe how to apply the model to extract words and transliterations in a parallel corpus. The model was first trained on a modest list of names and transliteration. The training resulted in a set of 'syllabus' to character transliteration probabilities, which are subsequently used to extract proper names and transliterations in a parallel corpus. These named entities are crucial for the development of named entity identification module in CLIR and QA.

We carried out experiments on an implementation of the word-transliteration alignment algorithms and tested on three sets of test data. The evaluation showed that very high precision rates were achieved.

A number of interesting future directions present themselves. First, it would be interesting to see how effectively we can port and apply the method to other language pairs such as English-Japanese and English-Korean. We are also investigating the advantages of incorporate a machine transliteration module in sentence and word alignment of parallel corpora.

Acknowledgement

We acknowledge the support for this study through grants from National Science Council and Ministry of Education, Taiwan (NSC 90-2411-H-007-033-MC and MOE EX-91-E-FA06-4-4).

References

Al-Onaizan, Y. and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 400-408.

Chen, H.H., S-J Huang, Y-W Ding, and S-C Tsai. 1998. Proper name translation in cross-language information retrieval. In *Proceedings of 17th COLING and 36th ACL*, pages 232-236.

Chuang, T., G.N. You, J.S. Chang (2002) Adaptive Bilingual Sentence Alignment, *Lecture Notes in Artificial Intelligence 2499*, 21-30.

Cibelli, J.B. R.P. Lanza, M.D. West, and C. Ezzell. 2002. What Clones? SCIENTIFIC AMERICAN, Inc., New York, January. http://www.sciam.com.

Dagan, I., Church, K. W., and Gale, W. A. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1-8, Columbus Ohio.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1-38.

Forney, G.D. 1973. The Viterbi algorithm. *Proceedings of IEEE*, 61:268-278, March.

Knight, K. and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599-612.

Lee, J.S. and K-S Choi. 1997. A statistical method to generate various foreign word transliterations in multilingual information retrieval system. In *Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages (IRAL'97)*, pages 123-128, Tsukuba, Japan.

Lin, W-H Lin and H-H Chen. 2002. Backward transliteration by learning phonetic similarity. In *CoNLL-2002*, *Sixth Conference on Natural Language Learning*, Taipei, Taiwan.

Manning, Ch. and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*, MIT Press; 1st edition.

Oh, J-H and K-S Choi. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan.

Proctor, P. 1988. *Longman English-Chinese Dictionary of Contemporary English*, Longman Group (Far East) Ltd., Hong Kong.

Sinorama. 2002. *Sinorama Magazine*. http://www.greatman.com.tw/sinorama.htm.

Stalls, B.G. and K. Knight. 1998. Translating names and technical terms in Arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*.

Tsujii, K. 2002. Automatic extraction of translational Japanese-KATAKANA and English word pairs from bilingual corpora. *International Journal of Computer Processing of Oriental Languages*, 15(3):261-279.

應用混淆音矩陣之中英文音譯詞組自動抽取

郭金喜[1,2] 楊英魁[2]

[1]中華電信研究所 [2]國立台灣科技大學電機系

jskuo@cht.com.tw ykyang@mouse.ee.ntust.edu.tw

摘要

機器音譯(Machine Transliteration)是機器翻譯中重要的一環，因爲許多文章中常有人名、地名及組織名等專有名詞夾雜其中，雖然經由查閱預先整理之詞典可以解決部分的問題，但是這些專有名詞數量隨時間不斷的增加及成長，而辭典的整理既費時又費力，透過音譯詞組自動抽取(Transliterated-Term Pair Extraction)，可動態補充辭典內容之不足。有足夠的中英文音譯詞組做爲訓練語料之後，則可建立一中英文音節對應(Syllable Mapping)系統，應用於中英文詞組音譯，但問題是該如何快速獲取足夠的中英文音譯詞。本文提出一方法，自網頁中抽取出大量的中英文音譯詞組，利用中文語音辨認系統在辨認過程所產生的混淆矩陣(Confusion Matrix)來克服發音變異(Pronunciation Variation)。從實驗結果發現本文所提出的方法可達到 32.26%的檢出率(Recall)及 95.23%的準確率(Precision)，足以證明所用方法確實可有效的應用於音譯詞組自動抽取。

1. 簡介

當國際交流日益頻繁，各國間的資訊傳遞也更加迅速，許多的媒體必須在短時間將所收到的外國資訊儘可能完善的翻譯成本國文字，以滿足讀者的需求，在現今媒體開放、競爭的台灣這種步調更加快速。這些外國資訊常包含有許多的專有名詞(Proper Noun)如人名、地名及組織名等夾雜其中，同一名詞出現在不同文章中但由同一人員翻譯可能會出現不同的譯名，同一名詞由不同人員翻譯也可能會出現不同的名稱。這些問題主要是因爲所接收的外國資訊涵蓋非常廣泛，發生的地點及所使用的語言更是廣佈於全世界，實在是非單一個人可以準確的音譯出由不同語言所發聲的的專有名詞。專有名詞的音譯並不在本文的探討範圍，但音

譯詞組的自動抽取卻是建構機器音譯系統不可或缺的一步。

機器音譯常用來處理人名、地名等，其作法乃是將這些專有名詞經由發音方式自一語言轉換至另一語言。它是機器翻譯中重要的一環，因為在許多文章中常有人名、地名及組織名等專有名詞夾雜其中，雖然經由查閱預先整理之詞典可以解決部分的問題，但是這些專有名詞數量隨時間不斷的增加及成長，而辭典的整理既費時又費力，透過音譯詞組自動抽取可動態補充辭典內容之不足。

想要自動抽取音譯詞組，必須要能自足夠大的語料庫中抽取出多樣且量多的音譯詞組，一般測試用語料庫大多無法滿足這樣的需求。網際網路是現今世界上最大的分散式資料庫，其所包含的資料雖然缺乏有系統的整理 (Systematically Organized) ，但卻包羅萬象而且源源不絕不斷有新的內容產生，這樣具有動態特性的資料是許多研究不可或缺的素材。本文的目的是要自這些網頁資料中抽取出許多可能的中英文音譯詞組，做為未來發展機器音譯的基礎。

英文是目前國際上最通用的語言之一，許多資訊是透過英文翻譯或音譯至其他語言去，有許多的名詞先被引進至英文，其他語言的使用者，透過再從英文引進這些名詞。因此造成許多語言自英文引進的外來語，其原來的字源(Word Origin)[Llitjos2001] 並非來自英文，因此若不了解外來語的字源，常會有發音不一致的其行產生。例如義大利地名 Firenze 及其英文音譯 Florence [Lin2000]，究竟應該採用哪一種讀法，實在很難決定。即使對常用的英文字如 Mary/meɪri/ 、marry/mæri/及 merry/mɛri/ ，有些人把這三者均唸成不同或某兩者相同，但大多數的美國卻把這三者均唸成/mɛri/[Jurafsky2000]。這表示有發音變異的問題存在。而機器音譯則用來將人名、地名等專有名詞經由發音方式自一語言轉換至另一語言，所以為了克服不同人的發音變異問題，必須抽出足夠的音譯詞組，進而建構不同語言間的音節轉換關係。

鄰近的日本與韓國也極力引進及吸收國外資訊，日本文字[NIHOGO90] 中有片假名用於表達外國人的國名、地名、人名(中國、韓國人名 除外)、外來語及專門術語等等，因此英文及日文中的音譯詞組抽取[Brill2001] 可以較清楚的區分何者為外來語。韓文也引進大量的外

來語 ，因此有許多人致力於英文及韓文間之音譯研究[Jeong97][Lee98][Jung2000][Kang2000][Oh2002]。韓文中雖然沒有像日文中特別以片假名來表示外來語，但韓文中的外來語在其字尾多包含有 “Josa” 及 “Eomi” 等符號。這些特性是中文中所沒有的，也增加了中英文音譯詞組自動抽取的困難度。

有關中文方面的音譯詞組自動抽取研究，曾有研究[Xiao2002]利用某些特定的外國人名所出現的音譯中文字如『夫』、『斯』、『基』等找尋其他的中文音譯詞，這些字也許常出現於外國人名的中文音譯中，特別是斯拉夫民族的人名，但是中文人名也有可能會用到這些字，如『李斯』及『郭李建夫 』等。[Lee2003] 則使用統計音譯模型於中英文雙語語料庫(Parallel Corpora)音譯詞組抽取，由於雙語語料庫的資料量相對小於散佈於網際網路上的網頁資料，故能夠取得的中英文音譯詞組數量相對較少。

根據機器音譯模組化學習法(Modular Learning Approach)[Knight98]，在一堆不同語言的音譯詞組候選詞中，最後的音譯詞 $\hat{C}$ 可由(1)決定

$$\hat{C} \equiv \underset{C_w}{argmax}\, p(C_w|E_w) = \underset{C_w}{argmax}\, p(E_s|E_w)\, p(C_s|E_s)\, p(C_w|C_s) \qquad (1)$$

，其中 C_w 及 E_w 分別為目標語言(Target Language)及來源語言(Source Language)文字串， C_s 及 E_s 是目標語言及來源語言文字串所對應之音素(Phoneme)串， $p(C_w|E_w)$ 為一條件機率函數 ，下標 C_w 是在所有音譯詞中使得 $p(C_w|E_w)$ 最大的音譯詞。在本文中來源語言是指英文，目標語言則是指中文。公式(1)意思為要從一個英文字串所對應的中文候選詞中挑出一個最有可能的中文字串，乃是找出一個中文字串使得英文字轉音的機率、英文對中文音轉音的機率以及中文的音轉字的機率三者連乘最大。本文的目的在於音譯詞組抽取，故著重於音素相似程度的評估。由於中英文隸屬於不同語系，且並無類似韓文EKSCR(English-to-Korea Standard Conversion Rules)[Oh2002] 的規則可供遵循，而且一個英文字(Word)可有多個音節，但一個中文字則只有一個音節，其對應關係應不易決定。

基於上述理由，本文提出基於語音辨認的混淆音矩陣，來解決中英文音譯詞自動抽取過程中必須克服的發音變異以及中英文音節對應不易問題。音譯詞組自動抽取的目的在於蒐集

足夠的中英文音譯詞組，以處理中英文音譯的音節轉換問題，這樣經由大量統計而產生的中英文音節轉換系統，納入了許多中英文音節的各種對應關係，對於進一步的中英文音譯研究有很大的裨益。

本文內容第二節將討論如何自動抽取中英文音譯詞組，第三節爲實驗結果與討論，最後是結論。

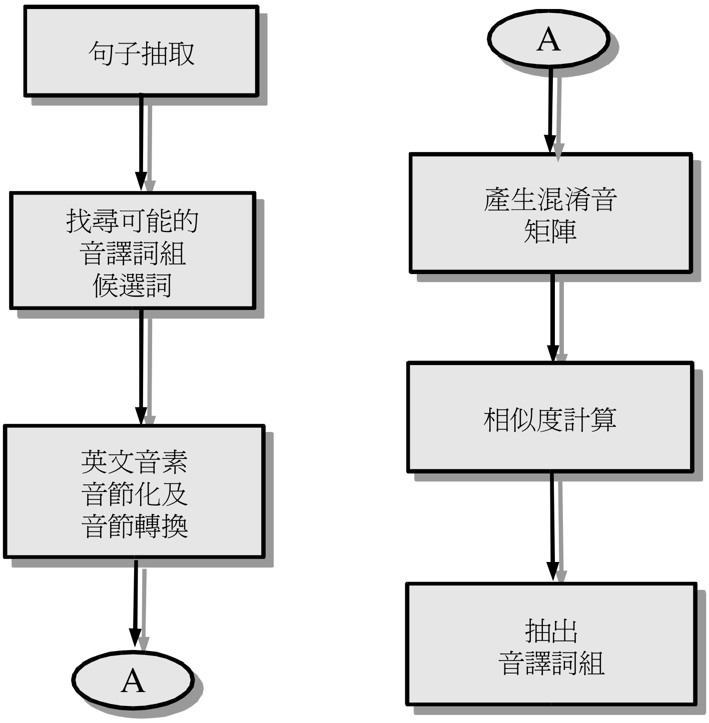


圖 1 中英文音譯詞組自動抽取流程圖

2. 使用方法

網際網路是現今世界上最大的分散式資料庫，每年以極快速的速度成長，其內容五花八門而且不斷有內容更新及產生。本文的目的是要自這個最大的資料庫中抽取出許多可能的中英文音譯詞組，處理中英文音譯的音節轉換問題，做爲進一步發展機器音譯的基礎。這個準結構化的(Semi-Structured)資料庫的確包羅萬象、應有盡有，但如何能夠在其中有效率且快速的獲取所需要的資訊，訂出適當的搜尋或比對範圍，減少不必要的計算，過濾不必要的雜訊，進而抽取出大量可能的中英文音譯詞組，是一項艱難但必須克服的問題。

音譯詞組係由不同語言的文字串所對應而成，如果要在非雙語對應(Non-Parallel)的網頁中抽取出音譯詞組，則表示網頁中包含有混合兩種語言以上的文字資料。因此重要的是要找到分屬目標語言與來源語言的對應文字串，如果無法找到非常明確的對應文字串，則必須縮小範圍或以其他有效的特徵過濾不必要的文字雜訊。

本文在中英文音譯詞組自動抽取的方法上，如圖 1. 中英文音譯詞組抽取流程圖所示，主要可分爲四個步驟即 1) 找尋可能的音譯詞組候選詞，2) 英文音素音節化(Syllabification)及音節轉換，3) 產生混淆音矩陣，4)相似度計算。中英文音譯詞組自動抽取模組首先自龐大的文字語料庫中找到一個由標點符號隔開的句子，並在此句子中找到一連續的英文字串，這個英文字串可能包含一個或一個以上的英文字，再以此字串爲中心往兩旁延伸，訂出中文詞尋找範圍，並經由音節相似度比對，過濾掉不符合要求的候選詞。在音節相似度比對之前，必須將英文字串中的每一個英文字，經由英文字轉音及音素音節化程序，將每一個英文音節轉化至中文音節，並產生相對應的混淆音矩陣，這些中文化後的英文音節再與中文候選詞的音節進行相似度比對，找出適當的中文候選詞，並決定是否爲可能的中英文音譯詞組。

以下分別就每一步驟進行更詳細的說明：

2.1 找尋可能的音譯詞組候選詞

[Nagata2001]提出自近乎雙語(Partially Bilingual)的網頁中抽取出翻譯詞組，他們觀察到在日文的網頁中有許多英文詞與日文詞夾雜，大部分的日文相對應翻譯詞即英文詞被包括在括號內，而且緊接在日文詞後面。此種現象不僅出現在日文中，也同時出現在東方語系的中文及韓文中。有許多情形的確是如此，而且括號也常暗示強烈的詞組對應關係，但並非所有的翻譯詞組或音譯詞組皆以這種對應型態出現。以下面這段取自報章的文章說明，

『...MP3 所引起長久以來「版權」的問題，訴訟不斷，爭議不休，始終沒有一個確定的解決方案，經營 Kuro 庫洛 P2P 音樂交換軟體的飛行網，3 日發表 P2P 與版權爭議的解決方案—C2C(Content to Community)，希望能在使用端、科技業者與唱片業者三方中間找到一

個平衡點。…』

在文章中，出現的中英文詞的關係可分爲以下幾種情形，1) 用來形容或補充說明相關詞，如 P2P 與音樂交換軟體的關係。2)日常生活中常用之英文詞，如 MP3。3)無適當的翻譯詞或音譯詞，如 C2C。同時在此段文章亦出現有一音譯詞組『 Kuro 庫洛』，此音譯詞組在文章是緊鄰在一起，但並非以括號突顯音譯詞的型態出現；相反的，文中出現以括號突顯的翻譯詞『 C2C(Content to Community) 』，其關係恰如 1) 用來進一步說明前面所出現的詞。

因此本文參考[Nagata2001]的觀察，使用類似的方法，但並不僅限於處理括號所給予的提示。假設自語料庫中抽出一句子 S= $(s_1 s_2 \ldots s_m)$ ，其中每一個 $s_{i,i=1..m}$ ，是一中文字(Character)或英文字元(Alphabet)，自 S 中先找到可能的英文字串(Word String) EWS ，EWS 包含一個或一個以上的字元，EWS 可表示爲 EWS ∈S, EWS = $(t_1 t_2 \ldots t_n)$ ，其中個別 $t_{i,i=1..n}$ 是以空白斷開來或強迫被切開的英文詞(Word 或 Token)，也是後續處理的基本單位。而可能的中文候選字串則可沿著 EWS 左右兩邊尋找而得，若 l_{ss} 和 l_{se} 分別爲 S 之起始與結束之位置(Location)， l_{es} 和 l_{ee} 分別爲 EWS 之起始與結束之位置， lnc_l 和 lnc_r 分別爲沿 EWS 左右兩邊所遇到的第一個非中文字的位置(括號在此是可被忽略的)。故 EWS 左邊的中文候選字串 CW_l 的範圍爲從 $max(l_{ss}, lnc_l)$ 至 l_{es} ，而 EWS 右邊的中文候選字串 CW_r 的範圍爲從 l_{ee} 至 $min(l_{se}, lnc_r)$ 。這個方法不需經過斷詞程序，即可找到適當的中文候選字串 ，再透過後面所敘述的音節相似度計算，進一步找到正確的音譯詞組。但缺點是可能會將一些音節相似的候選字串納入考慮，不只增加計算量，也使得錯誤率提高。

以上一段文章中部分字串『經營 Kuro 庫洛』例，先找到這個字串所屬的句子，繼而找到句子中的來源語言(即英文)字串『Kuro』，找到英文字串後則沿此英文字串的左右兩傍找尋目標語言(即中文)候選字串，故可找到『經營』以及『庫洛』兩中文候選字串，做爲 Kuro 可能的音譯候選字串。而依上述的方法，雖可找到英文字串『 C2C 』，但卻無法

找到相對應的中文候選字串。

如果要決定『經營』以及『庫洛』兩候選字串是否為 Kuro 的音譯候選詞，則必須經過相似度計算。但來源語言字串與目標語言字串分屬於不同的語言，該如何計算彼此的相似度變成一項問題。直接輸入兩種不同語言的音素資料，並試著找出兩者的關係是一種方式，但問題是如何找出不同語言的音素資料關係。當抽取出大量的音譯詞組時，這樣的轉換關係是很容易可以獲得的；另一種方式則是將兩種語言的音素資料轉換至其中一方或其他第三種表示方式，使得資料表示形式一致，相似度計算相對變成較簡單。待解決的問題則是不同語言的音素資料轉換以及如何對應轉換過程的一對多或多對一關係。

2.2. 英文音素音節化及音節轉換

英文的音素約分為四種類型，即子音(Consonant)、母音(Vowel)、半母音(Semi-Vowel)及鼻音(Nasal) ，其中子音約有 17 個，母音約有 16 個，半母音有 4 個，鼻音有 3 個[Jurafsky2000]。由這些音素所組成的英文音節總數可達數千種，而中文僅有約 414 個音節。由中文音節對應至英文音節則會有一對多的對應問題，使得對應關係更加複雜。而由英文音節對應至中文音節會有多對一的對應問題，但一個中文音節可拆成聲母(Initial)及韻母(Final)兩部份，透過子音與聲母及母音與韻母的對應，可較簡化其對應的複雜度。中英文音節對應關係則可應用現有語言學相關資料[NTNU82] 。

要達到音節轉換之前，必須先將前一節所找到的英文字串 EWS 中的每一個英文詞經由英文字轉音系統 MBRDICO[Pagel98] 轉換成一串的音素，經由 MBRDICO 所產生的音素係以 SAMPA(Speech Assessment Methods Phonetic Alphabet)表示法表示，隨後則將這些音素轉換至以 IPA(International Phonetic Association)表示法表示。將音素轉至以 IPA 表示法表示的目的，是希望能沿用 IPA 與中文聲母與韻母的關係，這種使用 IPA 與中文音節對應關係的方式，將來也可應用至其他語言與中文間的音譯詞組自動抽取。

為求找到中英文音譯詞組，先將轉換後的英文音素音節化，音節化後的英文音素係以子音母音對(Consonant-Vowel Pair) 的方式呈現，[Wan98] 曾類似的音節化演算法，但其方

法伴隨著英文字轉音系統，並非直接針對音素處理。因此本文的音節化程序則直接以音素爲主。

以英文字 Kuro 爲例，首先利用英文字轉音系統轉成音素串/kurə/，然後利用音節化程序將此音素串切割成/ku/ 及 /rə/ 等音節，這些切割後的音節則再利用語言學上的規則，將每一個音節中的音素轉換成相對應的中文聲母或韻母。

如前所述，音譯過程乃是將文字經由發音的方式自一語言轉換至另一語言，但發音時常會因爲腔調或發音部位的不同，導致對於不同的翻譯人員對同一字詞有不同的發音，如 /rə/ 可能會有人發成 /lo/ 或 /ra/，這樣會使得音譯時所處理的音很接近但實際上不相同，更使得相似度比對時無法達到預期效果，音譯詞組自動抽取的成效連帶受到影響。可能的解決方式是設法收集相關混淆音並建立這些混淆音之間的關係，既可處理確實相關的混淆音，又可排除不相關的雜訊音。若以人工方式收集並區分這些混淆音，既曠日又費時，而且不知該從何處著手。如果能夠充分運用電腦的計算能力，快速的收集到這些資訊並建立彼此的關係，對於音譯詞組自動抽取的進行必有很大的幫助。

2.3 產生混淆音矩陣

語音辨認系統可被視爲一有雜訊的通道[Wang2002] ，當一信號輸入至辨認系統時，原本應該辨認到正確的信號，卻可能因爲雜訊與輸入的信號混合，使得混合後的信號被誤認爲其他的信號，導致辨認結果錯誤。這些雜訊對於語音辨認效影響極大，因此通道雜訊的消除是語音辨認研究中重要的課題。混淆音矩陣是語音辨認過程的副產品，它表列了某一音常與某些音混淆在一起，這些混淆音可能是原本正確的音，也可能是常被誤認的音，因此常被用來分析辨認結果，進而改善辨認效能。

取自語音辨認系統的混淆音矩陣，可用來解決如何避免人工方式收集並區分混淆音費時費力的問題，這是因爲這些混淆音矩陣本身即表列了正確及常被混淆的音。但問題是要如何去控制混淆音矩陣品質，將容易混淆的音納入，而將確實因爲雜訊原因產生的音排除。採用同時考慮主音與混淆音是否同時出現及彼此的相依程度的方式，可達到上述目標。

由語音辨認系統產生的混淆音矩陣，並不適合直接拿來應用於音譯詞組抽取上，原因是部份混淆音並非取自良好的語料，剔除這些不良語料後，可較準確的計算出相對應的混淆音。本文所使用的混淆音矩陣有兩種，一是根據全音節計算而得的全音節混淆音矩陣。另一是將全音節混淆音矩陣分別拆成以聲母及韻母爲主的音素混淆音矩陣。因爲發音變化時可能只有聲母、韻母或者兩者同時產生變化，將音節拆成由聲母及韻母分別的對應關係，可以用來克服這種問題。

因爲發音方式不同所產生的問題，可以引用混淆音來解決。除此之外，還有在來源語言中會有發音變異的問題，例如 /t/ 和 /d/ 在英文中的發音，在子音之前或在一串子音群時常會被忽略而不發聲[Jurafsky2000]，這種問題在音譯詞組抽取時必須加以考慮，因爲如果這時候不納入考慮，則會影響到以後中英文音節對應的研究。這種發音變異的問題相當不易處理，原因是在來源語言的音素或音節中這些音是存在的，只是在發音的時候，它因人而異可以發音也可以不發音。如何規範這些不確定性是很大的挑戰。

2.4 相似度計算

公式(1)是機器音譯模組化學習法中用來處理機器音譯的數學模型，但本文的重點是在音譯詞組自動抽取，而且因爲使用 MBRDICO 英文字轉音系統，因此重點將是公式(1)的音素計算 $\boldsymbol{p}(\boldsymbol{C}_s|\boldsymbol{E}_s)$ 上。[Brown93]曾提出一系列的統計式機器翻譯模型應用於機器翻譯詞的相似度計算，這些統計模型稍加修改也可適用於機器音譯上，但問題是該如何將發音變異問題同時納入於相似度計算上。而在相似度計算時，中文候選字串只是大約訂出大約範圍，並不知道英文詞所對應中文候選詞的真正位置。在發音變異問題時，某些被轉換至中文音節後的英文音節可能被忽略，因此必須將英文音節的所有可能組合列出，而對應中文候選詞音節的選取則採取滑動視窗(Sliding Window)的方式，滑動視窗的大小即爲英文音節的數目，中英文候選詞選出後再進行音節間的相似度計算。

在相似度計算之前，先定義以下符號：

EWS：爲一英文字串。
EW：爲 *EWS* 中之一英文詞(Word)。

ES：爲 EW 所對應的音素串。

ECS：爲 ES 被轉換至中文音節的音節串。

ECS_i：爲 ECS 第 i 個音節子集合。

E_{ij}：爲 ECS_i 中第 j 個音節。

CWS：爲抽取音譯詞組時的中文候選字串。

CS：爲 CWS 所對應的音節串。

CW_i：爲 CWS 第 i 個字串子集合。

CS_i：爲 CW_i 所對應的音節串。

C_{ij}：爲 CS_i 中第 j 個音節。

$$p(CWS|EW)=\sum_{CW_i} p(CW_i|EW) \tag{2}$$

音譯詞組抽取時是以計算音節相似度做爲抽取與否的依據，故直接將中英文字串轉換成相對的音素或音節，做進一步的比對。則公式(2)變成

$$p(CWS|EW)\approx\sum_{CS_i} p(CS_i|ES) \tag{3}$$

爲了使音節相似度計算的進行，可以在同一基準上，故將英文音素經音節化並轉換至中文音節。並同時將發音變異問題納入考慮，則公式(3)變成

$$\begin{aligned} &p(CWS|EW) \\ &\approx\sum_{CS_i} p(CS_i|ECS) \\ &=\sum_{CS_i}\sum_{ECS_j} p(CS_i|ECS_j) \end{aligned} \tag{4}$$

而使得公式(4)機率最大的中文詞 $\hat{C}$ 及英文詞 $\hat{E}$，$\hat{J}=(\hat{C},\hat{E})$ 則可下式決定

$$\hat{J}\approx\underset{CS_i\,ECS_j}{argmax}\; p(CS_i|ECS_j) \tag{5}$$

$$\begin{aligned} &=\underset{CS_i\,ECS_j}{argmax}\; p(C_{i1}C_{i2}\ldots C_{in}|E_{j1}E_{j2}\ldots E_{jn}) \\ &\approx\underset{CS_i\,ECS_j}{arrmag}\prod_{k=1}^{n} p(C_{ik}|E_{jk}) \end{aligned} \tag{6}$$

，其中 $p(C_{ik}|E_{jk})$ 爲 C_{ik} 與 E_{jk} 兩個音節間的機率 。

公式 (6) 中的 $p(C_{ik}|E_{jk})$ 是兩個中文音節的機率，這個機率的計算可直接由混淆音矩陣得到，若同時將全音節混淆音矩陣(ASR-Syllable, AS) 、音素混淆音矩陣(ASR-Phoneme, AP) 以及根據語言學規則[NTNU82] 所訂定的音素混淆音矩陣(Rule-based

Phoneme, RP) 納入考慮，則 $p(C_{ik}|E_{jk})$ 可寫爲，

$$p(C_{ik}|E_{jk})=\alpha t_s(C_{ik}|E_{jk})+\beta t_p(C_{ik}|E_{jk})+\gamma t_r(C_{ik}|E_{jk}),\alpha+\beta+\gamma=1, \tag{7}$$

，其中 $t_s(C_{ik}|E_{jk})$ 可直接利用 AS 求得， $t_p(C_{ik}|E_{jk})$ 可利用 AP 求得， $t_r(C_{ik}|E_{jk})$ 是可利用 RP 求得， α 、 β 及 γ 則分別爲 $t_s(C_{ik}|E_{jk})$ 、 $t_p(C_{ik}|E_{jk})$ 及 $t_r(C_{ik}|E_{jk})$ 的權重(Weighting)。

因爲AP是將一個中文音節拆成聲母及韻母，假設聲母及韻母的產生彼此沒有關聯，所以 $t_p(C_{ik}|E_{jk})$ 可由下式求得，

$$t_p(C_{ik}|E_{jk})\approx p(CI_{ik}|EI_{jk})\,p(CF_{ik}|EF_{jk}) \tag{8}$$

，其中 CI_{ik} 及 EI_{jk} 分別爲 C_{ik} 及 E_{jk} 之聲母部份， CF_{ik} 及 EF_{jk} 分別爲 C_{ik} 及 E_{jk} 之韻母部份。RP 是根據的語言學規則所訂定的音素混淆音矩陣，例如國音聲學可以發音部位及發音方法分成兩類，若以發音部位爲分類依據，b、p及m(以IPA表示)都受到上下唇而發音，是屬於同部位的音，故可視爲同一群[NTNU82]。故 $t_r(C_{ik}|E_{jk})$ 可定義爲，

$$\begin{aligned} t_p(C_{ik}|E_{jk}) &\equiv 1 \text{，如果} C_{ik} \text{與} E_{jk} \text{在RP 的同一混淆音群中} \\ &\equiv 0 \text{，其他} \end{aligned} \tag{9}$$

3. 實驗結果與討論

實驗結果是取自台灣區域的繁體中文網頁，過濾後的純文字檔案大小約有 500MB，自其中抽出 80,094 個句子，其檔案大小約爲 5MB，最後共抽出 10,225 個可能的音譯詞組(Transliterated-Term Pair)。在計算檢出率(Recall) 及準確率(Precision)時，採用隨機選擇(Randomly Selected) 方式，自 80,094 個句子中挑出 200 個句子爲評估樣本，在 200 個句子中共產生 488 個候選音譯詞組，可抽出 21 個音譯詞組，經人工確認，其中有 20 個相關。此外經人工確認在 488 個候選音譯詞組中，有 62 個相關的音譯詞組。故檢出率爲 32.26%，準確率爲 95.23%。

表 2 所示爲部份的實驗結果，其中括號內的數字代表出現的次數，這也可得知對某一英文字而言，大部分相對的中文音譯詞爲何。有趣的是許多中文詞僅出現一次，這些低

頻詞並無法以詞共現(Word Co-occurrence)的翻譯詞抽取方法抽取出來，但透過以發音爲特徵的音譯詞組抽取是可以將這些低頻詞抽取出來。

由表 2 的實驗結果可發現發音變異的問題確實存在於音譯處理中，例如英文的 /t/ 常對應至中文的『t』或『t'』(以 IPA 符號表示)，而英文的/d/ 則也可能會對應到中文的『t』或『l』，這表示應用於混淆音矩陣於音譯詞組抽取是合理的。另外表 2 中的 Charles 對應至其中的一個中文音譯詞『策略師』，從以發音爲特徵的相似度計算的角度來看，並沒有錯誤，但對於音譯的角度而言卻是個錯誤。這個錯誤是因爲在找尋可能的音譯詞組時範圍設定過寬，導致將相連接的其他文字也納入範圍之內，再加上相似度計算彼此特性相近，故將此錯誤納入。

英文字	音譯中文詞 1	音譯中文詞 2	音譯中文詞 3	音譯中文詞 4
Robert	霍伯德(1)	羅伯特(4)		
Charles	查理斯(1)	查爾斯(5)	察爾斯(1)	策略師(1)
Michael	麥可(4)	麥克(6)	邁可(1)	邁克(1)
Richard	李查德(1)	李查羅(1)	李察(1)	理查(3)

表 2 經由音譯詞組自動抽取程序所得的部份實驗結果

在公式(7) 中使用了 AS、AP 及 RP 三種方法，但三者的效能如何尚不得而知。表 3 爲分別單獨使用 AS、AP 及 RP 三種方法來計算音節相似度，個別所產生的數量及佔總數的比重，其中第二列初步抽出數量(Raw Count)是表示音譯詞組抽取過程中未先以一般詞典先行濾除常用的英文字，故有一些非人名地名的詞彙被抽出，如『Homework』即有相對應的音譯詞『洪沃客』，這可以顯現出網路中無奇不有的特性。第三列含獨特(Unique)英文詞的音譯詞組是指僅計算不重複的英文詞時所得的音譯詞組數量及比重，因爲一個音譯詞組包含一個中文詞及一個英文詞，僅計算不重複英文詞數量，可了解平均一個英文詞所對應到的中文詞數量。第四列含獨特中英文詞的音譯詞組是指同時計算不重複的中英文音譯詞組數量及比重，以了解平均一個中英文音譯詞組的重複狀況。

從表中可看出使用語言學規則所歸納而得的音素混淆音矩陣(RP)的效果最好，而由語音辨認所得的混淆音矩陣並不如預期來的好，由語音辨認結果所產生的混淆音矩陣中分析

發現 ，若中文音節符號以 IPA 表示，如果 A 是 B 的混淆音，並不保證 B 一定是 A 的混淆音，例如『p'』常被念成『p』，因此『p'』常是『p』的混淆音，但『p』並不見得會常被念成『p'』，所以『p』不見得會是『p'』的混淆音。這也使得應用混淆音矩陣於音譯詞組抽取時，有些正確的音譯詞組可能無法被抽出。但使用語言學規則所歸納而得的音素混淆音矩陣是經過不斷的微調所產生的結果，例如上述之非雙向規則，就可以在這時候由人工直接加入規則中。

項目及方法	三種方法均使用	AS	AP	RP
初步抽出數量	10,225	4,964(485.%)	8,254(80.7%)	9,175(89.7%)
含獨特(Unique)英文詞的音譯詞組	3,742	1,887(50.4%)	3,086(82.5%)	3,412(91.2%)
含獨特中英文詞的音譯詞組	4,779	2,400(50.2%)	3,798(79.5%)	4,224(88.4%)

表 3 分別單獨使用 AS，AP 及 RP 三種方法所得的數量及比重

爲了進一步瞭解 AS，AP 及 RP 三者的影響如何。表 4 爲交叉驗證分別單獨使用 AS，AP 及 RP 三種方法所得資料之數量及比例，以了解三種方法彼此間的差異程度，其中可發現 AP 與 RP 在含獨特英文詞的音譯詞組與含獨特中英文詞的音譯詞組交集部份兩者重疊性均很高(分別爲 AP 的 98.3%及 RP 的 88.9%)，雖然在含獨特英文詞音譯詞組部分 AP 差集僅佔 1.7% ，RP 差集僅佔 22.4% ，但在含獨特中英文詞的音譯詞組 AP 的差集則上升至佔 2% ，RP 的差集則下降至 12% 。就音譯詞組自動抽取的角度而言，由 AP 來取代經過微調過的 RP 應該是可以合理的，這是因爲微調過的 RP 須具備許多的語言學知識，而且也必須耗費許多時間觀察許多可能的情形。此外雖然 RP 在含獨特英文詞的音譯詞組交集部份可以包含 86%的 AS，在含獨特中英文詞音譯詞組交集部份可以包含 82.6%的 AS，但 AS 在含獨特英文詞的音譯詞組仍有 13.9% 的獨特詞組，在含獨特中英文詞的音譯詞組仍有 17.4% 的獨特詞組，顯示 AS 不應被完全捨棄。

項目及方法	AS vs AP	AP vs RP	RP vs AS
含獨特(Unique)英文詞的音譯詞組(交集)	1,514 (80.2% to AS) (49.1% to AP)	3,034 (98.3 % to AP) (88.9% to RP)	1,624 (86% to AS) (47.6% to RP)

含獨特英文詞的音譯詞組(差集)	373(19.8% to AS)
	1,572(51% to AP)
		52(1.7% to AP)
		764(22.4% to RP)
			263(13.9% to AS)
			1,788(52.4 to RP)

含獨特(Unique)中英文詞的音譯詞組(交集)	1,824 (76% to AS) (48% to AP)	3,721 (98% to AP) (88.1% to RP)	1,982 (82.6% to AS) (46.9% to RP)

含獨特中英文詞的音譯詞組(差集)	576(24% to AS)
	1,974(52% to AP)
		77(2% to AP)
		503(12% to RP)
			418(17.4% to AS)
			2,242(53.1% to RP)

表 4 交叉驗證分別單獨使用 AS，AP 及 RP 三種方法所得資料之數量及比例

在[Lin2000] 曾討論到中英文詞組音譯失敗的原因，其中有一項是約定成俗但聲音不相近的音譯，由本文實驗中可以發現，由於許多人名或地名係先被引進至英文，在處理中英文詞組音譯詞組抽取時，若能把字源因素納入考慮，以原始語言的發音規則發音，則有較大的機會被抽取出來。以 Bach (巴哈)例，因爲 Bach 是一個德國人名，若能以德語發音，則更能貼近音譯是以發音方式將一詞從一個文字轉換到另一文字的特性。

4. 結論

機器音譯常用來處理文章中許多人名、地名、組織名等專有名詞，是機器翻譯中重要的一環。本文提出基於語音辨認系統所產生的混淆音矩陣，用來解決中英文音譯詞組自動抽取所面臨的發音變異問題，並自網頁中抽取出大量的中英文音譯詞組。由實驗結果發現，本文所提出的方法確實有效的處理發音變異問題。中英文音譯詞組自動抽取是研究中英文詞組音譯的第一步，未來將繼續進行中英文音節的自動轉換等相關的中英文詞組音譯研究。

5. 致謝

中華電信研究所王文俊博士提供語音辨認系統所產生的混淆音矩陣，由於這個混淆

音矩陣資料，使得中英文音譯詞組自動抽取變成可行。中華電信研究所賴玟杏小姐提供許多語言學的資訊。中央研究院資訊所簡立峯博士提供許多寶貴的意見。在此一併致謝。

6. 參考文獻

[Brill2001] Eric Brill, Gary Kacmarcik, Chris Brockett, “Automatically Harvesting Katakana-English Term Paris from Search Engine Query Logs”, In *Proceedings of NLPRS'2001*

[Brown93] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer, “The Mathematics of Statistical Machine Translation: Parameter Estimation”, Computational Linguistics, 19(2), pp.263-311, 1993

[Jeong97] Kil-Soon Jeong, Yun-Hyung Kwon, and Sung-Hyun Myaeng, “Construction of Equivalence Classes of Foreign Words through Automatic Identification and Extraction, NLPRS'97

[Jung2000] SungYoung Jung, SungLim Hong, and Eunok Paek, “An English to Korea Transliteration Model of Extended Markov Windows”, In *Proceedings of COOLING'2000*

[Jurafsky2000] Daniel Jurafsky and James H. Martin, Speech and Language Processing, pp. 156-163, Prentice-Hall, New Jersey, 2000

[Kang2000] In-Ho Kang and GilChang Kim, “English-to-Korean Transliteration Using Multiple Unbounded Overlapping Phoneme Chinks”, In *Proceedings of COLING'2000*, 2000

[Knight98] Kevin Knight and Jonathan Graehl, “Machine Transliteration”, Computational Linguistics, 24(4), pp. 599-612, 1998

[Lee2003] Chun-Jen Lee and Jason S. Chang, “Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts Using a Statistical Machine Transliteration Model”, In *Proceedings of NAACL*, pp. 96-103, 2003

[Lee98] Jae-Sung Lee and Key-Sun Choi, “English to Korea Statistical Transliteration for Information Retrieval”, Computer Processing of Oriental Languages, Vol. 12, No. 1, pp. 17-37, 1998.

[Lin2000] Wei-Hao Lin and Hsin-Hsi Chen, “Similarity Measure in Backward Transliteration between Different Character Set and Its Application to CLIR”, In *Proceedings of Computational Linguistics Conference XIII*, pp. 97-113, 2000 (in Chinese)

[Llitojos2001]Ariadna Font Llitjos and A. Black, “Knowledge of Language Origin improves Pronunciation Accuracy of Proper Names, In *Eurospeech'2001* Vol. 3, Aalborg Denmark, pp.1919-1922, 2001

[Nagata2001] Masaaki Nagata, Teruka Saito, and Kenji Suzuki, “Using the Web as a Bilingual Dictionary”, In *Proceedings of ACL'2001 DD-MT Workshop*, 2001

[Oh2002] Jong-Hoon Oh and Key-Sun Choi, “An English-Korea Transliteration Model Using Pronunciation and Contextual Rules, In *Proceedings of COLING'2002*, Taipei, Taiwan, 2002

[Pagel98] Vincent Pagel, Kevin Lenzo, and Alan W. Black, "Letter to Sound Rules for Accented Lexicon Compression", In *Proceedings of the ICSLP'98*, Sydney, Australia, 1998

[Wan98] Stephen Wan and Cornelia Maria Verspoor, "Automatic English-Chinese Name Transliteration for Development of Multilingual Resources", In *Proceedings of 17th COLING and 36th ACL*, pp. 1352-1356, Montreal , Quebec, Canada, 1998

[Xiao2002] Jing Xiao, Jimin Liu and Tat-Seng Chua, "Extracting Pronunciation-translated Names from Chinese Texts Using Bootstrapping Approach", In *Proceedings of COLING'2002*, Taipei, Taiwan, 2002

[NIHOGO90] 日本語知識百科，和風語言雜誌，豪風出版社，1990

[NTNU82] 國音學，國立台灣師範大學國音教材編輯委員會，正中書局，1982

[Wang2002] 『雜訊語音辨認技術』，王文俊等，中華電信研究所技術報告編號 91-31-002，2002

Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses

Chien-Cheng Wu
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
g904374@oz.nthu.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
jschang@cs.nthu.edu.tw

Abstract

In this paper, we describe an algorithm that employs syntactic and statistical analysis to extract bilingual collocations from a parallel corpus. The preferred syntactic patterns are obtained from idioms and collocations in a machine-readable dictionary. Phrases matching the patterns are extract from aligned sentences in a parallel corpus. Those phrases are subsequently matched up via cross-linguistic statistical association. Statistical association between the whole collocations as well as words in collocations is used jointly to link a collocation and its counterpart collocation in the other language. We experimented with an implementation of the proposed method on a very large Chinese-English parallel corpus with satisfactory results.

1. Introduction

Collocations like terminology tend to be lexicalized and have a somewhat more restricted meaning than the surface form suggested (Justeson and Katz 1995). Collocations are recurrent combinations of words that co-occur more often than chance. The words in a collocation may appear next to each other (rigid collocations) or otherwise (flexible/elastic collocations). On the other hand, collocations can be classified into lexical and grammatical collocations (Benson, Benson, Ilson, 1986).

Lexical collocations are formed between content words, while the grammatical collocation has to do with a content word and function words or a syntactic structure. Collocations are pervasive in all types of writing and can be found in phrases, chunks, proper names, idioms, and terminology. Collocations in one language are usually difficult to translate directly into another language word by word, therefore present a challenge for machine translation systems and second language learners alike.

Automatic extraction of monolingual and bilingual collocations are important for many applications, including natural language generation, word sense disambiguation, machine translation, lexicography, and cross language information retrieval. Hank and Church (1990) pointed out the usefulness of mutual information for identifying monolingual collocations in lexicography. Justeson and Katz (1995) proposed to identify technical terminology based on preferred linguistic patterns and discourse property of repetition. Among many general methods presented by Manning and Schutze (1999), best results can be achieved by filtering based on both linguistic and statistical constraints. Smadja (1993) presented a method called EXTRACT, based on means variance of the distance between two collocates capable of computing elastic collocations. Kupiec (1993) proposed to extract bilingual noun phrases using statistical analysis of co-occurrence of phrases. Smadja, McKeown, and Hatzivassiloglou (1996) extended the EXTRACT approach to handling of bilingual collocation based mainly on the statistical measures of Dice coefficient. Dunning (1993) pointed out the weakness of mutual information and showed that log likelihood ratios are more effective in identifying monolingual collocations especially when the occurrence count is very low.

Both Smadja and Kupiec used the statistical association between the whole of collocations in two languages without looking into the constituent words. For a collocation and its paraphrasing translation counterpart, that is reasonable. For instance, with the bilingual collocation (“擠破頭”, “stop at nothing”) in Example 1, it is not going to help looking into the statistical association between “stopping” and “擠” [ji] (sqeeze) (or “破” [bo, broken] and “頭” [tou, head] for that matter). However, with the bilingual collocation (“減薪”, “pay cut”) in Example 2, considering the statistical association between “pay” and “薪” [xin] (wage) as well as “cut” and “減” [jian, reduce] certainly makes sense. Moreover, we have more data to make statistical inference between words than phrases. Therefore, measuring the statistical association of collocations based on constituent words will help to cope with the data sparseness problem. We will be able to extract bilingual collocations with high reliability even when they appear together in aligned sentences only once or twice.

Example 1

They are **stopping at nothing** to get their kids into "star schools"

他們**擠破頭**也要把孩子送進明星小學

Source: 1995/02 No Longer Just an Academic Question: Educational Alternatives Come to Taiwan

Example 2

Not only haven't there been layoffs or **pay cuts**, the year-end bonus and the performance review bonuses will go out as usual .

不但不虞裁員、**減薪**，年終獎金、考績獎金還都照發不誤

Source: 1991/01 Filling the Iron Rice Bowl

Since the collocations could be rigid or flexible in both languages, we can generally classify the match type of bilingual collocation into three types. In Example 1, (“擠破頭”, “stop at nothing”) is a pair of rigid collocations, and (“把…送

進”, “get … into”) is a pair of elastic collocations. In Example 3 ,(“走…的路線’, “take the path of”) gives the example for a pair of elastic and rigid collocations.

Example 3

Lin Ku-fang, a worker in ethnomusicology, worries too, but his way is not to take the path of revolutionizing Chinese music or making it more "symphonic"; rather, he goes directly into the tradition, looking into it for "good music" that has lasted undiminished for a hundred generations.

民族音樂工作者林谷芳也非不感到憂心，但他的方法是：不**走**國樂改革或「交響化」**的路**，而是直接面對傳統、從中尋找歷百代不衰的「好聽音樂」。

Source: 1997/05 A Contemporary Connoisseur of the Classical Age--Lin Ku-fang's Canon of Chinese Classical Music

In this paper, we describe an algorithm that employs syntactic and statistical analyses to extract rigid lexical bilingual collocations from a parallel corpus. Here, we focus on the bilingual collocations, which have some lexical correlation between them and are rigid in both languages. To cope with the data sparseness problem, we use the statistical association between two collocations as well as that between their constituent words. In Section 2, we describe how we obtain the preferred syntactic patterns from collocation and idioms in a machine-readable dictionary. Examples will be given to show how collocations matching the patterns are extracted and aligned for a given aligned sentence pairs in a parallel corpus. We experimented with an implementation of the proposed method for the Chinese-English parallel corpus of Sinorama Magazine with satisfactory results. We describe the experiments and evaluation in Section 3. The limitations and related issues will be taken up in Section 4. We conclude and give future directions in Section 5.

2. Extraction of Bilingual Collocations

In this chapter, we will describe how we obtain the bilingual collocation by using the preferred syntactic patterns and associative information. Consider a pair of aligned sentences in a parallel corpus such as Example 4 given below:

Example 4

The civil service rice bowl, about which people always said "you can't get filled up, but you won't starve to death either," is getting a new look with the economic downturn. Not only haven't there been layoffs or pay cuts, the year-end bonus and the performance review bonuses will go out as usual, drawing people to compete for their own "iron rice bowl."

以往一向被認爲「吃不飽、餓不死」的公家飯，值此經濟景氣低迷之際，不但不虞裁員、減薪，年終獎金、考績獎金還都照發不誤，因而促使不少人回頭競逐這隻「鐵飯碗」。

Source: 1991/01 Filling the Iron Rice Bowl

We are supposed to extract the following collocations and translation counterparts:

(civil service rice bowl, 公家飯)
(get filled up, 吃…飽)
(starve to death, 餓…死)
(economic downturn, 經濟景氣低迷) (pay cuts, 減薪)
(year-end bonus, 年終獎金)
(performance review bonuses, 考績獎金)
(iron rice bowl, 鐵飯碗)

In Section 2.1, we will first show how that process is carried out for Example 4 under the proposed approach. The formal description will be given in Section 2.2.

2.1 An Example of Extracting Bilingual Collocations

To extract bilingual collocations, we first run part of speech tagger on both sentences. For instance, for Example 4, we get the results of tagging in Example 4A and 4B.

In the tagged English sentence, we identify phrases that follow a syntactic pattern from a set of training data of collocations. For instance, “jj nn” is one of the preferred syntactic structures. So, ”civil service,” “economic downturn,” and “own iron,”…etc are matched. See Table 1 for more details. For Example 4, the phrases in Example 4C and 4D are considered as potential candidates for collocations because they match at least two distinct collocations listed in LDOCE:

Example 4A

The/at civil/jj service/nn rice/nn bowl/nn ,/, about/in which/wdt people/nns always/rb said/vbd "/`` you/ppss can/md 't/* get/vb filled/vbn up/rp ,/, but/cc you/ppss will/md 't/* starve/vb to/in death/nn either/cc ,/rb "/" is/bez getting/vbg a/at new/jj look/nn with/in the/at economic/jj downturn/nn ./. Not/nn only/rb have/hv 't/* there/rb been/ben layoffs/nns or/cc pay/vb cuts/nns ,/, the/at year/nn -/in end/nn bonus/nn and/cc the/at performance/nn review/nn bonuses/nn will/md go/vb out/rp as/ql usual/jj ,/, drawing/vbg people/nns to/to compete/vb for/in their/pp$ own/jj "/`` iron/nn rice/nn bowl/nn ./. "/"

Example 4B

以往/Nd 一向/Dd 被/P02 認爲/VE2 「/PU 吃/VC 不/Dc 飽/VH 、/PU 餓不死/VR 」/PU 的/D5 公家/Nc 飯/Na ，/PU 值此/Ne 經濟/Na 景氣/Na 低迷/VH 之際/NG ，/PU 不但/Cb 不虞/VK 裁員/VC 、/PU 減薪/VB ，/PU 年終獎金/Na 、/PU 考績/Na 獎金/Na 還都/Db 照/VC 發/VD 不誤/VH ，/PU 因而/Cb 促使/VL 不少/Ne 人/Na 回頭/VA 競逐/VC 這/Ne 隻/Nf「/PU 鐵飯碗/Na 」/PU

Example 4C

“civil service,” ” rice bowl,” ” iron rice bow,” ” fill up,” ” economic downturn,” ” end bonus,” ” year - end bonus,” ” go ut,” ” performance review,” ” performance review bonus,” ” pay cut,” ” starve to death,” ” civil service rice,” ” service rice,” ” service rice bowl,” ” people always,” ” get fill,” ” people to compete,” ” layoff or pay,” ” new look,” ” draw people”

Example 4D

“吃不飽,” “餓不死,” “公家飯,” “經濟景氣,” “景氣低迷,” “經濟景氣低迷,” “裁員,” “減薪,” “年終獎金,” “考績獎金,” “競逐,” ” 鐵飯碗.”

Although “new look” and “draw people” are legitimate phrases, they more like “free combinations” than collocations. That reflects from their low log likelihood ratio values. For that, we proceed to see how tightly the two words in overlapping bigrams within a collocation associated with each other; we calculate the minimum of the log likelihood ratio values for all bigrams. With that, we filter out the candidates that its POS pattern appear only once or has minimal log likelihood ratio of less than 7.88. See Tables 1 and 2 for more details.

In the tagged Chinese sentence, we basically proceed the same way to identify the candidates of collocations and based on the preferred linguistic patterns of the Chinese translation of collocations in an English-Chinese MRD. However, since there is no space delimiter between words, it is at time difficult to say whether the translation is a multi-word collocation or it is a single word and should not be considered as a collocation. For that reason, we take multiword and singleton phrases (with two or more characters) into consideration. For instance, in the tagged Example 2C, we will extract and consider the following candidates as the counterparts of English collocations:

Notes that at this point, we are not pinned down on the collocations and allow overlapping and conflicting candidates such as “經濟景氣,” “景氣低迷,” “經濟景氣低迷.” See Tables 3 and 4 for more details.

Table 1 The initial candidates extracted based on preferred patterns trained on collocations listed in LDOCE.

E-collocation Candidate	Part of Speech	Pattern Count	Min LLR
civil service	jj nn	1562	496.156856

rice bowl	nn nn	1860	99.2231161
iron rice bowl	nn nn nn	8	66.3654678
filled up	vbn rp	84	55.2837871
economic downturn	jj nn	1562	51.8600979
*end bonus	nn nn	1860	15.9977283
year - end bonus	nn - nn nn	12	15.9977283
go out	vb rp	1790	14.6464925
performance review	nn nn	1860	13.5716459
performance review bonus	nn nn nn	8	13.5716459
pay cut	vb nn	313	8.53341082
starve to death	vb to nn	26	7.93262494
civil service rice	jj nn nn	19	7.88517791
*service rice	nn nn	1860	7.88517791
*service rice bowl	nn nn nn	8	7.88517791
* people always	nn rb	24	3.68739176
get filled	vb vbn	3	1.97585732
* people to compete	nn to vb	2	1.29927068
* layoff or pay	nn cc vb	14	0.93399125
* new look	jj nn	1562	0.63715518
* draw people	vbg nn	377	0.03947748

* indicates invalid candidate

Table 2 The candidates of English collocation based on both preferred linguistic patterns and log likelihood ratio

E-collocation Candidate	Part of Speech	Pattern Count	Min LLR
civil service	jj nn	1562	496.156856
rice bowl	nn nn	1860	99.2231161
iron rice bowl	nn nn nn	8	66.3654678
filled up	vbn rp	84	55.2837871
economic downturn	jj nn	1562	51.8600979
*end bonus	nn nn	1860	15.9977283
year - end bonus	nn - nn nn	12	15.9977283
go out	vb rp	1790	14.6464925
performance review	nn nn	1860	13.5716459
performance review bonus	nn nn nn	8	13.5716459
pay cut	vb nn	313	8.53341082
starve to death	vb to nn	26	7.93262494
civil service rice	jj nn nn	19	7.88517791
*service rice	nn nn	1860	7.88517791
*service rice bowl	nn nn nn	8	7.88517791

* indicates invalid candidate

Table 3 The initial candidates extracted by the Chinese collocation recognizer.

C-collocation Candidate	POS	Patter Count	Min LLR
不少 人	Ed Na	2	550.904793
*被 認為	PP VE	6	246.823964
景氣 低迷	Na VH	97	79.8159904
經濟 景氣 低迷	Na Na VH	3	47.2912274
經濟 景氣	Na Na	429	47.2912274
公家 飯	Nc Na	63	42.6614685
*不 飽	Dc VH	24	37.3489687
考績 獎金	Na Na	429	36.8090448
不虞 裁員	VJ VA	3	17.568518
回頭 競逐	VA VC	26	14.7120606
*還都 照	Db VC	18	14.1291893
*發 不誤	VD VH	2	13.8418648
*低迷 之際	VH NG	10	11.9225789
*值此 經濟 景氣	VA Na Na	2	9.01342071
*值此 經濟	VA Na	94	9.01342071
*照 發	VC VD	2	6.12848087
*人 回頭	Na VA	27	1.89617179

* indicates invalid candidate

Table 4 The result of Chinese collocation candidates extracted which are picked out. (the ones which have no Min LLR are singleton phrases)

C-collocation Candidate	POS	Patter Count	Min LLR
不少 人	Ed Na	2	550.904793
*被 認為	PP VE	6	246.823964
景氣 低迷	Na VH	97	79.8159904
經濟 景氣 低迷	Na Na VH	3	47.2912274
經濟 景氣	Na Na	429	47.2912274
公家 飯	Nc Na	63	42.6614685
*不 飽	Dc VH	24	37.3489687
考績 獎金	Na Na	429	36.8090448
不虞 裁員	VJ VA	3	17.568518
回頭 競逐	VA VC	26	14.7120606
*還都 照	Db VC	18	14.1291893
*發 不誤	VD VH	2	13.8418648
*低迷 之際	VH NG	10	11.9225789
*值此 經濟 景氣	VA Na Na	2	9.01342071
*值此 經濟	VA Na	94	9.01342071
之際	NG	5	

經濟	Na	1408	
景氣	Na	1408	
年終獎金	Na	1408	
考績	Na	1408	
獎金	Na	1408	
鐵飯碗	Na	1408	
公家	Nc	173	
以往	Nd	48	
值此	VA	529	
裁員	VA	529	
回頭	VA	529	
減薪	VB	78	
競逐	VC	1070	
認爲	VE	139	
低迷	VH	731	
不誤	VH	731	
不虞	VJ	205	
促使	VL	22	
餓不死	VR	14	

To align collocations in both languages, we follow the idea of Competitive Linking Algorithm proposed by Melamed (1996) for word alignment. Basically, the proposed algorithm **CLASS**, Collocation Linking Algorithm based on Syntax and Statistics, is a greedy method that selects collocation pairs. The pair with higher association value takes precedence over those with a lower value. CLASS also imposes a one-to-one constraint on the collocation pairs selected. Therefore, the algorithm at each step considers only pairs with words not selected before. However, CLASS differs with CLA in that it considers the association between the two candidate collocations in two aspects:

- Logarithmic Likelihood Ratio between the two collocations in question as a whole.
- Translation probability of collocation based on constituent words

For Example 4, the CLASS Algorithm first calculates the counts of collocation candidates in the English and Chinese part of the corpus. The collocations are matched up randomly across from English to Chinese. Subsequently, the co-occurrence counts of these candidates across from English to Chinese are also tallied. From the monolingual collocation candidate counts and cross language concurrence counts, we produce the LLR values and the collocation translation probability derived from word alignment analysis.. Those collocation pairs with zero translation probability are ignored. The lists are sorted in descending order of LLR values, and the pairs with low LLR value are discarded. Again, for Example 4, the greedy selection process of collocation starts with the first entry in the sorted list and proceeds as follows:

1. The first, third, and fourth pairs, ("iron rice bowl," "鐵飯碗"), ("year-end bonus," "年終獎金"), and ("economic downturn," "經濟景氣低迷"), are selected first. And that would exclude conflicting pairs from being considered including the second, fifth pairs and so on.
2. The second, fifth entries ("rice bowl," "鐵飯碗") and ("economic downturn," "值此經濟景氣") and so on, conflict with the second and third entries that are already selected. Therefore, CLASS skips over those.
3. The entries ("performance review bonus," "考績獎金"), ("civil service rice," "公家飯"), ("pay cuts," "減薪"), and ("starve to death," "餓不死") are selected next.
4. CLASS proceeds through the rest of the list and the other list without finding any entries that do not conflict with the seven entries selected previously.
5. The program terminates and output a list of seven collocations.

Table 5 The result of Chinese collocation candidates extracted which are picked out. The shaded collocation pairs are selected by the CLASS (Greedy Alignment Linking E).

English collocations	Chinese collocations	LLR	Collocation Translation Prob.
iron rice bowl	鐵飯碗	103.3	0.0202
rice bowl	鐵飯碗	77.74	0.0384
year-end bonus	年終獎金	59.21	0.0700
economic downturn	經濟 景氣 低迷	32.4	0.9359

economic downturn	值此 經濟 景氣	32.4	0.4359
...	. . .	**...**	**...**
performance review bonus	考績 獎金	30.32	0.1374
economic downturn	景氣 低迷	29.82	0.2500
civil service rice	公家 飯	29.08	0.0378
pay cuts	減薪	28.4	0.0585
year-end bonus	考績 獎金	27.35	0.2037
performance review	考績	27.32	0.0039
performance review bonus	年終獎金	26.31	0.0370
starve to death	餓不死	26.31	0.5670
...	. . .	...	...
rice bowl	公家 飯	24.98	0.0625
iron rice bowl	公家 飯	25.60	0.0416
…	…	…	…

2.2 The Method

In this section, we describe formally how CLASS works. We assume availability of a parallel corpus and a list of collocations in a bilingual MRD. The sentences and words have been aligned in the parallel corpus. We will describe how **CLASS** extracts bilingual collocations in the parallel corpus. CLASS carries out a number of preprocessing steps to calculate the following information:

1. Lists of preferred POS patterns of collocation in both languages.
2. Collocation candidates matching the preferred POS patterns.
3. N-gram statistics for both languages, N = 1, 2.
4. Log likelihood Ratio statistics for two consecutive words in both languages.
5. Log likelihood Ratio statistics for a pair of candidates of bilingual collocation across from one language to the other.
6. Content word alignment based on Competitive Linking Algorithm (Melamed 1997).

Figure 1 illustrates how the method works for each aligned sentence pair (*C*, *E*) in the corpus. Initially, part of speech taggers process *C* and *E*. After that, collocation candidates are extracted based on preferred POS patterns and statistical association between consecutive words in a collocation. The collocation candidates are subsequently matched up across from one language to the other. Those pairs are sorting according to log likelihood ratio and collocation translation probability. A greedy selection process goes through the sorted list and selects bilingual collocations subject to one to one constraint. The detailed algorithm is given below:

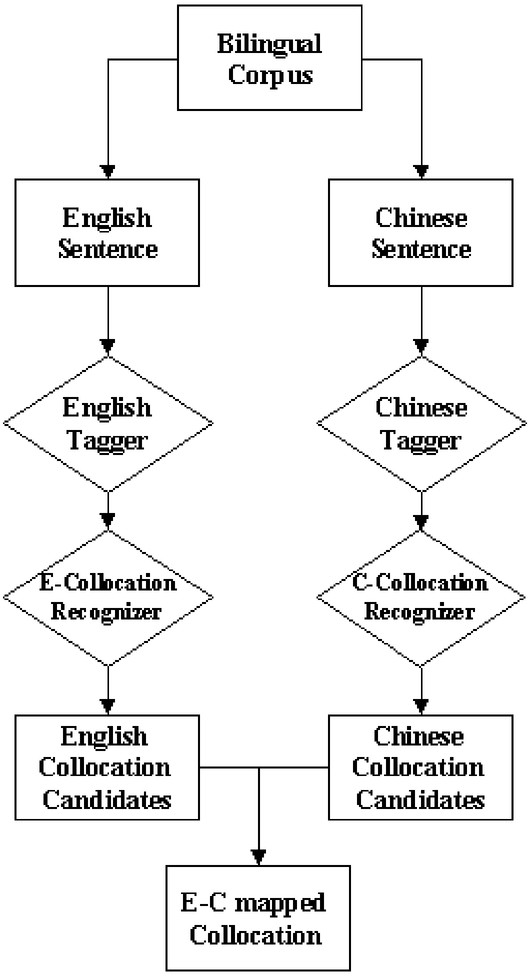


Figure 1 The major components in the proposed CLASS algorithm

Preprocessing: Extracting preferred POS patterns *P* and *Q* in both languages

Input: A list of bilingual collocations from a machine-readable dictionary

Output:

1. Perform part of speech tagging for both languages

2. Calculate the number of instances for all POS patterns in both languages
3. Eliminate the POS patterns with instance count 1.

Collocation Linking Alignment based on Syntax and Statistics

Extract bilingual collocations in aligned sentences.

Input:

(1) A pair of aligned sentences (C, E), C = (C_1 C_2 ... C_n) and E = (E_1 E_2 ... E_m)
(2) Preferred POS patterns P and Q in both languages

Output: Aligned bilingual collocations in (C, E)

1. C is segmented and tagged with part of speech information T.
2. E is tagged with part of speech sequences S.
3. Match T against P and S against Q to extract collocation candidates X_1, X_2,....X_k in English and Y_1, Y_2, ...,Y_e in Chinese.
4. Consider bilingual each collocation candidates (X_i , Y_j) in turn and calculate the minimal log likelihood ratio LLR between X_i and Y_j

   MLLR (D) = $\min_{i=1,n-1} LLR(W_\mathrm{i}\, , W_{\mathrm{i}+1})$
5. Eliminate candidates with LLR smaller than a threshold (7.88).
6. Match up all possible linking from English collocation candidates to Chinese ones: (D_1, F_1), (D_1, F_2), … (D_i, F_j), … (D_m, F_n).

Log-likelihood ratio: LLR(x;y)

$$LLR(x;y) = -2\log_2 \frac{p_1^{k_1}(1-p_1)^{n_1-k_1}\, p_2^{k_2}(1-p_2)^{n_2-k_2}}{p^{k_1}(1-p)^{n_1-k_1}\, p^{k_2}(1-p)^{n_2-k_2}}$$

k_1 : # of pairs that contain x and y simultaneously.
k_2 : # of pairs that contain x but do not contain y.
n_1 : # of pairs that contain y
n_2 : # of pairs that does not contain y
p_1 = k_1/n_1, $p_2 = k_2/n_2$,
p = $(k_1+k_2)/(n_1+n_2)$

7. Calculate LLR for (D_i, F_j), and discard pairs with LLR value lower than 7.88.

Collocation translation probability P(x | y)

$$P(D_\mathrm{i} \mid F_\mathrm{j}) = \frac{1}{k} \sum_{e \in F_j} \max_{c \in D_i} P(c \mid e)$$

k : number of words in the English collocation F_j

8. The candidate list of bilingual collocations is considered only the one with non-zero collocation translation probability P(D_i, F_j) values. The list is then sorted by the LLR values and collocation translation probability.
9. Go down the list and select a bilingual collocation if it is not

conflicting with previous selection.

10. Output the bilingual collocation selected in Steps 10.

3. Experiments and Evaluation

We have experimented with an implementation of CLASS based on Longman dictionary of Contemporary English, English-Chinese Edition and the parallel corpus of Sinorama magazine. The articles from Sinorama cover a wide range of topics, reflecting the personalities, places, and events in Taiwan for the past three-decade. We experiment on articles mainly dated from 1995 to 2002. Sentence and word alignment were carried out first for Sinorama parallel Corpus.

Sentence alignment is a very important aspect of the CLASS. It is the basis of a good collocation alignment. We using a new alignment method based on punctuation statistics (Yeh & Chang, 2002). The punctuation-based approach outperforms the length-based approach with precision rates approaching 98%. With the sentence alignment approach, we obtain approximately 50,000 reliably aligned sentences containing 1,756,000 Chinese words (about 2,534,000 Chinese characters) and 2,420,000 English words in total.

The content words were aligned based on Competitive Linking Algorithm. Alignment of content words resulted in a probabilistic dictionary with 229,000 entries. We evaluated 100 random sentence samples with 926 linking types, and the precision is 93.3%. Most of the errors occurred with English words having no counterpart in the corresponding Chinese sentence. The translators do not always translate the word for word. For instance, with the word “water” in Example 4, it seems that these is no corresponding pattern in the Chinese sentence. Another major cause of errors is collocations that are not translated compositionally. For instance, the word “State” in the Example 6 is a part of the collocation “United States”, and “美國” is more highly associated with “United” than “States”, therefore due to one-to-one constraint “States” will not be aligned with “美國”. Most often, it will be aligned incorrectly. About 49% error links belongs to this kind.

Example 5

The boat is indeed a vessel from the mainland that illegally entered Taiwan waters. The words were a "mark" added by the Taiwan Garrison Command before sending it back.

編按：此船的確是大陸偷渡來台船隻，那八個字只不過是警總在遣返前給它加的「記號」！

Source: 1990/10 Letters to the Editor

Example 6

Figures issued by the American Immigration Bureau show that most Chinese immigrants had set off from Kwangtung and Hong Kong, which is why the majority of overseas Chinese in the United States to this day are of Cantonese origin.

由美國移民局發表的數字來看，中國移民以從廣東、香港出海者最多，故到現在為止，美國華僑仍以原籍廣東者佔大多數。

Source: 1990/09 All Across the World: The Chinese Global Village

We obtained word-to-word translation probability from the result of word alignment. The translation probability P(*c*|*e*) is given below:

$$\mathrm{P}(c|e) = \frac{count(e,c)}{count(e)}$$

count(e,c) : number of alignment linking between a Chinese word *c* and an English word *e*

count(e): number of instances of e in alignment likings.

Let's take "pay" as an example. Table 6 shows the various alignment translations for "pay" and the translation probability.

Table 6 The aligned translations for the English word "pay" and their translation probability

Translation	Count	Translation Prob.	Translation	Count	Translation Prob.
代價	34	0.1214	花錢	7	0.025
錢	31	0.1107	出錢	6	0.0214
費用	21	0.075	租	6	0.0214
付費	16	0.0571	發給	6	0.0214
領	16	0.0571	付出	5	0.0179
繳	16	0.0571	薪資	5	0.0179
支付	13	0.0464	付錢	4	0.0143
給	13	0.0464	加薪	4	0.0143
薪水	11	0.0393	. . .	...	...
負擔	9	0.0321	積欠	2	0.0071
費	9	0.0321	繳款	2	0.0071
給付	8	0.0286			

Before running CLASS, we obtained 10,290 English idioms, collocations, and phrases together with 14,945 Chinese translations in LDOCE. After part of speech tagging, we had 1,851 distinct English patterns, and 4326 Chinese patterns. To calculate the statistical association of within words in a monolingual collocation and across the bilingual collocations, we built N-grams for the SPC. There were 790,000 Chinese word bigram and 669,000 distinct English bigram. CLASS identified around 595,000 Chinese collocation candidates (184,000 distinct types), and 230,000 English collocation candidates (135,000 distinct types) in the process.

We selected 100 sentences to evaluate the performance. We focused on rigid lexical collocations. The average English sentence had 45.3 words, while the average Chinese sentence had 21.4 words. The two human judges both master student majoring in Foreign Languages identified the bilingual collocations in these sentences.

We then compared the bilingual collocations produced by CLASS against the answer keys. The evaluation indicates an average recall rate = 60.9 % and precision = 85.2 % (See Table 7).

Table 7 Experiment result of bilingual collocation extracted from Sinorama parallel Corpus

# keys	#answers	#hits	#errors	Recall	Precision
382	273	233	40	60.9%	85.2%

4. Discussions

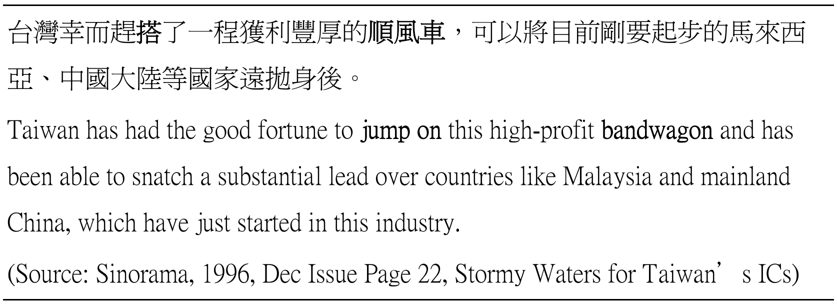
This paper describes a new approach to automatic acquisition of bilingual collocations from a parallel corpus. Our method is an extension of Melamed's Competitive Linking Algorithm for word alignment, combining both linguistic and statistical information for recognition of monolingual and bilingual collocations in a much simpler way than Smadja's work. We differ from previous work in the following ways:

1. We use a data-driven approach to extract monolingual collocations.
2. Unlike Smadja and Kupiec, we do not commit to two sets of monolingual collocations. Instead, we consider many overlapping and conflicting candidate and rely on the cross linguistic statistics to revolve the issue.
3. We combine both information related to the whole collocation as well as those of constituent words for more reliable probabilistic estimation of aligned collocations.

The approach is limited by its reliance on the training data of mostly rigid collocation patterns and is not applicable to elastic collocations such as "jump on …

bandwagon.” For instance, the program cannot handle the elastic collocation in following example:

Example 7



台灣幸而趕**搭**了一程獲利豐厚的**順風車**，可以將目前剛要起步的馬來西亞、中國大陸等國家遠拋身後。

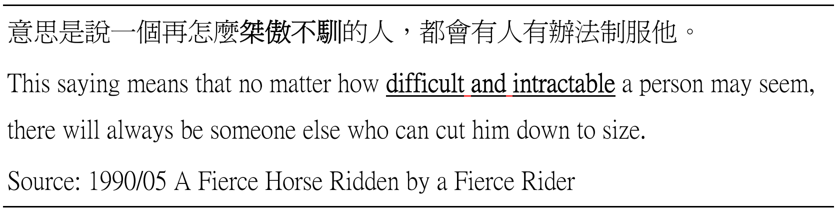
Taiwan has had the good fortune to **jump on** this high-profit **bandwagon** and has been able to snatch a substantial lead over countries like Malaysia and mainland China, which have just started in this industry.

(Source: Sinorama, 1996, Dec Issue Page 22, Stormy Waters for Taiwan’s ICs)

That limitation can be partially alleviated by matching nonconsecutive word sequence against existing lists of collocations for the two languages.

Another limitation has to do with bilingual collocations, which are not literal translations. For instance, “difficult and intractable” is not yet handled in the program, because it is not a word for word translation of “桀傲不馴”.

Example 8



意思是說一個再怎麼**桀傲不馴**的人，都會有人有辦法制服他。

This saying means that no matter how difficult and intractable a person may seem, there will always be someone else who can cut him down to size.

Source: 1990/05 A Fierce Horse Ridden by a Fierce Rider

In the experiment process, we found that the limitation may be partially solved by spliting the candidate list of bilingual collocations into two lists: one (NZ) with non-zero phrase translation probabilistic values and the other (ZE) with zero value. The two lists are then sorted by the LLR values. After extracting bilingual collocations from NZ list, we could continue to go downing the ZE list and select bilingual collocations if not conflicting with previously selection.

In the proposed method, we did no t take advantage of the correspondence of POS patterns from one language to the other. Some linking mistakes seem to be avoidable with the POS information. For example, the aligned collocation for "issue/**vb** visas/**nns**" is "簽證/**Na**", instead of "發/**VD** 簽證/**Na.**" However, the POS pattern "vb nn" appears to be more compatible with "VD Na" than "Na."

Example 9

一九七二年澳洲承認中共，中華民國即於此時與澳斷交。因爲無正式邦交，澳洲不能在台灣**發簽證**，而由澳洲駐香港的使館代辦，然後將簽證送回台灣，簽證手續約需五天至一周。

The Republic of China broke relations with Australia in 1972, after the country recognized the Chinese Communists, and because of the lack of formal diplomatic relations, Australia felt it could not **issue visas** on Taiwan. Instead, they were handled through its consulate in Hong Kong and then sent back to Taiwan, the entire process requiring five days to a week to complete.

Source: 1990/04 Visas for Australia to Be Processed in Just 24 Hours

A number of mistakes are caused with the erroneous word segments process of the Chinese tagger. For instance, "大學及研究生出國期間" should be segmented as "大學 / 及 / 研究生 / 出國 / 期間" but instead segment was "大學 / 及 / 研究 / 生出 / 國 / 期間 / 的 / 學業." Another major source of segmentation mistakes has to do with proper names and their transliterations. These name entities that are not included in the database are usually segmented into single Chinese character. For instance, "...一書作者劉學銚指出..." is segmented as " ... / 一 / 書 / 作者 / 劉 / 學 / 銚 / 指出 / ...," while "...在匈牙利地區建國的馬札爾人..." is segmented as "...在 / 匈牙利 / 地區 / 建國 / 的 / 馬 / 札 / 爾 / 人 / ...." Therefore, handling these name entities in a pre-process should be helpful to avoid segment mistakes, and alignment difficulties.

5. Conclusion and Future Work

In this paper, we describe an algorithm that employs syntactic and statistical analyses to extract rigid bilingual collocations from a parallel corpus. Phrases matching the preferred patterns are extract from aligned sentences in a parallel corpus. Those phrases are subsequently matched up via cross-linguistic statistical association. Statistical association between the whole collocations as well as words in collocations is used jointly to link a collocation and its counterpart. We experimented with an implementation of the proposed method on a very large Chinese-English parallel corpus with satisfactory results.

A number of interesting future directions suggest themselves. First, it would be interesting to see how effectively we can extend the method to longer and elastic collocations and to grammatical collocations. Second, bilingual collocations that are proper names and transliterations may need additional considerations. Third, it will be interesting to see if the performance can re improved cross language correspondence between POS patterns.

References

1. Benson, Morton., Evelyn Benson, and Robert Ilson. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam, Netherlands, 1986.
2. Choueka, Y. (1988) : "Looking for needles in a haystack", Actes RIAO, Conference on User-Oriented Context Based Text and Image Handling, Cambridge, p. 609-623.
3. Choueka, Y.; Klein, and Neuwitz, E.. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, 4(1):34-8, (1983)
4. Church, K. W. and Hanks, P. Word association norms, mutual information, and lexicography. Computational Linguistics, 1990, 16(1), pp. 22-29.
5. Dagan, I. and K. Church. Termight: Identifying and translation technical terminology. In *Proc. of the 4th Conference on Applied Natural Language Processing (ANLP)*, pages 34-40, Stuttgart, Germany, 1994.
6. Dunning, T (1993) Accurate methods for the statistics of surprise and coincidence, Computational Linguistics 19:1, 61-75.
7. Haruno, M., S. Ikehara, and T. Yamazaki. Learning bilingual collocations by word-level sorting. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, Copenhagen, Denmark, 1996.
8. Huang, C.-R., K.-J. Chen, Y.-Y. Yang, Character-based Collocation for Mandarin Chinese, In ACL 2000, 540-543.
9. Inkpen, Diana Zaiu and Hirst, Graeme. ``Acquiring collocations for lexical choice between near-synonyms." SIGLEX Workshop on Unsupervised Lexical Acquisition, 40th meeting of the Association for Computational Lin
10. Justeson, J.S. and Slava M. Katz (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1):9-27.
11. Kupiec, Julian. An algorithm for finding noun phrase correspondences in bilingual corpora. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993.
12. Lin, D. Using collocation statistics in information extraction. In *Proc. of the Seventh Message Understanding Conference (MUC-7)*, 1998.

13. Manning and H. Schutze. Foundations of Statistical Natural Language Processing (SNLP), C., MIT Press, 1999.

14. Melamed, I. Dan. "A Word-to-Word Model of Translational Equivalence". In Procs. of the ACL97. pp 490-497. Madrid Spain, 1997.

15. Smadja, F. 1993. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143-177

16. Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1-38, 1996.

17. Kevin C. Yeh, Thomas C. Chuang, Jason S. Chang (2003), Using Punctuations for Bilingual Sentence Alignment- Preparing Parallel Corpus

LiveTrans: Translation Suggestion for Cross-Language Web Search from Web Anchor Texts and Search Results

Wen-Hsiang Lu[1,2], Lee-Feng Chien[1] and Hsi-Jian Lee[2]

1. Institute of Information Science, Academia Sinica, Taiwan, ROC

2. Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, ROC

{whlu, lfchien}@iis.sinica.edu.tw, {whlu, hjlee}@csie.nctu.edu.tw

Abstract

In this paper we will present a system, called LiveTrans, which can generate translation suggestions for given user queries and provide an English-Chinese cross-language search service for the retrieval of both Web pages and images. The system effectively utilizes two kinds of Web resources: anchor texts and search results. The developed anchor-text-based and search-result-based methods are complementary in the precision and coverage rates and promising in extracting translations of unknown query terms that were not included in general-purpose translation dictionaries. Experimental results demonstrate the feasibility of the system.

1. Introduction

To deal with automatic construction of translation lexicons, conventional research on machine translation (MT) [3] and cross-language information retrieval (CLIR) [1, 5, 7, 10, 13, 18] has generally used statistical techniques to automatically extract word translations from domain-specific parallel/comparable bilingual texts, such as bilingual newspapers [4, 11, 12, 20, 21]. However, only a certain set of their translations can be extracted through corpora with limited domains. In our research, we are interested in extracting translations of technical terms and proper names in diverse subjects, which are especially needed in performing CLIR services for Web users, e.g., "Hussein" (海珊/哈珊/侯賽因), "SARS" (嚴重急性呼吸道症候群). Existing CLIR systems usually rely on bilingual dictionaries for query translation [1, 13, 15]. Unfortunately, our analysis of Dreamer query log collected in Taiwan (see Section 3.1) showed that 74% of the 20,000 high frequent Web queries can not be found in general-purpose English-Chinese dictionaries (they are called *unknown terms* in this paper). How to automatically find translations for unknown terms, therefore, has become a major challenge for cross-language Web search.

Different from previous works, we focus on investigating new approaches to mining multilingual Web resources [19]. We have proposed a novel approach to extracting translations of Web queries through the mining of Web anchor texts and link structures [16, 17]. An anchor text is the descriptive part of an out-link of a Web page used to provide a brief description of the linked page. A variety of anchor texts in multiple languages might link to the same pages from all over the world. For example, Figure 1 shows a typical example, in which there are a variety of anchor texts in multiple languages linking to the Yahoo! from all over the world. Such a bundle of anchor texts pointing together to the same page is called an *anchor-text set*. Web anchor-text sets may contain similar description texts in multiple languages. Thus, for an unknown term appearing in some anchor-text sets, it is likely that its corresponding target translations appear together in the same anchor-text sets.

However, discovering translation knowledge from the Web has not been fully explored. In this paper, we intend to investigate another kind of Web resource, *search results*, and try to combine them with the anchor texts to benefit term translation. Chinese pages on the Web consist of rich texts in a mixture of Chinese (main language) and English (auxiliary language), and many of them contain translations of proper nouns. According to our observations, many search result pages in Chinese Web usually contain snippets of summaries in a mixture of Chinese and English. For example, Figure 2 illustrates the search-result page of the English query "National Palace Museum," which was submitted to Google for searching Chinese pages, could obtain many relevant results containing both the query itself and its Chinese aliases. To explore search results on extraction of term translation, we have employed two methods: the chi-square test and context-vector analysis.

Based on a novel integration of the developed anchor-text- and search-result-based methods, we implemented an experimental system, called LiveTrans, to provide English-Chinese translation suggestion and cross-lingual retrieval of both Web pages and images. The purpose of this paper is to introduce our experiences in developing the methods and implementing the system.

2. Related Work

Term translation extraction is an important research problem in the context of MT. A number of related researches [12, 21] have used sentence-aligned parallel corpora to extract translations since the advent of statistical translation model [3]. Although high accuracy can be easily achieved by these techniques, sufficiently large parallel corpora for various subject domains and language-pairs are still hard to be available. On the other hand, some work has been done on term translation extraction from comparable or even unrelated texts [11, 20]. However, using non-parallel corpora is more difficult to effectively extract translations than parallel corpora due to the lack of parallel correlation aligned between documents or sentence pairs.

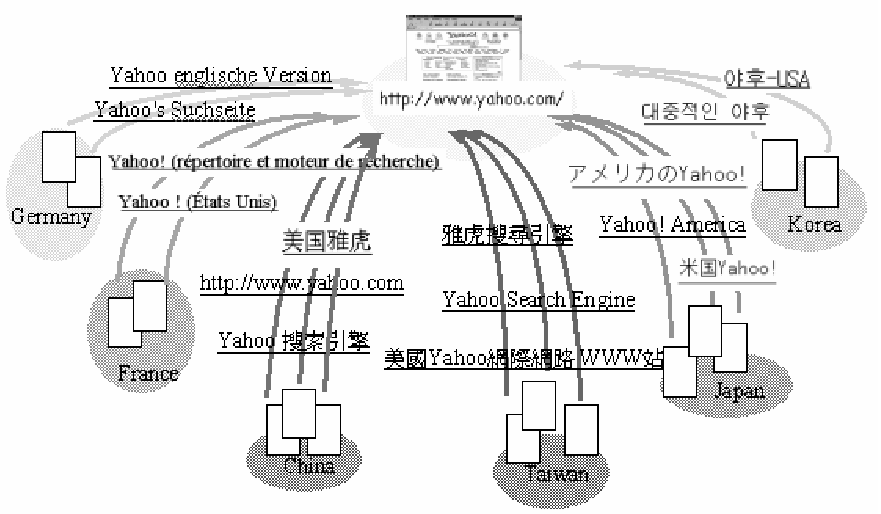
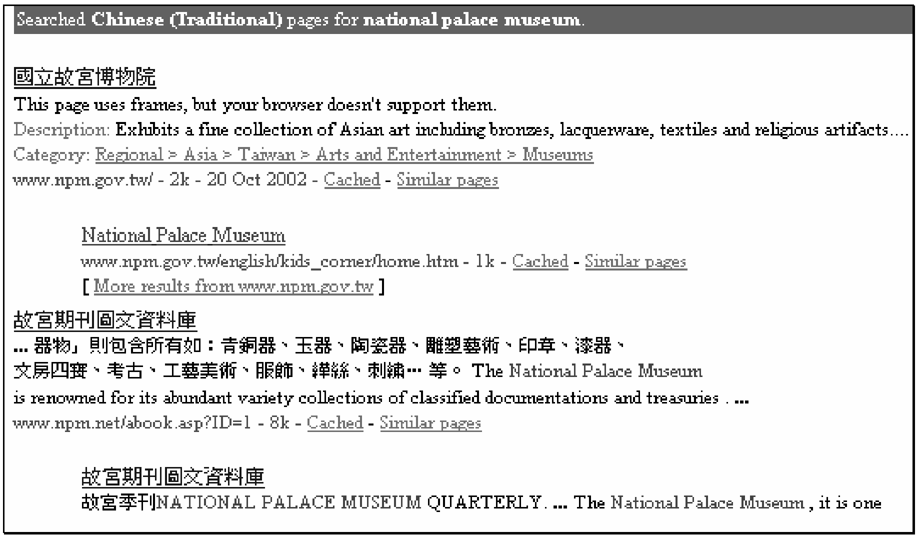


Figure 1. An illustration showing various anchor texts in multiple languages linking to Yahoo! from all over the world [17].



Searched **Chinese (Traditional)** pages for **national palace museum**.

國立故宮博物院
This page uses frames, but your browser doesn't support them.
Description: Exhibits a fine collection of Asian art including bronzes, lacquerware, textiles and religious artifacts....
Category: Regional > Asia > Taiwan > Arts and Entertainment > Museums
www.npm.gov.tw/ - 2k - 20 Oct 2002 - Cached - Similar pages

National Palace Museum
www.npm.gov.tw/english/kids_corner/home.htm - 1k - Cached - Similar pages
[More results from www.npm.gov.tw]

故宮期刊圖文資料庫
... 器物」則包含所有如：青銅器、玉器、陶瓷器、雕塑藝術、印章、漆器、文房四寶、考古、工藝美術、服飾、緙絲、刺繡… 等。 The National Palace Museum is renowned for its abundant variety collections of classified documentations and treasuries . ...
www.npm.net/abook.asp?ID=1 - 8k - Cached - Similar pages

故宮期刊圖文資料庫
故宮季刊NATIONAL PALACE MUSEUM QUARTERLY. ... The National Palace Museum , it is one

Figure 2. An illustration showing translation equivalents, such as "National Palace Museum"/國立故宮博物院 (故宮), which are included in a search result page returned from Google.

On the other hand, CLIR has become an important topic in recent research on information retrieval, however, practical cross-language Web search services have not lived up to expectations. This task must face a number of challenges, especially the problem of query translation. To deal with such problem, existing CLIR systems mostly rely on bilingual dictionaries. These dictionary-based techniques are limited in real-world applications since queries often contain unknown query terms, such as personnel names and technical terms [15]. Although some methods integrating dictionary-based techniques with parallel-corpus disambiguation, technology have been proposed and achieved performance improvements [1, 13]. Nonetheless, the unavailability of translations of unknown Web queries in diverse subjects is still a thorny problem.

A page[1] modified by Oard lists some CLIR retrieval systems, which can be either used on the Internet or obtained from commercial sources. For example, the Multilingual Summarization and Translation (MuST[2]) system is a Web-accessible CLIR system that uses English queries to search Indonesian, Spanish, Arabic and Japanese. MTIR is a demonstration search system that accepts queries in Chinese, finds documents in English, and then translates the selected documents into Chinese [1]. These systems generally rely on built-in bilingual dictionaries for query translation. To our knowledge, the proposed LiveTrans system is one of the few CLIR systems which allow the translations of unknown queries to be extracted through the mining of Web resources.

3. LiveTrans System

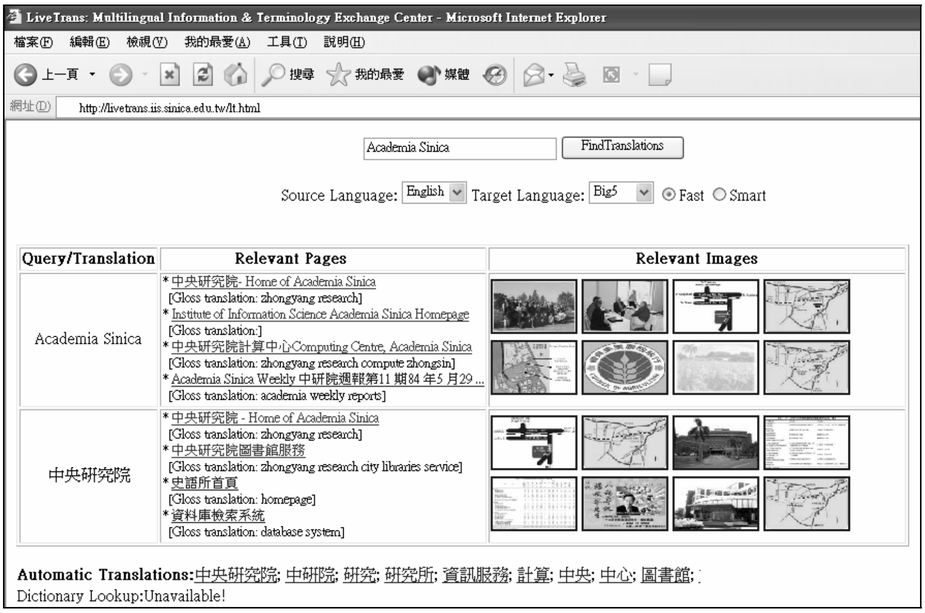
The LiveTrans[3] system is an experimental meta-search engine that provides English-Chinese translation suggestion and cross-language search for retrieval of both Web pages and images. It was implemented based on a novel combination of the developed Web mining methods. To use the system, users may select either English, traditional Chinese or simplified Chinese as the source/target language. For each input source query, the system will suggest a list of target translations. Since real queries are often short, there is a lack of context information needed to perform query translation. The system combines the term translation extraction methods and bilingual lexicons to make suggestions. The users can select the preferred translation and the system will return the retrieved Web pages and images, and sort them in their order of decreasing relevance to the corresponding translated queries. The titles of the retrieved pages are also translated word by word to the source languages for reference (i.e. gloss translation). Like most of the meta-search engines, backend engines can be chosen and the retrieved results can be merged using a data fusion technique. The system has been used to collect translation equivalents of a certain portion of users' queries. Many of the obtained translations are really not easy for human indexers to compile. For example, in the case shown in Figure 3, the user selected English as the source language and Chinese as the target language. In this example, the given query was "Academia Sinica" and its translations were extracted, i.e., 中央研究院 and 中研院.

We sometimes refer to the Web as a globally interconnected information infrastructure. At present, however, for someone who reads only English, it is presently the English-Wide-Web, and a reader of only Chinese sees only the Chinese-Wide-Web. With the LiveTrans system, it is easy to see that there are a number

[1] http://raveb.umd.edu/ddlrg/clir/systems.html

[2] http://www.isi.edu/natural-language/projects/C-ST-RD.html

[3] http://livetrans.iis.sinica.edu.tw/lt.html


LiveTrans: Multilingual Information & Terminology Exchange Center - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
媒體
網址(D)
FindTranslations

http://livetrans.iis.sinica.edu.tw/lt.html

Academia Sinica

Source Language: English Target Language: Big5 ● Fast ○ Smart

Query/Translation	Relevant Pages	Relevant Images
Academia Sinica	* 中央研究院- Home of Academia Sinica [Gloss translation: zhongyang research] * Institute of Information Science Academia Sinica Homepage [Gloss translation:] * 中央研究院計算中心Computing Centre, Academia Sinica [Gloss translation: zhongyang research compute zhongsin] * Academia Sinica Weekly 中研院週報第11 期84 年5 月29 ... [Gloss translation: academia weekly reports]	
中央研究院	* 中央研究院 - Home of Academia Sinica [Gloss translation: zhongyang research] * 中央研究院圖書館服務 [Gloss translation: zhongyang research city libraries service] * 史語所首頁 [Gloss translation: homepage] * 資料庫檢索系統 [Gloss translation: database system]	

Automatic Translations:中央研究院; 中研院; 研究; 研究所; 資訊服務; 計算; 中央; 中心; 圖書館;

Dictionary Lookup:Unavailable!

Figure 3. An example showing the search results retrieved by the LiveTrans system, where the given query was "Academia Sinica" and its translations extracted were 中央研究院, 中研院.

of cases where Chinese users need English-Chinese cross-language translation. In fact, the LiveTrans system was found to be effective in increasing the recall rate of Web search, especially for the retrieval of Web images. Requests for images often are not limited to the local environment. For example, for the original query 羅浮宮 (Louvre) in Chinese, it could retrieve only hundreds of Web images, but it could retrieve hundreds of thousands images through its English translation.

With the novel combination of the developed Web mining methods (see Section 4), the LiveTrans system could provide effective translation suggestions for users selecting the 'Smart' mode; however, it cannot perform efficiently in real time due to its computation complexity. To obtain query translation instantly, the user is recommended selecting the 'Fast' mode with a little loss of accuracy. To remain the accuracy, the system can constantly update translations for new queries in the query log in a batch. Therefore, the system can effectively provide translation suggestions and cross-lingual search services.

4. Query Translation from Anchor Texts and Search Results

To implement a query translation process via mining the Web resources: anchor texts and search results, three major processing steps are required:

(1) Corpus collection: Collect bilingual Web data as a comparable corpus.

(2) Translation candidate extraction: Extract translation candidates from the collected corpus.

(3) Translation selection: Estimate the similarity for each candidate and determine the most possible translations.

To effectively handle this process, we have developed two kinds of methods: the anchor-text-based method and the search-result-based method. The details regarding the two methods will be presented in the following.

4.1 The Anchor-Text-Based Method

Query translation from anchor texts contains three major computational modules: anchor-text extraction, translation candidate extraction, and translation selection. The anchor-text extraction module was constructed to collect pages from the Web and build up a corpus of anchor-text sets. For each given query term, the translation candidate extraction module extracts key terms in the target language as the translation candidates from the anchor-text sets containing the query term. The effectiveness of the adopted term extraction methods greatly affects the performance in extracting correct translations. Three different methods have been tested in our previous work [17]: the PAT-tree-based (a statistics-based n-gram model [9]), query-set-based and tagger-based methods. Among them, the query-set-based method has been adopted in this paper because it could extract longer terms (i.e. multi-words) and have less problems of Chinese term segmentation than the other methods. This method uses query logs in the target language as the translation vocabulary set to segment anchor texts and extract key terms. The pre-condition for using this method is that the coverage of the query set should be high. Finally, the translation selection module selects the possible translation that maximizes the estimation based on the probabilistic inference model described below.

4.1.1 The Probabilistic Inference Model

To find the most probable translation t for a query term s, we have proposed probabilistic inference model to utilize Web anchor texts and hyperlink structures. This model is used to estimate the probability value between a query term and each translation candidate that co-occurs with the query term in the same anchor-text sets. The estimation assumes that anchor texts linking to the same pages may contain similar terms with analogous concepts. Therefore, a candidate has a higher chance of being an correct translation if it is written in the target language and frequently co-occurs with the query term in the same anchor-text sets. In addition, in the field of Web research, it has been proven that link structures can be used effectively to estimate the authority of Web pages [2, 14]. Our model further assumes that the translation candidates in the anchor-text sets of pages with higher authority may be more reliable. For a Web page (or URL) u_i, its anchor-text set $AT(u_i)$ is defined as consisting of all of the anchor texts of the links pointing to u_i, i.e., u_i 's in-links.

The similarity estimation function based on the probabilistic inference model is called model S_{AT} for the sake of usage consistency in the consequent sections and is defined below:

$$S_{AT}(s,t) = P(s \leftrightarrow t) = \frac{P(s \cap t)}{P(s \cup t)} = \frac{\sum_{i=1}^{n} P(s \cap t \cap u_i)}{\sum_{i=1}^{n} P((s \cup t) \cap u_i)} = \frac{\sum_{i=1}^{n} P(s \cap t \mid u_i)P(u_i)}{\sum_{i=1}^{n} P(s \cup t \mid u_i)P(u_i)}. \quad (1)$$

The above measure is adopted to estimate the degree of similarity between source term *s* and target translation *t*. The measure is estimated based on their co-occurrence in the anchor text sets of the concerned Web pages **U** = $\{u_1, u_2, \ldots u_n\}$, in which u_i is a page of concern and $P(u_i)$ is the probability value used to measure the authority of page u_i. By considering the link structures and concept space of Web pages, $P(u_i)$ is estimated along with the probability of u_i being linked, and its estimation is defined as follows: $P(u_i)= L(u_i)/\Sigma_{j=1,n}\ L(u_j)$, where $L(u_j)$ indicates the number of in-links of page u_j.

In addition, we assume that *s* and *t* are independent given u_i; then, the joint probability $P(s \cap t/u_i)$ is equal to the product of $P(s/u_i)$ and $P(t/u_i)$, and the similarity measure becomes

$$S_{AT}(s,t) \approx \frac{\sum_{i=1}^{n} P(s \mid u_i)P(t \mid u_i)P(u_i)}{\sum_{i=1}^{n} [P(s \mid u_i) + P(t \mid u_i) - P(s \mid u_i)P(t \mid u_i)]P(u_i)}. \quad (2)$$

The values of $P(s/u_i)$ and $P(t/u_i)$ are estimated by calculating the fractions of the numbers of u_i's in-links containing *s* and *t* over $L(u_i)$, respectively. Therefore, a candidate translation has a higher confidence value for being an effective translation if it frequently co-occurs with the source term in the anchor-text sets of those pages having higher authority. For details about the probabilistic inference model, readers may refer to our previous work [17].

4.2 The Search-Result-Based Method

Query translation from search results also contains three major computational modules: search-result collection, translation candidate extraction, and translation selection. In the search-result collection module, a given source query is submitted to a real-world search engine to collect the top search result pages. In the translation candidate extraction module, we use the same term extraction method adopted in the anchor-text-based method. In the translation selection module, our idea is to utilize co-occurrence and context information between source queries and target translation candidates to estimate their semantic similarity and to determine the most possible translations. We have investigated several different methods of estimation and found that the chi-square test and context vector analysis achieve better performance.

4.2.1 The Chi-Square Test

A number of statistical measures have been proposed for estimating the association between words/phrases based on co-occurrence analysis, including mutual information, the DICE coefficient, and statistical tests, such as the chi-square test and the log-likelihood ratio test [12, 20, 21]. Although the log-likelihood ratio test is suitable for dealing with the data sparseness problem, in our preliminary experiments on 430 popular Web queries (see Section 5.1), we found that the chi-square test performs better than the log-likelihood ratio test. One of the possible reasons is that the required parameters for the chi-square test can be effectively obtained from real-world search engines, and is enough to avoid the data sparseness problem. The chi-square test was, therefore, adopted as the major method for co-occurrence analysis in our work. Its similarity measure is defined as

$$S_{X}2(s,t) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)}, \; (3)$$

where a, b, c and d are the numbers in the four cells of the contingency table (see Table 1) for the source term *s* and target term *t* and are defined as follows:

a: the number of pages containing both terms *s* and *t*;

b: the number of pages containing term *s* but not *t*;

c: the number of pages containing term *t* but not *s*;

d: the number of pages containing neither term *s* nor *t*;

N: the total number of pages, i.e., *N*= *a*+*b*+*c*+*d*.

Table 1. A contingency table.

	t	*~t*
s	*a*	*b*
~s	*c*	*d*

The required parameters for the chi-square test can be computed using the search results returned from real-world search engines. Most search engines accept Boolean queries and can report the number of pages matched.

4.2.2 The Context-Vector Analysis

Co-occurrence analysis is applicable to frequent query terms because these terms are more likely to appear with their translation candidates. On the other hand, infrequent query terms have little chance of appearing with translation candidates in the same pages. The context-vector-based method has been used to extract translations from comparable corpora [11, 20], and is thus adopted to deal with this problem. Different from previous works using a translation lexicon to bridge the features with the same meaning in different languages, we use only popular query terms as the feature set, because of the advantage of updating the feature set with queries in diverse subjects continuously supplied by Web users. This is a suitable way to provide effective feature sets to represent context vectors of diverse unknown query terms and their translation candidates. For each query or candidate term, we take the co-occurring feature terms as its context vector since translation equivalents may share the same occurring feature terms. The similarity between a query term and each translation candidate can be computed based on their context vectors. Thus, infrequent query terms still have a chance of extracting translations.

Like Fung et al.'s vector space model, we also use the TF-IDF weighting scheme to estimate the significance of each feature in the context vector and use the cosine measure to calculate the translation similarity of each query term and its translation candidates. The weighting scheme is defined as follows:

$$w_{t_i} = \frac{f(t_i, d)}{\max_j f(t_j, d)} \times \log(\frac{N}{\mathrm{n}}), (4)$$

where $f(t_i,d)$ is the frequency of t_i in search result page d, N is the total number of Web pages in the collection of search engines, and n is the number of pages including t_i.

Given the context vectors of a query term and each translation candidate, their similarity measure is estimated as follows:

$$S_{CV}(s,t) = \frac{\sum_{i=1}^{m} w_{s_i} \times w_{t_i}}{\sqrt{\sum_{i=1}^{m} (w_{s_i})^2 \times \sum_{i=1}^{m} (w_{t_i})^2}}. (5)$$

It is not difficult to construct context vectors for query terms and their translation candidates. For a query term, we can obtain search results by submitting it as a query to real-world search engines. Basically, we can use a fixed number of the top retrieved results (snippets) to extract translation candidates. The co-occurring feature terms of each query can also be extracted, and their weights calculated based on the retrieved snippets. The context vector of the query is, thus, constructed. The same procedure is used to construct a context vector for each translation candidate.

4.3 The Combined Method

Our previous experiments show that the anchor-text-based method can achieve a good precision rate for popular Web queries in other language pairs besides Chinese and English [17], but it has a major drawback; that is, the cost is relatively high to collect sufficient pages to extract anchor texts. Benefiting from real-world search engines, the search-result-based method can achieve a good coverage rate for diverse query terms. However, method using the chi-square test has difficulty in dealing with infrequent query terms, and the method using context-vector analysis needs to carefully handle the issue of feature selection. Intuitively, a more complete solution is to integrate the three different methods. Under consideration of the large difference of ranges of similarity values among the three methods, we use a linear combination weighting scheme to compute the similarity measure as follows:

$$S_{COMBINED}(s,t)=\sum_{m}\frac{\alpha_m}{R_m(s,t)}, \quad (6)$$

where α_m is an assigned weight for each similarity measure S_m, and $R_m(s,t)$, which represents the similarity ranking of each translation candidate t with respect to the source term s, is assigned to be from 1 to k (candidate number) in decreasing order by similarity measure $S_m(s,t)$.

5. Experimental Results

5.1 The Test Bed

To determine the effectiveness of the developed methods to Web query translation, we conducted several experiments on extracting English translations for Chinese queries. We collected real query terms along with the logs from two real-world Chinese search engines in Taiwan, i.e., Dreamer and GAIS. The Dreamer log contained 228,566 unique query terms from a period of over 3 months in 1998, and the GAIS log contained 114,182 unique query terms from a period of two weeks in 1999. A query set, called the *popular-query set*, was prepared to test the translation effectiveness for unknown Web queries. There were 9,709 most popular query terms whose frequencies were above 10 in the two logs, and 1,230 of them were English terms. After checking the logs, we obtained 430 terms whose Chinese translations appeared together in the logs and took their Chinese translations as the popular-query set. Table 2 lists some examples of the test query terms, which were divided into two types, where type Dic (the terms existing in the dictionary) made up about 36% (156/430) of the test queries, and type OOV (out of vocabulary; the terms not in the dictionary) made up about 64% (274/430).

In addition, to further investigate the translation effectiveness for proper names and technical terms, we also prepared two different query sets containing 50 scientist names and 50 disease names in English, which were randomly selected from the 256 scientists (Science/People) and 664 diseases (Health/Diseases and Conditions) in the Yahoo! Directory, respectively. It should be noted that 76% (38/50) scientist names and 72% (36/50) disease names are not included in the general-purpose translation dictionary which contains 202,974 entries collected from the Internet.

Table 2. Some sample test queries.

Type	Number	Sample test queries
Dic	156	銀行 (bank) 亞洲 (Asia) 愛滋病 (AIDS) 白宮 (White House) 世界貿易組織 (WTO)
OOV	274	電子商務 (E-commerce) 個人數位助理(PDA) 雅虎 (Yahoo) 太空總署 (NASA) 星際大戰 (Star War)

5.2 Web Data Collection

We had collected 1,980,816 traditional Chinese Web pages in Taiwan and then extracted 109,416 pages (URLs), whose anchor-text sets contained both traditional Chinese and English terms, and which were taken as the anchor-text-set corpus for testing the anchor-text-based method. In addition, for testing the search-result-based method, we obtained search results of queries by submitting them to real-world Chinese search engines, such as Google Chinese[4] and Openfind[5]. Basically, we used only the first 100 retrieved results (snippets) to extract translation candidates. The context vector of each query was also extracted from the snippets. Also, the required parameters for the chi-square test were computed using the search results returned from the utilized search engines.

5.3 Performance of the Proposed Methods for Popular Query Terms

We carried out experiments to determine the performance of the proposed methods in extracting translations for the bilingual query set. To evaluate the performance of translation extraction, we used the *average top-n inclusion rate* as a metric. For a set of test queries, its top-n inclusion rate was defined as the percentage of queries whose effective translations could be found in the first *n* extracted translations. Also, we wished to know if the coverage of effective translations was high enough in the top search result pages for the real queries. The coverage rate was the percentage of queries whose effective translations could be found in the extracted translation candidate set.

Table 3 shows the obtained results in terms of top 1-5 inclusion rates and coverage rate. In this table, CV, χ^2, AT and Combined represent the context-vector analysis, chi-square test, anchor-text-based, and combined methods, respectively. In addition, Dic, OOV and All represent the terms existing in a dictionary, the terms not

[4] http://www.google.com/

[5] http://www.openfind.com/

in a dictionary, and the total query set, respectively. It is clear that the AT method and the combined method performed better than the χ^2 and CV methods in almost every case. The weights of the combined method were assigned according to the top-1 inclusion rates achieved by the three other methods, i.e., α_{cv} = 56.3%/(56.3%+49.5%+66.5%) ≈ 0.33. In fact, the obtained coverage rates were very high. This shows that the Chinese Web is rich in texts with a mixture of Chinese and English.

Table 3. Coverage and inclusion rates for popular Chinese queries using the different methods.

Method	Query Type	Top-1	Top-3	Top-5	Coverage
CV	Dic	56.4%	70.5%	74.4%	80.1%
	OOV	56.2%	66.1%	69.3%	85.0%
	All	56.3%	67.7%	71.2%	83.3%
χ^2	Dic	40.4%	61.5%	67.9%	80.1%
	OOV	54.7%	65.0%	68.2%	85.0%
	All	49.5%	63.7%	68.1%	83.3%
AT	Dic	67.3%	78.2%	80.8%	89.1%
	OOV	66.1%	74.5%	76.6%	83.9%
	All	66.5%	75.8%	78.1%	85.8%
Combined	Dic	68.6%	82.1%	84.6%	92.3%
	OOV	66.8%	85.8%	88.0%	94.2%
	All	67.4%	84.4%	86.7%	93.5%

Table 4. Coverage and inclusion rates for popular English queries using the different methods.

Method	Top-1	Top-3	Top-5	Coverage
CV	50.9%	60.1%	60.8%	80.9%
χ^2	44.6%	56.1%	59.2%	80.9%
AT	57.1%	70.0%	71.9%	85.4%
Combined	59.4%	74.3%	76.2%	89.9%

The above popular-query set contained only Chinese queries. To determine the performance of the proposed methods in translating English queries into Chinese, we carried out another experiment which used the English translations of the same popular-query set as the test set. The results are shown in Table 4. The achieved performance was a little worse than achieved using the Chinese query set. The reason for this result was that the English queries had to deal with more ambiguous Chinese translation candidates since the search result pages returned from Chinese search engines normally contain mostly Chinese texts.

5.4 Performance of the Combined Method for Proper Names and Technical Terms

To further deal with the translation of proper names and technical terms, we conducted an experiment on the test sets of scientist names and medical terms mentioned in Section 5.1. According to our analysis of the test terms, many of scientist and disease names were not included in our collected query-log set, and some disease names were multi-words, e.g., "Hypoplastic Left Heart Syndrome" (左心發育不全症候群), "Lactose Intolerance" (乳糖不耐症), "Nosocomial Infections" (院內感染). Thus, we slightly modified the method of query-set-based translation candidate extraction by augmenting a simplified technique of unknown term and multi-word identification [6, 8]. As a result, the top-1 inclusion rate was obtained at 40% and 44% for the scientist and disease names, respectively (see Table 5). Some examples of the correct translations extracted using the combined method are shown in Table 6.

Table 5. Inclusion rates for proper names and technical terms using the combined method.

Query Type	Top-1	Top-3	Top-5
Scientist Name	40.0%	52.0%	60.0%
Disease Name	44.0%	60.0%	70.0%

5.5 Discussion

The translation accuracy achieved using the combined method is very promising, especially for popular queries. According to our analysis, this good performance was primarily due to the fact that the Chinese Web has a mixed language characteristic: many pages mainly consist of texts in Chinese (main language) with parts of texts in English (auxiliary language). The Chinese Web is considerably rich in texts containing English-Chinese translations of proper nouns, such as personal names and technical terms. As a result, this characteristic makes it possible to automatically extract English-Chinese translations of a large number of unknown query terms.

In fact, the translation process based on the search-result-based method might not be very effective for language pairs that do not exhibit the language-mixed characteristic on the Web. For this reason, the anchor-text-based method is still attractive while it achieves good precision rates for popular queries in other language pairs besides Chinese and English, even though not every particular pair of languages has sufficient texts on the Web.

The performance achieved using the combined method looks very promising, but it still has limitations. For example, it is less reliable in extracting translations of multi-word terms. To enhance the accuracy in translating multi-word or unknown terms, it should be worthy to employ more effective techniques, such as word segmentation and language model, to filter out noise terms and extract complete translation candidates. Currently, the LiveTrans system cannot perform efficiently in real time due to its computation complexity. This is a real challenge to improve the response time of query translation in our future work. However, the system can constantly update translations for new queries in the query log in a batch. Therefore, the system still can provide translation suggestions and cross-lingual search services.

Table 6. Some examples of the test proper names and technical terms, and their extracted translations.

Query Type	English Query	Extracted Chinese Translations
Scientist Name	Aldrin, Buzz (Astronaut)	艾德林
	Hadfield, Chris (Astronaut)	哈德菲爾德
	Galilei, Galileo (Astronomer)	伽利略/伽里略/加利略
	Ptolemy, Claudius (Astronomer)	托勒密
	Earhart, Amelia (Aviators)	鄂哈特
	Tibbets, Paul (Aviators)	第貝茲/迪貝茨
	Crick, Francis (Biologists)	克立克/克里克
	Drake, Edwin Laurentine (Earth Scientist)	德拉克
	Aryabhata (Mathematician)	阿耶波多/阿利耶波多
	Kepler, Johannes (Mathematician)	克卜勒/開普勒/刻卜勒
	Dalton, John (Physicist)	道爾頓/道耳吞/道耳頓
	Feynman, Richard (Physicist)	費曼
Disease Name	Ganglion Cyst	腱鞘囊腫
	Gestational Diabetes	妊娠糖尿病
	Hypoplastic Left Heart Syndrome	左心發育不全症候群
	Lactose Intolerance	乳糖不耐症
	Legionnaires' Disease	退伍軍人症
	Muscular Dystrophy	肌肉萎縮症
	Nosocomial Infections	院內感染
	Shingles	帶狀皰疹/帶狀庖疹
	Stockholm Syndrome	斯德哥爾摩症候群
	Sudden Infant Death Syndrome (SIDS)	嬰兒猝死症

6. Conclusion

Practical cross-language Web search services have not lived up to expectations since they suffer from a major problem where up-to-date multilingual lexicons containing the translations of popular Web queries, such as proper names and technical terms, are lacking. In this paper we present a promising system, called LiveTrans, which can generate translation suggestions for given user queries and provide an English-Chinese cross-language search service for the retrieval of both Web pages and images. The system effectively utilizes two kinds of live Web resources: anchor texts and search results, which are contributed continuously by a huge number of volunteers (page authors) around the world. The developed anchor-text-based and search-result-based methods are complementary in the precision and coverage rates and promising in extracting translations of query terms that were not included in general-purpose translation dictionaries.

References

[1] Bian, G. W. and Chen, H. H. (2000) Cross-Language Information Access to Multilingual Collections on the Internet, Journal of the American Society for Information Science, 51(3), 281-296.

[2] Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proceedings of the 7th International World Wide Web Conference, 107-117.

[3] Brown, P., Pietra, S. A. D., Pietra, V. D. J., Mercer, R. L. (1993) The Mathematics of Machine Translation, Computational Linguistics, 19(2), 263-312.

[4] Chang, J. S., Yu, D., Lee, C. J. (2001) Statistical Translation Model for Phrases, Computational Linguistics and Chinese Language Processing, 6(2), 43-64.

[5] Chang, J. S., Ker, S. J. and Chen, M. H. (1998) Cross Language Information Retrieval and Data Mining, Proceedings of the Conference on Information Science and Technology-1998: Perspectives in the 21st Century, 153-166.

[6] Chang, J. S. and Su, K. Y. (1997) A Multivariate Gaussian Mixture Model for Automatic Compound Word Extraction, Proceeding of ROCLING X, 123-142.

[7] Chen, K. H. and Chen, H. H. (2001) The Chinese Text Retrieval Tasks of NTCIR Workshop 2, Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization.

[8] Chen, K. J. and Bai, M. H. (1998) Unknown Word Detection for Chinese by a Corpus-Based Learning Method, International Journal of Computational Linguistics and Chinese Language Processing, 3(1), 27-44.

[9] Chien, L. F. (1997) PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval, Proceedings of ACM-SIGIR '97, 50-59.

[10] Dumais, S. T., Landauer, T. K., Littman, M. L. (1996) Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing, Proceedings of ACM-SIGIR'96 Workshop on Cross-Linguistic Information Retrieval, 16-24.

[11] Fung, P. and Yee, L. Y. (1998) An IR Approach for Translating New Words from Nonparallel, Comparable Texts, Proceedings of the 36th Annual Conference of the Association for Computational Linguistics, 414-420.

[12] Gale, W. A. and Church, K. W. (1991) Identifying Word Correspondances in Parallel Texts, Proceedings of DARPA Speech and Natural Language Workshop.

[13] Hiemstra, D. and de Jong, F. (1999) Disambiguation Strategies for Cross-language Information Retrieval, Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, pp. 274-293.

[14] Kleinberg, J. (1998) Authoritative Sources in a Hyperlinked Environment, Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 46(5), 604-632.

[15] Kwok, K. L. (2001) NTCIR-2 Chinese, Cross Language Retrieval Experiments Using PIRCS, Proceedings of NTCIR workshop meeting.

[16] Lu, W. H., Chien, L. F., Lee, H. J. (2001) Anchor Text Mining for Translation of Web Queries, Proceedings of the 2001 IEEE International Conference on Data Mining, 401-408.

[17] Lu, W. H., Chien, L. F., Lee, H. J. (2002) Translation of Web Queries using Anchor Text Mining, ACM Transactions on Asian Language Information Processing (TALIP), 159-172.

[18] Lvarenko, V., Choquette, M., Croft, W. B. (2002) Cross-lingual Relevance Model, Proceedings of ACM-SIGIR 2002 Conference, 175-182.

[19] Nie, J. Y., Isabelle, P., Simard, M., and Durand, R. (1999) Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web, Proceedings of ACM-SIGIR’99 Conference, 74-81.

[20] Rapp, R. (1999) Automatic Identification of Word Translations from Unrelated English and German Corpora, Proceedings of the 37th Annual Conference of the Association for Computational Linguistics.

[21] Smadja, F., McKeown, K., Hatzivassiloglou, V. (1996) Translating Collocations for Bilingual Lexicons: A Statistical Approach, Computational Linguistics, 22(1), 1-38.

從語域及借詞觀點探討台語文寫作風格

楊允言
花蓮 大漢技術學院資訊工程系
ungian@ms01.dahan.edu.tw
http://203.64.42.21

摘要

本文利用台語語詞中移借語層詞彙的使用情形來探討台語文的寫作風格，將移借語層的詞彙視爲日語借詞、華語借詞和教會用語三個主要不同的語域(register)。把不同年代的文本拿來比較，可以看出其差異性。

本文的實驗，以李勤岸的研究爲基礎，其語料爲 1920 年代及 1990 年代的台語文小說作品，他將語料中出現的詞彙分出所屬的各語域。本文則利用各語域的台語詞彙，探討台語文的寫作風格。除上述語料，並加上卓緞女士白話字歌詩，以計算詞彙所屬語域的方式，來量化寫作風格，並說明台語文的特殊性，這特殊性基於政治環境及宗教因素等。此外，並針對實驗結果，台語文的語域、文類與寫作風格的關係，以及書面語和漢字制約問題對台語文的影響等，做進一步討論。

如果利用台語詞彙，從語域及借詞觀點來探討台語文寫作風格的方法可行，那麼本實驗應可做爲進一步研究的基礎。

關鍵詞 台語文(Written Taiwanese)、語域(Register)、借詞(Loanword)、寫作風格(Writing Style)、文類(Genre)、台語羅馬字(Taiwanese Romanization)

1. 台語文的特色與特殊性

在台灣，台語是許多人的母語，然而大部分人使用台語僅止於「聽」與「說」，並不包括「讀」和「寫」，究其原因，與台語所身處的歷史背景或政治現實有很大關係。

若以漢字角度來談，台語一字多音現象比華語更加明顯。以台語線上字典的資料爲例，此線上字典共有 22,080 個項目（entry，每個項目包括一個漢字及其讀音），11,635 個不同的漢字，平均每個漢字有將近兩個讀音（最高紀錄是 1 個漢字 12 個不同讀音）；如果只考慮 BIG-5 的常用字集，則比例更高（13,176 個項目，5,337 個不同漢字，平均每個漢字將近 2.5 個讀音）。[1]其主要原因爲，台語的位階較低，主體性較差，導致在不同時期需吸收不同的漢字讀音。

上述問題，若以詞的角度來看，兩音節或以上的詞彙，部分的讀音模糊性問題得以解決（但不是全部）。而台語的詞彙又是另外一個錯綜複雜的問題。在我們日常生活中所使用的台語，夾雜了不少華語、日語，[2]還包括原住民語，此外，更底層的百越語、中國歷代的官話系統也在其中。1935 年，郭一舟發表〈福佬話〉一文，形容台語的詞彙現象是「九重粿」，不管是從歷時或共時的角度看皆如此。

另外一個問題是詞彙變遷，因爲政治環境變動大，以及台語位階低，其詞彙變遷速度遠比華語快。變遷包括同一個詞，現在和從前的用法不同，以及從前使用的詞彙現在被另一個新詞彙取代。

台語文自然語言處理，其中一個重要的基礎就是台語詞彙的整理，然而，要把「九重粿」中的每一層次逐層抽離出來，實非易事。

[1] 台語線上字典的資料以 1913 年出版，甘爲霖的《廈門音新字典》爲主要來源，這本字典應該是台灣最通行的台語字典，有其代表性。台語線上字典於 2003 年 1 月上線，可以用漢字查發音（羅馬拼音），也可以反向查詢，網址爲 http://203.64.42.21/TG/jitian。

[2] 在台灣，語碼混合(code-mixing)的情形普遍，不過，所夾雜的日語或華語能不能算是台語，有一個簡單的檢驗規則：對於不黯日語或華語的台語使用者(native speaker)是否使用這個華語或日語詞，如果是，則

2. 台語詞彙變遷的研究與寫作風格

李勤岸(2000)曾對台語的詞彙變遷做較深入的研究。他認爲某一語言和其它語言的接觸，一開始是語碼轉換(code-switching)或是語碼混合(code-mixing)的方式出現，之後成爲借詞，再漸漸深入本土語層；以目前而言，借詞的主要來源爲廈門話、日語和華語。他將台語的詞彙分爲兩大語層：本土語層(local layer)及非本土語層(non-local layer)（又稱爲移借語層，loanword layer），本土語層包括澳亞語層、南島語層、古漢語以及中古漢語，移借語層則主要包括教會用語(church register)、日語借詞和華語借詞。日語借詞和華語借詞很明顯是因爲台灣的歷史、政治因素，而教會用語則較爲特殊，屬於宗教因素，而教會用語又可分爲廈門話和禮拜儀式用語(liturgical register)。

教會用語的特殊性在於，日語借詞和華語借詞是兩個語言接觸時，強勢語言（日語、華語）影響弱勢語言（台語）的口語及書面語，教會用語卻是同一類語言（如果把廈門話和台語視爲方言的關係）以書面語來影響口語，而如果以台語和華語相比較，書面語影響口語的現象，在台語非常明顯。

他所用的語料包括 1916 年及 1935 年出版的台語聖經（用來分析教會用語中的廈門話），[3]1920 年代鄭溪泮及賴仁聲用台語羅馬字寫的台語小說（用來分析日語借詞），以及 1990 年代陳雷及陳明仁用漢羅合用方式[4]所寫的台語小說（用來分析華語借詞），將這些語料逐一打字建檔校對；利用斷詞系統將語料斷詞，[5]並經過人工校對。接著將其中非本土語層的詞彙挑出來，日語借詞和華語借詞較易分辨，而教會用語中的廈門話則

此詞彙可視爲台語，如日語的”khi-mo”、”tha-ma-to”，或華語的「大哥大」、「小姐[che]」等。

[3] 這是巴克禮(Thomas Barclay)翻譯的版本，1916 年出版新約，1935 年出版舊約，雖然是以廈門話寫成，部分詞彙和語法和台語不同，不過主要閱讀人口都是台灣人。真正「台語」版本的聖經，在 1972 年才由高積煥和陳邦鎮以漳州腔編寫出版，稱爲紅皮聖經，卻被當時政府以影響國語推行的理由而沒收。

[4] 漢羅合用，指的是台語文的一種書寫方式，目前可算是台語文書寫的主流，基本觀念是，若這個漢字一般人很容易知道台語要怎麼讀，就用漢字，否則就用台語羅馬字。

[5] 如果是台語羅馬字的語料，詞彙之間以空白分隔，沒有斷詞的問題。

需細加斟酌，除了本身爲台語使用者的語言學者的語感和直覺，還需藉助相關資料。[6]

李勤岸的語料量是：1920 年代有 112,964 個詞次（word token）、12,941 個詞型（word type），1990 年代有 92,539 個詞次、12,969 個詞型。這些書面語中，移借語層詞彙的數量及變化請參看表一：

表一 兩個時代移借語層詞彙數量的變化

	詞型數	教會用語	日語借詞	華語借詞	總計
1920 年代文本	12,941	563	178	27	768
1990 年代文本	12,969	97	239	202	538

資料來源：李勤岸(2000) Table 4.4、Table 5.3、Table 5.16
說明：以詞型來計算，非詞次。

台語詞彙在這段時間的變化，教會用語大量減少，日語借詞增加小部分，華語借詞則大幅度增加，而整體來說，使用移借語層詞彙的數量減少。[7]

另外一個度量是詞彙豐富度（lexical richness），計算方式爲：

$$\text{詞彙豐富度} = \frac{\text{詞型(word type)}}{\text{詞次(word token)}}$$

其計算出的結果請參看表二：

表二 兩個時代詞彙豐富度的比較

	詞型	詞次	詞彙豐富度
1920 年代	12,941	112,964	11.46%
1990 年代	12,969	92.539	14.01%
變化情形			+2.55%

資料來源：李勤岸(2000) Table 5.31

我們可以說，這些是以小說爲語料所算出來的結果。

李勤岸的研究，主要是從移借語層詞彙的使用情形來探討台語語詞變遷。而本文則

[6] 根據王毯仁的研究，閩南語 2,000 個常用詞中，有 200 多個台語不使用；而閩南語一般以廈門話爲代表。1916 年及 1935 年所出版的台語聖經，其實是以廈門話爲本，當中所用的詞彙，有些並未通行於台灣。不過，是因爲此詞彙從來沒有在台灣被使用，還是以前曾經有人使用而目前已被別的詞彙取代，這又是另一值得深入探討的題目。王毯仁的研究引用自李勤岸(2000)。

[7] 這裡要說明的是，兩個不同年代有時空背景的差異。1920 年代，台語羅馬字的閱讀人口主要爲天主教及基督教徒，而 1990 年代台語羅馬字已經世俗化，非教徒的閱讀人口應該遠超過教徒，不過如前所述，這時候的文字形式爲漢羅合用。

是利用移借語層詞彙的使用，來探討寫作風格。由於作者基本背景的差異，這個差異包括世代、宗教信仰等，導致其作品偏好使用的移借語層詞彙有所不同，從而產生不同的寫作風格。

3. 特殊的教會用語寫作風格

這是頗特殊的例子。2001 年 2 月，報端刊載一則百歲人瑞卓緞女士過生日的消息。對地方而言，人瑞可傳爲美談，不過，報導中還提到，卓緞女士沒有受過體制內教育，全然不識漢字，[8]卻在 51 歲無師自通地使用在教會所習得的台語羅馬字（又稱爲教會羅馬字、白話字）創作白話字歌詩，並持續到她九十多歲時。其寫作時間爲 1950 年代至 1990 年代。

因爲地利之便，我與卓緞女士家屬取得聯繫，並著手進行白話字歌詩的整理。經過半年多，總共整理 25 首白話字歌詩及兩篇自傳性的文章。[9]

卓緞女士的白話字歌詩與一般台灣民間的七字仔類似，七字一句，有押韻，教會稱爲七言文，早期的《台灣教會公報》裡有不少這種文類的文章，[10]我認爲這是當時基督教爲融入台灣社會所發展出的一種文學創作方式。以下舉一小段白話字歌詩爲例，其中左邊是原文，右邊是以漢羅合用方式整理後的文稿。

Gō͘-ko sìn-gióng chin kian-tēng	五哥信仰真堅定
Kóng-ōe sì-chià^{n} chòe kng-teng	講話四正做光燈
Iā ū un-jiû ê sèng-chêng	Iā 有溫柔 ê 性情
Ho̍k-sāi Siōng-tè iā khiân-sêng	服侍上帝 iā 虔誠

[8] 相關報導刊登在東部發行的更生日報，以及自由時報、中國時報、聯合報的花蓮地方版。其中，中國時報的標題，竟然出現「目不識丁」。台語羅馬字也是文字，不識漢字並不表示目不識丁。這點令我感受到漢字教育所造成的觀念偏差。

[9] 因爲卓緞女士不識漢字，也聽不懂華語及日語，家中只有幾本泛黃的台語聖經、聖詩，這其間，我常常走訪，並將手邊的台語羅馬字資料放大影印供她閱讀，而她非常高興。卓緞女士的作品，目前已完成台語羅馬字原文和台語漢羅合用的兩個版本，家屬希望還有華文、日文及英文三個版本的整理再行出版，不過我傾向加註華文註解，若翻譯成別種語文，可能風味盡失。目前正尋找出版社出版。

[10] 《台灣教會公報》於 1885 年開始發行，原名爲《台灣府城教會報》，是台灣最早發行的報紙。歷經幾次改名和整併，目前名稱爲《台灣教會公報》。使用台語羅馬字書寫，在日本時代曾因太平洋戰爭而被迫停刊；戰後復刊，但後來政府推行國語政策，於 1970 年被迫全面改成華文。以上資料引用自張裕宏(2001)。

如果讀者並非教徒，會發現其中有些詞彙較少出現在其日常用語中，例如「信仰」、「光燈」、「服侍」、「上帝」、「虔誠」等。而短短 4 句、28 個音節，總共 17 個詞彙中，就出現了 5 個這樣的詞彙，比例之高，很容易令人察覺到其特殊的寫作風格，即大量運用教會用語的寫作風格。[11]

本文利用李勤岸整理的移借語層詞彙，從這些詞彙使用所透露的訊息，來探討卓緞女士作品的寫作風格，並與李勤岸的語料所得結果做比較，探討相關問題。

4. 實驗步驟及結果

實驗的步驟如下：

(1) 語料爲我已經整理好的卓緞女士白話字歌詩 25 首，因爲是台語羅馬字寫作，不用經過斷詞的處理；

(2) 計算的結果，總共有 4,594 個音節，[12]2,878 個詞次，829 個詞型；

(3) 將這 829 個詞，與李勤岸整理出來的移借語層的所有詞彙逐一比對，其中發現，除了大量的教會用語外，只有極少數的日語再借詞（如「環境」、「博士」等），[13]和華語借詞；

(4) 比對結果，將這 829 個詞型標記類別，區分爲教會用語（標記爲「C」）和本土語層詞彙（標記做「L」）。

表三列出詞頻較高的前 30 個詞型（詞頻>=14），表四則列出教會用語中詞頻較高的

[11] 另外，若以創作題材而言，除了和一般七字仔相同，都有勸世的題材外，卓緞女士白話字歌詩還有許多慶祝節慶（聖誕、生日、新年）的，七字仔則幾乎沒有這類題材。這應該是宗教氣氛環境的影響。

[12] 如果討論華文，一字就是一音節，不過台語文若用羅馬字書寫，一音節包括好幾個字元，用「字」來等同音節並不妥當。

[13] 日語再借詞（deep-relexified Japanese loanwords）指的是由古漢文創造出來的日語新詞，華語和台語再從日語借用過來，也算是台語和華語的共通詞（但是台語不是從華語借用過來），屬於移借語層中的日語借詞。而卓緞女士的作品中，只有出現三個日語再借詞，沒有一般的日語借詞。而詞頻出現最高的「博士」（5 次），經由台語文語詞索引程式(http://203.64.42.21/TG/concordance)的查詢結果，「博士」一詞在台語聖經中出現 23 次。這說明，雖然「博士」一詞是日語再借詞，不過卓緞女士應該就是由教會的台語羅馬字資料學到這個詞彙。

前 28 個詞型（詞頻>=4）。

表三 詞頻較高的30個詞型

詞頻	詞型	漢羅	類別
107	Chú	主	C
107	Siōng-tè	上帝	C
77	ê	ê	L
68	chin	真	L
66	lâi	來	L
60	lán	咱	L
54	ū	有	L
47	hō͘	hō͘	L
44	góa	我	L
37	hoaⁿ-hí	歡喜	L
37	i	伊	L
31	chin-chiàⁿ	真正	L
30	chiok-hok	祝福	L
26	sī	是	L
25	kám-siā	感謝	L
23	kiù-chú	救主	C
22	beh	beh	L
22	hok-khì	福氣	L
21	chòe	做	L
21	tāi-ke	大家	L
21	tio̍h	著	L
20	chhōa	chhōa	L
20	Sè-kan	世間	L
19	lâng	人	L
18	bô	無	L
18	o-ló	呵咾	L
17	kiâⁿ	行	L
16	hó	好	L
16	tī	tī	L
14	chiū	就	(C)

表四 教會用語詞頻較高的28個詞型

詞頻	詞型	漢羅	類別
107	Chú	主	C
107	Siōng-tè	上帝	C
23	kiù-chú	救主	C
14	chiū	就	(C)
11	Bo̍k-su	牧師	C
11	ōe	會	(C)
10	êng-kng	榮光	C
8	Iâ-so͘	耶穌	C
8	sūn-ho̍k	順服	C
7	hok-im	福音	C
7	sèng-kiáⁿ	聖囝	C
7	thiⁿ-pē	天父	C
6	Ha-le-lu-iah	(哈里路亞)	C
6	ín-chhōa	引chhōa	C
5	chóng-hōe	總會	(C)
5	sio̍k-hôe	贖回	C
5	Thian-kok	天國	C
4	bo̍k-chiá	牧者	C
4	gîm-si	吟詩	C
4	kî-tó	祈禱	C
4	kiù-un	救恩	(C)
4	kng-teng	光燈	C
4	lāi-kiàm	利劍	C
4	mô͘-kúi	魔鬼	(C)
4	si̍p-jī-kè	十字架	C
4	siúⁿ-sù	賞賜	(C)
4	tiúⁿ-ló	長老	C
4	tōa-tek-sèng	大得勝	C

其中，類別標記為「(C)」的詞型，沒有出現在李勤岸整理的移借語層詞彙中，不過我認為是教會用語。[14]會有這樣的分歧，主要原因包括：(i) 在他整理的語料中沒有出現此詞彙，(ii)「會」、「就」兩詞，如果用漢字書寫，我會將之視為本土語層詞彙，然而卓

[14] 標記為「(C)」的詞彙，除了表四中所列出的，還有詞頻為 3 的「該哉[kai-chài]」、「交託[kau-thok]」、「禮物[lé-mi̍h]」等 3 個，詞頻為 2 的「滅無[bia̍t-bô]」、「至微細[chì-bî-sòe]」、「尊崇[chun-chông]」、「近倚[kūn-óa]」、「跟 tè[kun-tè]」...等 12 個，詞頻為 1 的共有 29 個。

緞女士的作品以台語羅馬字書寫，一般台語使用者，這兩個詞分別唸成”ē”、”tȯh”，而非”ōe”、[15]”chiū”，(iii) 分類很難避免模糊性的問題。

表五列出卓緞女士作品中各類詞彙的數量及所佔比例。

表五 卓緞女士作品中各類詞彙的數量及比例

	本土語層		移借語層						總計
			教會用語		日語借詞		華語借詞		
詞型	677	81.66%	144	17.37%	3	0.36%	5	0.60%	829
詞次	2,291	79.60%	569	19.77%	8	0.28%	10	0.35%	2,878

5. 相關問題探討

本文中採用的卓緞女士作品的文本，或許較特殊，與相同文類的其它作品相比是個極端（使用過多的教會用語）。不過，這樣的文本卻也提供了一個視野，讓我們感受到，在幾乎不受日語及華語的影響下，台語文可以表現出什麼樣的風格。以下打算針對前述的實驗結果，做相關問題的探討，包括實驗結果的進一步分析討論，語域、文類與寫作風格的關係，以及台語文受書面語及漢字制約的問題。

5.1 不同文本移借語層詞彙使用比較

表六列出三個不同文本移借語層詞彙的數量及比例。

表六 三個文本移借語層數量及比例

	詞型數	教會用語		日語借詞		華語借詞		總計	
1920 年代文本	12,941	563	4.35%	178	1.37%	27	0.21%	758	5.93%
卓緞作品	829	144	17.37%	3	0.36%	5	0.60%	152	18.35%
1990年代文本	12,969	97	0.75%	239	1.85%	202	1.56%	538	4.15%

部分資料來源：李勤岸(2000) Table 4.4、Table 5.3、Table 5.16
說明：以詞型來計算。

若以 1920 年代文本及 1990 年代文本做比較，我們發現：(i)華語借詞大量增加，這是政治因素；(ii)日語借詞不減反增，但若詳加檢視，兩個時代所用的日語借詞差異頗大，

[15] 「會」讀做”ē”或”ōe”是方言差的問題，事實上就我所知，目前仍有少部分泉州腔讀做”ōe”，但是我很確定，許多教會人士是因爲聖經等台語書面語的關係而讀成”ōe”，而他們的腔調是”ē”。

1920 年代的日語借詞應以政治因素居多，而這些詞彙戰後可能被台語對等詞彙或華語借詞取代（例如「辯護士」被「律師」取代、「探偵」被「偵探」取代、「無料」被「免費」取代、「案內」被「招待」、「引 chhōa」取代...等），1990 年代的日語借詞，應該是整個社會、經濟因素所造成；(iii)教會用語大量減少，這與台語羅馬字的社會脈動有關，1920 年代文本的讀者主要爲教徒，1990 年代文本的讀者爲一般社會大眾，顯示台語羅馬字在整個台灣的台語文運動中，有世俗化的趨勢，不再僅限於教會人士。

若考慮卓緞女士的作品，除了大量使用教會用語，令人感受到很不一樣的寫作風格外，極少量的日語、華語借詞，也多少反映出的她的生活背景。若對照她的自傳性文章，確實發現她的生命歷程很艱辛，憑藉著宗教信仰而得到心靈的寄託。[16]

一般人眼中，卓緞女士生活在社會階層最底層，他們沒有機會受教育，導致不識漢字，失去了書寫媒介，更不易被社會大眾接受或關注。卓緞女士卻因爲有機會接觸基督教，得以學習台語羅馬字，並在 30 年後以此爲媒介，將自己的際遇、心情呈現出來。她是一個特例，讓我們得以一探他們所處的生活環境，而且是由自己書寫而非透過他人轉述。她所提供的語料之所以特別，在於我們所建構出的書面語語料極少這樣的材料，但是處於這樣生活環境的人卻不算少數。

5.2 詞彙豐富度的探討

詞彙豐富度的比較請參考表七：

[16] 在她出生後三天母親難產死亡，父親無力撫養，於是轉給貧苦人家做童養媳，二十歲左右在花蓮做牧師的五哥想起曾有這一位妹妹，找到她後，將撫養她的家庭一起從桃園帶來花蓮鳳林，五哥引導她接觸基督教，卻被其養父母禁止；後來生了一場大病，養父母只讓她接受「王祿仔仙」的治療、喝「符仔水」，拖了一年多未見起色，曾痛苦得想要自殺，後來幾位哥哥共同出錢籌措醫藥費，帶她去花蓮醫院醫治才得以痊癒，而後養父母才不再禁止她接觸基督教。婚姻不美滿，歷經許多波折，40 多歲時帶著兩名年幼的女兒來花蓮，爲了養家活口必須身兼數個工作，直到女兒長大有工作後才得以稍稍喘息。而面臨生活上的困頓，宗教信仰提供她心靈的唯一寄託。

表七　詞彙豐富度的比較

	詞型	詞次	詞彙豐富度
1920年代文本	12,941	112,764	11.48%
卓綴作品	829	2,878	28.80%
1990年代文本	12,969	92,539	14.01%

部分資料來源：李勤岸(2000) Table 5.31

卓綴女士沒有受過體制內教育，作品中極少日語和華語借詞，但是詞彙豐富度仍有1990年代文本的兩倍以上。這個結果與我們預期的不同，沒有受過體制內教育，應會導致所使用詞彙受限而降低詞彙豐富度，但是大量的教會用語卻彌補了這項缺憾，反而大大提高詞彙豐富度。從這裡可以看出台語書面語的強大影響力。

當然，這牽涉到不同文類(genre)的問題，卓綴女士的作品爲歌詩，作品中所使用的虛詞(function word)應該比小說少得多，而詞頻較高的詞彙以虛詞較多，減少虛詞使用自然容易使詞彙豐富度增加。

關於這部分，應該要以相同文類不同作者的文本做進一步計算，才能釐清這個問題。

5.3 語域、文類與寫作風格的計算

語域、文類等名詞的定義，在社會語言學界並不一致。本文中，語域指的是同一語言在不同世代、不同領域、不同行業中的語言變體(language variation)，並把範圍縮小到書面語；而文類指的是書面語的各種體裁類型，例如小說、詩歌、學術論文等。

不同的語域或文類，可以產生不同的寫作風格(writing style)。風格是抽象的概念，我們將這些文本中的詞彙抽出，利用計量的方式，量化成文本的風格。而計量的方式，就是先將詞彙做分類，尤其著重移借語層的分類。

台語曾經歷兩次有系統、全面性的壓制，一次是日語，一次是華語，這是政治力量促成的變化，比自然演變的速度快很多，影響的層面除了詞彙，還包括句法，不過詞彙的影響最爲顯著。

以各種不同類型的借詞來區分不同世代的語域，對台語是有用的。舉例來說，曾受日本統治的老一輩台灣人，其日常用語中所使用的日語詞彙會比較多，年輕一輩則較能運用華語詞彙。這在口語應可充分表現出來，而在書面語呈現時可能會被淡化，[17]不過仍可由上述的分析中看出端倪。從華語借詞的數量可以清楚看出1920年代和1990年代文本的差異。而卓緞女士作品所呈現出來的詞彙運用，顯然是沒有受到體制內教育影響的結果，對照不同世代的教育普及率，這是較早世代的寫作風格。

5.4 書面語及漢字問題

從卓緞女士的作品，我們清楚看到台語羅馬字書面語所產生的影響。這樣的例子雖然不多，但是並非唯一，只是沒有透過漢字或是華文的媒介，這些人、這些文本不易引人側目。

較特殊的是，台灣雖然已經遠離日本統治，但是90年代文本的日語借詞不減反增，我們假設台灣與日本的交流頻繁導致日語持續走入台語的日常用語中，可是對照同樣在台灣發展的華語，這個情形卻不明顯。這個現象應該是華語受漢字制約的原因，漢字的特性導致其不易借用外來語。台語文以漢羅合用方式，羅馬字的部分很適合用來處理借詞；此外，台語相較於華語，有較多音節，[18]這是有利於吸收外語借詞的因素。事實上，某些行業如建築業、製造業等日語的借詞特別多。若有詳盡的台語外來語辭典可供參考，這部分可做進一步的探討。

漢字制約另一個情形是，目前台語文作家，幾乎都受過完整且長時間的華文教育，華文的運用毫無問題，並在日常生活中大量接受華文訊息：報紙、書籍或是電視字幕等。慣用華語思考，極可能不自覺將許多華文詞彙直接帶入其台語文作品中，除非以純粹的

[17] 因為書面語在書寫的過程中，作者可能會儘量過濾掉這些借詞。

[18] 台語的音節數，我們以台語線上字典為例，共有2,728個音節；華語的部分，以詞庫小組技術報告98-01詞頻詞典的索引來計算，共有1,081個音節；台語約是華語的2.5倍。音節的組成單位是聲母、韻母及聲

台語羅馬字書寫。如果把台語分爲口語和書面語兩部分，日語借詞的數量，應該是口語多於書面語，而華語借詞的數量則相反，應該是書面語多於口語。這部分若要進一步探討，需藉助台語口語語料庫的建立。

6. 我們還需要什麼？

本文只做了一個初步的實驗，如果將李勤岸整理的1920年代、1990年代台語文文本視爲那個時代台語書面語一般的寫作風格，則其它文本，若有其特殊性，可以透過比較的方式，說明其特殊點。如果從語域、借詞來看台語文寫作風格這個觀點可行，我們需要更完整的資料來完成這件事。這些資料包括：

- ⇨ 完整的詞彙資料：除了基本的詞性等訊息之外，還要有語源的資料，最起碼能像李勤岸將本土語層和移借語層區分出來，並將移借語層再做區分，除了上述的日語借詞、華語借詞和教會用語外，還會有英語借詞、南島語借詞、客語借詞...等。[19]最理想的情形是把「九重粿」的每一層都剝離出來。這是大工程，期待關心台語的語言學者有此雄心壯志。若從另一個角度考慮，如果某些詞彙實在很難斬釘截鐵地確認其爲某一語域專用的詞彙，或許可以用統計語料的方式，來計算此詞彙在每一語域出現的比例，這是計算語言學者可以發揮的地方。
- ⇨ 完整的台語書面語語料庫：目前正在建立中，不過離「完整」的目標還有一段遙遠的距離。[20]目前可找到的最早的台語文資料是1885年所書寫的，[21]如果資

調。有較多的音節，意謂著對於借詞，更容易找到較相近的音節來處理。

[19] 英語借詞不多說；南島語借詞，例如我在花蓮，麵包果我們稱爲 pa-chi-lu，這是阿美語；客語借詞，例如 se-moe（小姐）、an-chu-se（非常謝謝）等。

[20] 台語文語料庫，由台灣羅馬字協會發起，從 2003 年 1 月開始著手建立台語文語料庫，分台語羅馬字文本及漢羅合用文本兩大部分，截至 2003 年 8 月底止，分別蒐集到了 92 萬音節及 164 萬音節的語料。目前提供一個簡單的語詞檢索程式(Concordancer)供使用者查詢，有 8 千 6 百多次的查詢人次記錄。期待未來以申請計畫的方式，加速此台語文語料庫的建置速度。請參考 http://203.64.42.21/TG/guliaukhou

[21] 中研院的語言典藏計畫中，台語文獻追溯到 16 世紀的歌仔冊。歌仔冊對於台語的保存，確實有其不可磨滅的貢獻，不過那個時候的書面語和口語有相當大的差距。這裡所說的 1885 年，是指用台灣的台語（而非廈門話等地的閩南語）所出版的《台灣府城教會報》，這份書面語書寫台語的口語。

料完全建檔，我們可以更詳細地討論同文類歷時(diachronic)的差異，或是不同文類共時(synchronic)的差異，或以性別、時間等因素做爲語域的區分等。目前世界上較有名的語料庫，其文類的分類標準並不一致，我們還可以計算同一文類中文本的差異和不同文類文本的差異，來檢視這個分類是否恰當。

⇨ 相關的工具：例如斷詞工具、詞性標示工具、語詞索引工具...等。

⇨ 台語口語語料庫：因爲至今台語文還沒有一套官方承認的正式書寫系統，這導致書面語的寫法不一致，也使口語少了文字的制約。所以，相較於華語，台語的書面語和口語之間的差異可能要大些。如果同時有書面語語料庫及口語語料庫可供比對，也許我們能更清楚台語的語言現象。

誌謝

本文寫作期間，感謝哈佛東亞語言文明系李勤岸助理教授、台東師院語教系張學謙助理教授、台語文作家吳景裕（陳雷）醫師對本文部分論點提出討論及建議。

參考資料

Douglas Biber, 1995,《Dimensions of register variation》, Cambridge Univ. Press.

楊允言, 2001,〈白話字傳奇〉,《台文通訊》86 期頁 1-3, 2001/4。(http://203.64.42.21/iug/Ungian/ Chokphin/ sanbun/POJthoanki/poj-ama.htm)

楊允言, 2002, 〈七字仔 tú 著上帝—論卓緞 ê 白話字歌詩〉,《第四屆台灣語言及其教學國際學術研討會》頁 489-497, 高雄：中山大學中文系，2002/4/27-28。(http://203.64.42.21/iug/Ungian/Chokphin/Lunbun/7jia/7ji-a-Tiongsan.htm)

楊允言, 2002, 〈論卓緞白話字歌詩 ê 寫作風格〉,《第八屆北美洲台灣研究學會年會》, 美國：芝加哥大學, 2002/6/27-30。(http://203.64.42.21/iug/Ungian/Chokphin/Lunbun/7jia/7ji-a-NATSC.htm)

李勤岸, 2000, 《Lexical Change and Variation in Taiwanese Literary Texts, 1916--1998 —A Computer-Assisted Corpus Analysis》, 美國夏威夷大學語言學研究所博士論文。

(2003 年 9 月由金安出版社出版)

郭一舟, 1935,〈福佬話〉,《台灣文藝》二卷 6 號、10 號, 1935/6 & 10.

黃昌寧、李涓子, 2002, 《語料庫語言學》, 北京：商務印書館。

黃佳惠, 2000,《白話字資料中的台語文學研究》, 台南師範學院鄉土文化研究所碩士論文。 (論文全文可至全國博碩士論文網下載 http://datas.ncl.edu.tw/cgi-bin/theabs/flywebi.cgi?p=3535&i=5229789&t=603&o=v5)

張學謙, 1998, 〈Ho-lo 台語的語層及語用〉, 《第二屆台灣語言暨語言學國際研討會論文選集》頁 451-463, 台北：文鶴。

張裕宏, 2001, 《白話字基本論：台語文對應&相關的議題淺說》, 台北：文鶴。(第一章導論有上網 : http://203.64.42.21/iug/Ungian/patlang/ POJkpl/POJkpl01.htm)

張德明, 1995, 《語言風格學》, 高雄：麗文文化。

卓, 1972, 《卓公宗慶子孫家譜—北部設教百週年暨卓家信主八十週年紀念》。

卓綴手稿，未出版。

ECONOMY IS A PERSON:
A Chinese-English Corpora and Ontological-based Comparison Using the Conceptual Mapping Model

Siaw Fong Chung	Kathleen Ahrens	Chu-Ren Huang
National Taiwan University	*National Taiwan University*	*Academic Sinica*
claricefong6376@hotmail.com	kathleenahrens@yahoo.com	churen@gate.sinica.edu.tw

Abstract

This paper proposes a corpora-based approach in comparing the Mapping Principles for economy metaphors in English and Chinese. The Mapping Principles are validated using an upper ontology (SUMO). This research extends on the work of Ahrens, Chung and Huang (2003) by examining the 'economy' metaphors in Chinese and English. In Ahrens, Chung and Huang (2003), they proposed to delimit the Mapping Principle via two steps: First, they used a corpora-based analysis on the word *jingji* 'economy' to find out the most prototypical mappings in a metaphor Second, they used an upper ontology (SUMO) to examine whether the mapping principle is a representation of conceptual knowledge in the ontology. This paper goes a step further by examining the similarities and differences of source domains in English and Chinese. Using the Conceptual Mapping Model, this paper looks particularly into the example of ECONOMY IS A PERSON. This paper observes the representation of shared knowledge in the source domain in different languages and explains the similarities and differences by looking into the definition of inference rules in the upper ontology of SUMO.

Key Words: Corpora, Conceptual Mapping Model, Mapping Principle, SUMO, ontology

1.0 Introduction

In the framework of Lakoff and Johnson (1980) and Lakoff (1993), conceptual metaphors are mappings from a concrete source domain to an abstract target domain. Lakoff proposes a "general principle" which is "part of the conceptual system underlying English" (1993:306). Ahrens (2002), however, suggested that this 'general principle' can be formulated in the form of Mapping Principle, an intuitive-based principle stating the underlying reason for source-domain mappings. These rules were verified with offline experiments (Ahrens 2002 and Lu 2002) in which they successfully predicted the reading times for metaphors that follow the mapping principles and metaphors that do not. Therefore, the 'general principle' can be delimited by providing Mapping Principle, which is specific for a particular metaphor to reason the mappings between source and target domains.

Ahrens, Chung and Huang (2003) proposed to delimit the Mapping Principle via two steps: First, they used a corpora-based analysis on the word *jingji* 'economy' to find out the most prototypical mappings in a metaphor and hence formed the mapping principle. Second, they used an upper ontology (SUMO http://ontology.teknowledge.com/) to examine whether the Mapping Principle is a representation of conceptual knowledge in the ontology. For example, in examining ECONOMY IS COMPETITION, the knowledge of 'competition' has a corresponding node with Contest in SUMO and “a War is kind of ViolentContest, which in term is a kind of Contest” (Ahrens, Chung and Huang 2003). Therefore, the metaphors ECONOMY IS COMPETITION and ECONOMY IS WAR can be subsumed under the same knowledge representation. These findings support the mapping principles that there are specific principles governing the source-target domain mappings.

In this paper, we will focus on one metaphor -- ECONOMY IS A PERSON – and compare the cross-linguistic data for the source domains of PERSON in English and Chinese. With these data, we also compare Mapping Principles cross-linguistically in both English and Mandarin. Two research questions are posed – (a) How similar or different the metaphor of ECONOMY IS A PERSON represented in English and Mandarin? (b) Are there differences in the representation of knowledge domains in English and Mandarin metaphor of ECONOMY IS A PERSON at the upper ontology level? To answer these questions, this paper adopts a similar methodology adopted by Ahrens, Chung and Huang (2003) by examining the corpora data as well as extracting the knowledge representation from SUMO to compare with the corpora data. However, this paper extends on previous research by examining the mapping in two languages. By comparing two languages, we can further investigate whether the similar Mapping Principle is extracting for the similar metaphor in two different languages. We foreshadow that if a similar metaphor with the same type of prototypical linguistic expressions is found in two different languages, the Mapping Principle should be the same. If the Mapping Principles are the same, the knowledge representations for both speech communities in describing that metaphor are also the same. In this paper, we will demonstrate this hypothesis by using corpora analysis of both Chinese and English metaphor of ECONOMY IS A PERSON.

2.0 Economy and Conceptual Metaphors

Metaphors are present in every day's language use. Some of these metaphors are so often used that the speakers are unaware of their metaphoric meanings. Charteris-Black (2000), for instance, carried out a comparative language analysis of the *Economist* magazine and the economist section of the Bank of English corpus. The results suggested that the metaphoric lexis in the *Economist* were higher in frequency than in the general magazines. This suggested that the ESP learners are dealing with more specific types of metaphors as part of their 'technical' register.

Incorporating this idea in teaching, Boers (2000) carried out an experiment comparing the teaching of economy metaphors to two groups of learners – one with special attention to the metaphoric meanings and the other with dictionary definitions of the metaphors. The subjects were the French-speaking university students of business and economics in Belgium. The targeted items for his experiment were *overcoming a hurdle*, *bailing out*, *weaning off*, *shifting tack* and *weeding out*. The different inputs for both groups were claimed to have affected the understandings of the learners – with the groups shown the metaphoric meanings performing better than the other group.

However, Boer's (2000) analysis of the metaphors lacks theoretical criterion in categorizing the metaphorical linguistic expressions. For instance, the examples of Health and Fitness (Boers, 2000:139) range from *sickly company* to *an acute shortage*. In addition, the target domain was unstated -- the term *storage* is ambiguous – i.e., it could have literally meant the shortage of medicine in some place or shortage of workforce. In order to define and delimit the target domain, this paper has chosen to look at economy metaphors appearing with the term 'economy.' By doing so, the target domain can be delimited. In regards of the source domain, we suggest the use of a single term and avoid overlapping scopes such as 'Health and Fitness.'

In what follows, this paper suggests the use of the Conceptual Mapping Model (Ahrens 2002), which provides a clearer theoretical analysis of metaphors.

The Conceptual Mapping Model

The CMM is a model based within the Contemporary Theory of Metaphor (CTM) (Lakoff and Johnson 1980, Lakoff 1993). It supports the idea that metaphors have systematic source to target domain mapping. However, the CMM goes beyond the CTM by postulating a principle connecting the mapping between the source and target domains. The CMM can also be used in analyzing metaphors linguistically by dividing the metaphorical expressions into entities (nouns), qualities (adjectives) and functions (verbs).

In Ahrens (2002), the metaphor IDEA IS BUILDING was analyzed. There were five steps to this analysis. These five steps are listed in Table 1:

Table 1: Analysis of IDEA IS BUILDING using the Conceptual Mapping Model

Step1	Given the target domain of IDEA, native speakers generated all items related to IDEA
Step 2	These generated items were categorized into similar source domains such as BUILDING and WAR
Step 3	For each source domain, the conceptual real world knowledge was generated. This was done by asking the following three questions: **1. What entities does the source domain (SD) have?** -- (for BULDINGS: foundation, structure, model, base, etc.) **2. What quality does the SD or the entity in the SD have?** -- (for BUILDING: shaky, high, short, strong, etc.) **3a. What does the SD do?** -- (for BUILDING: to protect, to shield, etc.) **b. What can somebody do to the SD?** -- (for BUILDING: to live in, to build, etc.)
Step 4	Non-conventional expressions generated in Step 1 were filtered out
Step 5	The actual mapping between the target (IDEA) and source (BUILDING) were compared with what could possibly be mapped in the real world.

For the metaphor of IDEA IS BUILDING, Ahrens (2001:279) proposed the following connection between the source and target domain pairings:

> *Idea (originally capitalized) is understood as building because buildings involve a (physical) structure and ideas involve an (abstract) structure.*

This connection is called 'Mapping Principle' (Ahrens 2001:279), which specifies the underlying reason for the mapping of source to target domains.

3.0 SUMO Ontology

SUMO (Suggested Upper Merged Ontology) is a shared upper ontology developed by the IEEE Standard Upper Ontology Working Group. It consists of concepts, relations and axioms that address a broad range of domains and interests. All concepts in SUMO are structured in the form of hierarchy with the root of Entity, which represents the most general concept. The Entity is divided into Physical and Abstract. These Physical and Abstract entities are then further divided into more specific nodes.

Applying ontology in linguistics, Niles (2003) suggested that the incorporation of the SUMO ontology with WordNet allows ontology to be used "automatically by applications (e.g. Information Retrieval and Natural Language Processing applications) that process free text.” The interest of this paper lies in observing the automated processing of Mapping Principles in the source-target domain mappings in conceptual metaphors.

In this paper, we demonstrate how SUMO helps delimit the source domain knowledge of metaphorical mappings. We also want to demonstrate how the source domain knowledge differs (or show similarities) across languages. In order to examining the similarities and differences cross-linguistically, the following section first displays our corpora analyses for economy metaphors in English and Chinese. These analyses help extracting the Mapping Principles of economy metaphors in both these languages. The concepts represented by the Mapping Principles will then be examined using the SUMO ontology. This incorporation of SUMO into our analysis allows the source domain knowledge (identified in the corpora analyses) to be defined at the upper ontology level.

The following section first presents the analyses of English and Chinese economy metaphors.

4.0 Corpora Data

Methodology

The Chinese data were extracted from the Academic Sinica Balanced Corpus, a tagged corpus with over 5 million words of Mandarin usage in Taiwan. The URL address for this corpus is http://www.sinica.edu.tw/SinicaCorpus/. 2000 search results of the Chinese term *jingji* 'economy' were analyzed for conceptual metaphors.

The English data were extracted from the corpora of the Linguistic Data Consortium (LDC), University of Pennsylvania. The URL address for LDC is http://www.ldc.upenn.edu/ldc/online/index.html. From the lists of corpora, term 'economy' was searched within the Wall Street Journal 1994, a corpus with the size of 14.3 MB (about 5 million words). This makes the size of both corpora almost the same for both English and Chinese. For each search, a maximum of 100 pages were extracted. Each page contains 100 instances. This paper selected the first 5 pages to look at, which constitutes approximately 500 instances of 'economy' in the corpus.

This paper has chosen to delimit the target domain of economy metaphors by running a search on the term 'economy' or *jingji* only. Other related terms such as 'currency' and 'market' are not the concerns of this current paper.

For both Chinese and English corpora, all instances were read through and metaphorical uses of 'economy' or *jingji* were marked manually. A metaphor was identified when the term ‘economy’ was expressed using concrete idea. For example, in the Chinese corpus, occurrences such as *jingji chengzhang* 經濟成長 ‘economy grew’ and *jingjizhan* 經濟戰 ‘economic battle’ were identified as metaphorical instances because there are the concrete domains of ‘growth’ and ‘war’ in the description of the economy[1]. Similarly, for English, instances such as ‘growing economy’ and ‘sputtering economy’ are identified as metaphorical due to the mapping of the concrete ideas of ‘growth’ and ‘engine’ in the metaphors. These metaphors were then collected and categorized according to different source domains (GROWTH CYCLE, WAR, COMPETITION, etc.) in Chinese and English respectively.

Results

The English corpus data produce a total of 209 recurring economy metaphors. Comparatively, in the Chinese data, a total of 311 recurring metaphors were found. The breakdowns of the data are shown in Table 2.

[1] In the next paper, we will demonstrate that linguistic expressions such as ‘growth’ and ‘war’ are definable as metaphors if they are hypernyms for at least one concrete and one abstract concept in the Wordnet. This incorporation of Wordnet strengthens the automation of the Conceptual Mapping Model in metaphors processing.

Table 2: Distributions of Economy Metaphors in the English and Chinese Corpora

Economy metaphors	Chinese *jingji*		English 'economy'	
	Types	Tokens	Types	Tokens
1. **ECONOMY IS A PERSON**	**11**	**121**	**26**	**131**
2. ECONOMY IS BUILDING	10	102	8	12
3. ECONOMY IS COMPETITION	11	40	3	15
4. ECONOMY IS WAR	12	23	--	--
5. ECONOMY IS JOURNEY	9	15	--	--
6. ECONOMY IS AEROPLANE	3	10	--	--
7. ECONOMY IS MOVING VEHICLES	--	--	25	51
TOTAL	**56**	**311**	**62**	**209**

There are three recurring source domains shown in Table 2, i.e., PERSON, BUILDING and COMPETITION (shaded above). Among these source domains, PERSON constitutes the majority of the total instances in both languages. In English, there are 131 tokens and 26 types of linguistic expressions found; In Chinese, there are 121 tokens and 11 types of linguistic expressions found. The types in the English data are more robust than in the Chinese data. Examples (1) and (2) below show examples of English and Chinese metaphor of ECONOMY IS A PERSON respectively.

(1) The immediate plate holds an **economy** with little **growth** and
low salaries, acute unemployment, expensive financing

(2) 國家 爲 促進 經濟 成長 （資本 累積、
guojia wei zujing jingji chengzhang zhiben leiji
country for improve economy grow capital accumulate
增殖） 的 使命，
zengzhi de shiming
multiply DE mission
“In order to improve the mission of making economy grows (accumulating and multiplying capital), the country…”

When we discuss ECONOMY IS A PERSON in detail, we will refer to more linguistic expressions in both languages.

The second source domain that appears in both languages is BUILDING. However, in Chinese, the use of the knowledge domain of ‘Building’ is far more frequent than the English data. In Table 2, we can see that there are 102 tokens in Chinese data and in the English data, there are only 12 tokens. This suggests that the Chinese prefer to use the knowledge (source) domain of BUILDING when describing economy metaphorically. This preference is not shown in the English data. Examples of ECONOMY IS BUILDING in both languages are shown in examples (3) and (4).

(3) being overbuilt needs to be taken in perspective of all the other parts of the
economy that are **overbuilt**, too."

(4) 爲 貴國 的 經濟 建設 盡 一 份 力量
wei guiguo de jingji jianshe jing yi fen liliang
for your contry DE economy building finish one CLASS power
“Contribute to the building of your nation’s economy.”

The third source domain is COMPETITION. As discussed in Ahrens, Chung and Huang (2003), the knowledge representation of ‘competition’ is corresponded with the node of ‘Contest,’ the same node that represents the concept of ‘War.’ If this is the case, the metaphors related to ‘Contest’ in Chinese is far more frequent than those in English. As we can see from Table 2, ECONOMY IS COMPETITION and ECONOMY IS WAR constitute 63 tokens in total whereas in the English data, ECONOMY IS COMPETITION only constitutes 15 tokens. This also shows that the concept of ‘ViolentContest’ is more viewed as a representation of ECONOMY by the Chinese speakers than the English speakers. Examples of these metaphors are shown in (5) to (7).

ECONOMY IS COMPETITION

(5) just as it is reshaping the **economy** to become more service-oriented , fragmented and **competitive** .

(6) 誰 能 掌握 **經濟** **競爭** 的 優勢，
shui neng zhangwo jingji jingzheng de youshi
who can control economy competition DE advantage
誰 就 能 立足 世界 舞台，
shui jiu neng lizu shijie wutai
who then can stand word stage
“Whoever can control the advantages of economy competition, that person can then stand on the stage of the world.”

ECONOMY IS WAR

(7) 一向 在 經濟 攻防戰 上
yixiang zai jingji fanggungzhan shang
always at economy attack-and-defend-war above
無堅不摧 的 日本
wujianbucui de riben
to-overrun-all-fortifications DE Japan
"Japan that is always overrunning fortifications at the economic battle…"

In addition to the source domains of PERSON, BUILDING, COMPETITION and WAR, there are other source domains of lower frequency. The English speakers also use the source domain of MOVING VEHICLES, which is not found in the Chinese economy metaphors. Contrastingly, the Chinese data show instances with the source domains of JOURNEY and AEROPLANE, which are also not used in the English data. Nevertheless, a comparison of these three source domains reviews that there are still similarities in these seeming different source domains. First, all these source domains are either referring to engine or moving vehicles or persons in the vehicles. Second, there are emphases on either directionality or speed when movements are concerned. For instance, the source domain of AEROPLANE in Chinese only refers to upwards movements whereas the source domain of MOVING VEHICLES refers particularly to speed of moving forwards. Examples are shown below.

ECONOMY IS AEROPLANE

(8) 臺灣 經歷 了 經濟 起飛， 成就 非凡
taiwan jingli le jingji chifei chenjiu feifan
Taiwan experience ASP economy take off results NEG-ordinary
“Taiwan has experienced the rises of economy and the results are extraordinary.”

ECONOMY IS MOVING VEHICLE

(9) the **economy** is going to **slow down** ,

(10) the U.S. **economy** were **barreling down the highway** at 100 miles

However, we will leave this portion under future research. In the next paper when we incorporate Wordnet into account, we will examine all linguistic expressions and compare their hypernyms so that the determination of metaphors and the selection of the source domains can become automated and hence overcome the limitations of the manual analysis.

For this current paper, we focus specifically on the source domain of PERSON, which obtained the most frequents scores in both languages. The following section will address this issue.

ECONOMY IS A PERSON

The details of the Chinese metaphors are shown in Table 3 and the English ones are shown in Table 4. In both Tables 3 and 4, the most frequent linguistic expressions are shaded. Expressions that appear in both Chinese and English are marked with a star (*) in both Tables 3 and 4.

Table 3: ECONOMY IS A PERSON in Chinese
M.P.: Economy is person because people have a life cycle and economy has growth cycle.

	Metaphor	**Frequency**
Entities	***成長 (growth)**	**67**
	衰退 (dysfunction)	8
	成長期 (growth period)	2
	病狀 (symptoms)	1
	命脈 (lifeblood)	2
	*衰頹(weakness and degeneration)	1
Functions	*成長 (grow)	21
	衰退 (to become dysfunctional)	5
	復甦 (regain consciousness)	9
	惡化 (deteriorate)	4
	*恢復 (recover)	1

Table 4: ECONOMY IS A PERSON in English

M.P.: Economy is person because people have a life cycle and economy has growth cycle.

	Metaphor	**Frequency**
Entities	*growth	15
	*growing	1
	exuberance	2
	*weakness	2
	recovery	4
	cooling	1
Quality	mature	1
	growing	4
	weak	9
	healthy	5
	ailing	5
	anemic	2
	recovering	2
	strong	20
	tiring	1
	depressed	2
Functions	***grow**	**41**
	shrinking	1
	weakening	1
	*recover	5
	suffer	2
	shudder	1
	hurt	3
	cool	2
	cool down	1

The driving principle of the Conceptual Mapping model is that there should be a principled reason for Mapping Principles. Ahrens, Chung and Huang (2003) hypothesized that this Mapping Principle can be automatically determined on the basis of frequency. Comparing the most frequent expressions in Tables 3 and 4, therefore, metaphorical terms that appear in both languages are ‘growth,’ ‘grow,’ ‘weakness’ and ‘recover.’ Among these expressions, ‘grow’ and ‘growth’ are the most frequent occurrences of source domain knowledge in the English and Chinese respectively. These outstanding recurring occurrences allow us to formulate the mapping principle for the Chinese and English metaphor of ECONOMY IS A PERSON as: *Economy is person because people have a life cycle and economy has growth cycle.*

This Mapping Principle is reflected in both the Chinese and English data. The English data, however, display more types (26) than the Chinese data (11). This is due to the mapping of ‘emotions’ in addition to the ‘physical growth’ in the English data. Expressions such as ‘depressed’ and ‘hurt’ are found repeatedly in the English examples (with ‘hurt’ being an ambiguous word referring to either physical or emotional hurts). However, the mapping of the emotion of a person is less frequent compared to the physical growth. Since our hypothesis considers the most frequent instances as contributors to the Mapping Principles, the occurrences of ‘emotion of a person’ do not interfere with the results.

In the next section, we will refer to the SUMO ontology in delimiting the source domain knowledge of the metaphors. The next section will explain why the source domain of PERSON can map expressions relating to 'growth' and at the same time allows the mapping of 'emotion' to PERSON. Using the SUMO ontology, this paper explains the source-target domains mappings using representation of shared knowledge provided by SUMO.

The Knowledge domain of 'Person' in SUMO

In the previous sections, our corpora analyses show that both English and Chinese 'economy' metaphors display the most prototypical Mapping Principle relating to 'growth' of a PERSON. The knowledge representation of 'growth' (or 'life cycle') was found to be involving the defining knowledge of an 'Organism' in SUMO, as stated in Ahrens, Chung and Huang (2003):

> *[T]he linguistic realizations of this [PERSON] mapping do not involve any knowledge that is specific to Human. In fact, it only involves the notion of a life cycle, which is the defining knowledge involving an Organism.* [Capital and word in square brackets added]

There are 16 inference rules for Organism in SUMO. All these inference rules were searched for and there is one that infers the shared knowledge of 'living object,' 'internal duration' and 'process.' These three concepts constitute the essential element of a 'growth' represented by the most prototypical linguistic expressions in the corpora. Hence, this inference rule was selected as reflection of the Mapping Principles of ECONOMY IS A PERSON. The inference rule reads as the following:

(=> (and (instance ?ORGANISM Organism) (agent ?PROCESS ?ORGANISM)) (holdsDuring (WhenFn ?PROCESS) (attribute ?ORGANISM Living)))

This rule encodes that 'An Organism is the agent of a living process that holds over a duration' (also stated in Ahrens, Chung and Huang (2003)). The consistency of this mapping ('growth') in English confirms the expectation of the Conceptual Mapping model that among the knowledge in the source domain, a particular aspect will show to be the most prototypical mappings. This prototypical mapping reflects the shared knowledge not only within a speech community, but across different speech communities. The data of the Chinese and English economy corpora analysis proves this point of view. In addition, the ability of an upper ontology to infer the similarity of prototypical mappings in two different languages also proposes the universality of the upper ontology.

However, in the previous section, we also observe that within the same source domain of PERSON, there are expressions referring mainly to aspect of 'living cycle' and there are also subsidiary frequencies of expressions relating to 'emotion' in the English data. The Organism, however, is defined as 'a living individual, including all plants and animals' in SUMO. With the occurrences of expressions relating to ‘emotions,’ we eliminate the possibility of Organism as referring to ‘a living plant’ in this metaphor. The definition of Emotion in SUMO is “the class of an attributes that denote emotional states of an Organisms.” This definition shows that ‘emotion’ is a state of an Organism and therefore a part of the shared knowledge of Organism. This complies with our analysis that categorizes expressions relating to ‘emotion’ to PERSON, which involves the node of Organism in SUMO.

From the Conceptual Mapping Model and SUMO inferences, we found that within a knowledge domain, the most prototypical mappings can be extracted using a corpus-based method. These prototypical mappings are formulated as Mapping Principles. Within two different languages, the existence of similar mapping principles can be explained using the inference rules of the shared knowledge in the upper ontology. This application of shared knowledge to similar Mapping Principles in different languages suggests the universality of the upper ontology. In addition, the inference rules also explain why there exist other aspects of knowledge aside from the most prototypical ones. This is because in different languages, a shared knowledge (such as Organism) may be chosen to express a similar metaphor (ECONOMY IS A PERSON), however, within this shared knowledge, there are elaborations of the conceptual nodes. For instance, in English, there are subsidiary elaborations referring to ‘state’ (EmotionalState) whereas in Chinese, there are elaborations referring only to ‘stage’ (living cycle) of an Organism. In general, however, the main mapping is the same (i.e., Organism) but the subsidiary mappings can differ. These results on main and subsidiary mappings are also reflected in the cross-linguistic study of TIME IS MOTION in Ahrens and Huang (2002). They proposed that when TIME IS A MOVING ENTITY the orientation of the ego is a conceptual subsidiary of the main mapping and can be parameterized differently in different languages.

In the case of ECONOMY IS A PERSON in English, the frequency of expressions relating to ‘emotions’ is low and therefore does not affect the most prototypical mapping – i.e., ‘growth.’

5.0 Conclusion

This paper provides a corpora-based analysis of the ‘economy’ metaphors in Chinese and English. The analysis supports a prototypical view of mappings that the most frequent mappings in a metaphor underlying the Mapping Principle (Ahrens 2002) for that metaphor. This paper also extends on the discussion of Ahrens, Chung and Huang (2003) in which they suggest a way of delimiting the source domain knowledge by using an upper ontology, i.e. SUMO. Looking into the example of ECONOMY IS A PERSON, we observe the representation of shared knowledge in the source domain in different languages and explain the similarities and differences by looking into the definition of inference rules in the upper ontology.

This paper contributes to further supporting the use of ontology and corpora data to automate the process of extracting Mapping Principles. This work provides a computational approach to refine Lakoff's (1993) statement that there is only 'general mapping principle' which exists between the mappings of source to target domain. This paper has shown that Mapping principles are not only specific but also extractable from corpora analysis.

In the corpora analysis, we constrain the Mapping Principle so that there is only one main Mapping Principle per source domain. We propose that this Mapping Principle is reflected by the prototypical (i.e. most frequent) mappings in the metaphor. If there is a subsidiary mapping in the same metaphor, as long as its frequency does not exceed the most prototypical mappings (such as 'stage'—i.e., 'living cycle'-- of a PERSON), the subsidiary mapping will not interfere with the main mapping. These main-and-subsidiary mappings can reflect cross-linguistic similarities and differences in conceptual metaphor mapping.

References

Ahrens, Kathleen. (2002). “When Love is Not Digested: Underlying Reasons for Source to Target Domain Pairings in the Contemporary Theory of Metaphor.” In Yuchau E. Hsiao (ed.). *Proceedings of the First Cognitive Linguistics Conference*. Cheng-Chi University. 273-302.

Ahrens, Kathleen and Huang Chu-Ren. (2002). “Time Passing is Motion.” *Language and Linguistics*. 3/3. 491-519.

Ahrens, Kathleen, Chung Siaw Fong and Huang Chu-Ren. (2003). “Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles.” *Proceedings of the ACL Workshop on The Lexicon and Figurative Language*. Sapporo, Japan. 53-41.

Boers, Frank. (2000). “Enhancing Metaphoric Awareness in Specialized Reading.” *English for Specific Purposes*, 19. 137-147.

Charteris-Black, Johnathan. (2000). “Metaphor and Vocabulary Teaching in ESP Economics.” *English for Specific Purposes*, 19. 149-165.

Lakoff, George and Mark Johnson. (1980). *Metaphors We Live By*. Chicago and London: The University of Chicago Press.

Lakoff, George. (1993). “The Contemporary Theory of Metaphor.” In Andrew Ortony (ed.). *Metaphor and Thought*. Second Edition. Cambridge: Cambridge University Press. 202-251.

Lu, Dora Hsin-yi. 2002. *Processing of Conceptual Metaphors in Mandarin Chinese A Conceptual-Mapping Model Based Study*. MA Thesis. Graduate Institute of Linguistics, National Taiwan University.

Niles, I. and Pease, A. (2003) “Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology.” In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering,*, Las Vegas, Nevada.

Automatic Pronominal Anaphora Resolution in English Texts

Tyne Liang and Dian-Song Wu
Department of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan
Email: tliang@cis.nctu.edu.tw; gis90507@cis.nctu.edu.tw

Abstract

Anaphora is a common phenomenon in discourses as well as an important research issue in the applications of natural language processing. In this paper, the anaphora resolution is achieved by employing WordNet ontology and heuristic rules. The proposed system identifies both intra-sentential and inter-sentential antecedents of anaphors. Information about animacy is obtained by analyzing the hierarchical relation of nouns and verbs in the surrounding context. The identification of animacy entities and pleonastic-it usage in English discourses are employed to promote the resolution accuracy.

1. Introduction

1.1 Problem description

Anaphora resolution is vital for areas such as machine translation, summarization, question-answering system and so on. In machine translating, anaphora must be resolved for languages that mark the gender of pronouns. One main drawback with most current machine translation systems is that the translation usually does not go beyond sentence level, and so does not deal with discourse understanding successfully. Inter-sentential anaphora resolution would thus be a great assistance to the development of machine translation systems. On the other hand, many of automatic text summarization systems apply a scoring mechanism to identify the most salient sentences. However, the task result is not always guaranteed to be coherent with each other. It could lead to errors if the selected sentence contains anaphoric expressions. To improve the accuracy of extracting important sentences, it is essential to solve the problem of anaphoric references in advance.

Pronominal anaphora is the most common phenomenon which the pronouns are substituted with previous mentioned entities. This type of anaphora can be further divided into four subclasses, namely,

Nominative: {he, she, it, they}

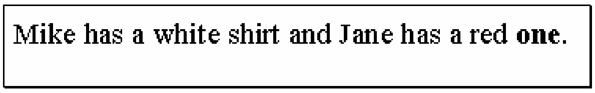
Reflexive: {himself, herself, itself, themselves}

Possessive: {his, her, its, their}

Objective: {him, her, it, them}

However, the usage of "it" can also be a non-anaphoric expression which does not refer to any items mentioned before and is called expletive or pleonastic-it [Lappin and Leass, 94]. Although pleonastic pronouns are not considered anaphoric since they do not have an antecedent to refer to, yet recognizing such occurrences is essential during anaphora resolution. In [Mitkov, 01], the non-anaphoric pronouns are in average of 14.2% from a corpus of 28,272 words.

Definite noun phrase anaphora occurs in the situation that the antecedent is referred by a general concept entity. The general concept entity can be a semantically close phrase such as synonyms or superordinates of the antecedent [Mitkov, 99]. The word *one* has a number of different uses apart from counting. One of the important functions is as an anaphoric form. For example:



Mike has a white shirt and Jane has a red **one**.

Intra-sentential anaphora means that the anaphor and the corresponding antecedent occur in the same sentence. Inter-sentential anaphora is where the antecedent occurs in a sentence prior to the sentence with the anaphor. In [Lappin and Leass, 94], there are 15.9% of Inter-sentential cases and 84.1% Intra-sentential cases in their testing result. In the report of [Mitkov, 01], there are 33.4% of Inter-sentential cases and 66.6% Intra-sentential cases.

1.2 Related works

Traditionally, anaphora resolution systems rely on syntactic, semantic or pragmatic clues to identify the antecedent of an anaphor. Hobbs' algorithm [Hobbs, 76] is the first syntax-oriented method presented in this research domain. From the result of syntactic tree, they check the number and gender agreement between antecedent candidates and a specified pronoun. In RAP (Resolution of Anaphora Procedure) proposed by Lappin and Leass [94], the algorithm applies to the syntactic representations generated by McCord's Slot Grammar parser, and relies on salience measures derived from syntactic structure. It does not make use of semantic information or real world knowledge in choosing among the candidates. A modified version of RAP system is proposed by [Kennedy and Boguraev, 96]. It depends only on part-of-speech tagging with a shallow syntactic parse indicating grammatical role of NPs and containment in an adjunct or noun phrase.

In [Cardie et al., 99], they treated coreference as a clustering task. Then a distance metric function was used to decide whether these two noun phrases are similar or not. In [Denber, 98], an algorithm called Anaphora Matcher (AM) is implemented to handle inter-sentential anaphora over a two-sentence context. It uses information about the sentence as well as real world semantic knowledge obtained from outer sources. The lexical database system WordNet is utilized to acquire the semantic clues about the words in the input sentences. He declared that most anaphora does not refer back more than one sentence in any case. Thus a two-sentence "window size" is sufficient for anaphora resolution in the domain of image queries.

A statistical approach was introduced by [Dagan and Itai, 90], in which the corpus information was used to disambiguate pronouns. It is an alternative solution to the syntactical dependent constraints knowledge. Their experiment makes an attempt to resolve references of the pronoun "it" in sentences randomly selected from the corpus. The model uses a statistical feature of the co-occurence patterns obtained from the corpus to find out the antecedent. The antecedent candidate with the highest frequency in the co-occurence patterns are selected to match the anaphor.

A knowledge-poor approach is proposed by [Mitkov, 98], it can also be applied to different languages (English, Polish, and Arabic). The main components of this method are so-called "antecedent indicators" which are used for assigning scores (2, 1, 0, -1) against each candidate noun phrases. They play a decisive role in tracking down the antecedent from a set of possible candidates. CogNIAC (COGnition eNIAC) [Baldwin, 97] is a system developed at the University of Pennsylvania to resolve pronouns with limited knowledge and linguistic resources. It presents a high precision pronoun resolution system that is capable of greater than 90% precision with 60% recall for some pronouns. [Mitkov, 02] presented a new, advanced and completely revamped version of Mitkov's knowledge-poor approach to pronoun resolution. In contrast to most anaphora resolution approaches, the system MARS, operates in fully automatic mode. The three new indicators that were included in MARS are Boost Pronoun, Syntactic Parallelism and Frequent Candidates.

In [Mitkov, 01], they proposed an evaluation environment for comparing anaphora resolution algorithms which is illustrated by presenting the results of the comparative evaluation on the basis of several evaluation measures. Their testing corpus contains 28,272 words, with 19,305 noun phrases and 422 pronouns, out of which 362 are anaphoric expressions. The overall success rate calculated for the 422 pronouns found in the texts was 56.9% for Mitkov's method, 49.72% for Cogniac and 61.6% for Kennedy and Boguraev's method.

2. System Architecture

2.1 Proposed System Overview

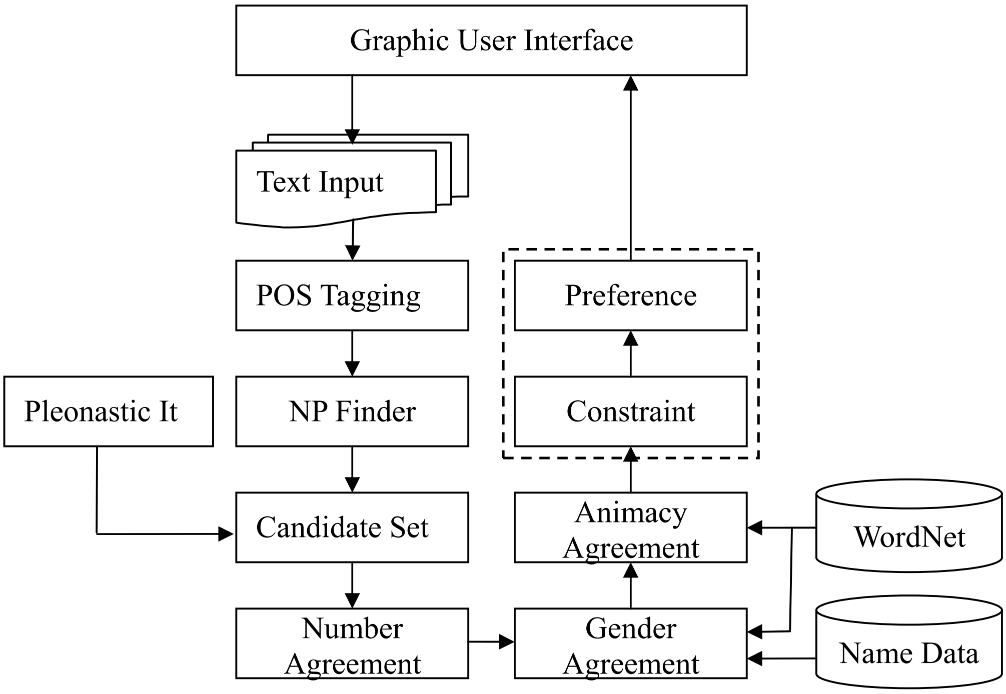


Figure 1: Architecture overview.

The procedure to identify antecedents is described as follows:

1. Each text is parsed into sentences and tagged by POS tagger. An internal representation data structure with essential information (such as sentence offset, word offset, word POS, base form, etc.) is stored.
2. Base noun phrases in each sentence will be identified by NP finder module and stored in a global data structure. Then the number agreement is implemented on the head noun. Testing capitalized nouns in the name gazetteer to find out the person names. The gender feature is attached to the name if it can be found uniquely in male or female class. In this phase, WordNet is also used to find out possible gender clues to increase resolution performance. The gender attribute is ignored to avoid the ambiguity while the noun can be masculine or feminine.
3. Anaphors are checked sequentially from the beginning of the first sentence. They are stored in the list with information of sentence offset and word offset in order. Then pleonastic-it is checked so that no further attempt for resolution is made.
4. The remaining noun phrases preceding the anaphor within predefined

window size are collected as antecedent candidates. Then the candidate set is furtherly filtered by the gender and animacy agreement.

5. The remaining candidates are evaluated by heuristic rules afterward. These rules can be classified into preference rules and constraint rules. A scoring equation (equation 1) is made to evaluate how likely a candidate will be selected as the antecedent.

$$score(can, ana) = (\sum_i rule\_pre_i - \sum_j rule\_con_j) \times \prod_k agreement_k \qquad (1)$$

where

can: each candidate noun phrase for the specified anaphor

ana: anaphor to be resolved

rule_pre$_i$: the *i*th preference rule

rule_con$_i$: the *i*th constraint rule

agreement$_k$: denotes number agreement, gender agreement and animacy agreement

2.2 Main Components

2.2.1 POS Tagging

The TOSCA-ICLE tagger [Aarts et al., 97] was used for the lemmatization and tagging of English learner corpora. The TOSCA-ICLE tagset consists of 16 major wordclasses. These major wordclasses may further be specified by features for subclasses as well as for a variety of syntactic, semantic and morphological characteristics.

2.2.2 NP Finder

According to part-of-speech result, the basic noun phrase patterns are found as follows:

base NP → modifier＋head noun

modifier → <article| number| present participle| past participle |adjective| noun>

In this paper, the proposed base noun phrase finder is implemented on the basis of a finite state machine (figure 2). Each state indicates a particular part-of-speech of a word. The arcs between states mean a word input from the sentence sequentially. If a word sequence can be recognized from the initial state and ends in a final state, it is accepted as a base noun phrase with no recursion, otherwise rejected. An example of base noun phrase output is illustrated in figure 3.

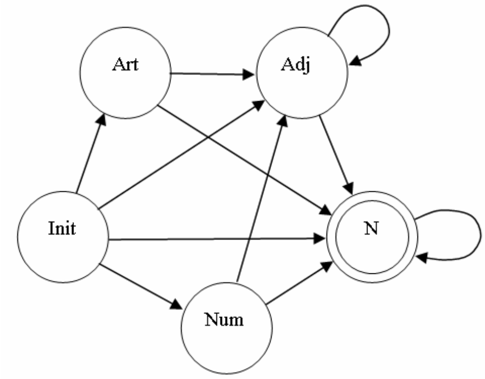
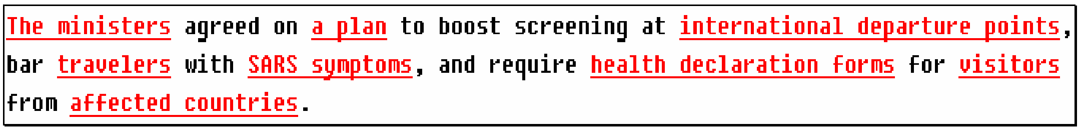


Figure 2: Finite state machine for a noun phrase.



```
The ministers agreed on a plan to boost screening at international departure points,
bar travelers with SARS symptoms, and require health declaration forms for visitors
from affected countries.
```

Figure 3: An Example output of base noun phrase.

2.2.3 Pleonastic-it Module

The pleonastic-it module is used to filter out those semantic empty usage conditions which is essential for pronominal anaphora resolution. A pronoun it is said to be pleonastic when it is used in a discourse where the pronoun has no antecedent.

The usage of “pleonastic-it” can be classified into state reference and passive reference [Denber, 98]. State references are usually used for assertions about the weather or the time, and it is furtherly divided into meteorological references and temporal references.

Passive references consist of modal adjectives and cognitive verbs. The modal adjectives (**Modaladj**) like advisable, convenient, desirable, difficult, easy, economical, certain, etc. are specified. The set of modal adjectives is extended with their comparative and superlative forms. Cognitive verbs (**Cogv**), on the other hand, are like anticipate, assume, believe, expect, know, recommend, think, etc.

Most of "pleonastic-it" can be described as the following patterns:

1. It is **Modaladj** that **S**
2. It is **Modaladj** (for **NP**) to **VP**
3. It is **Cogv-ed** that **S**
4. It seems/appears/means/follows (that) **S**
5. **NP** makes/finds it **Modaladj** (for **NP**) to **VP**
6. It is time to **VP**
7. It is thanks to **NP** that **S**

2.2.4 Number Agreement

Number is the quantity that distinguishes between singular (one entity) and plural (numerous entities). It makes the process of deciding candidates easier since they must be consistent in number. With the output of tagger, all the noun phrases and pronouns are annotated with number (single or plural). For a specified pronoun, we can discard those noun phrases whose numbers differ from the pronoun.

2.2.5 Gender Agreement

Gender recognition process can deal with words that have gender features. To distinguish the gender information of a person, we collect an English first name list from (http://www.behindthename.com/) covering 5,661 male first name entries and 5,087 female ones. Besides, we employ some useful clues from WordNet result by using keyword search around the query result. These keywords can be divided into two classes：

Class_Female= {feminine, female, woman, women}

Class_Male= {masculine, male, man, men}

2.2.6 Animacy Agreement

Animacy denotes the living entities which can be referred by some gender-marked pronouns (he, she, him, her, his, hers, himself, herself) in texts. Conventionally, animate entities include people and animals. Since we can hardly obtain the property of animacy with respect to a noun phrase by its surface morphology, we make use of WordNet [Miller, 93] for the recognition of animate entities. In which a noun can only have a hypernym but many hyponyms (an opposite relation to hypernym). In the light of twenty-five unique beginners, we can observe that two of them can be taken as the representation of animacy. These two unique beginners are {animal, fauna} and {person, human being}. Since all the hyponyms inherit the properties from their hypernyms, the animacy of a noun can be achieved by making use of this hierarchical relation. However, a noun may have several senses with the change of different contexts. The output result with respect to a noun must be employed to resolve this problem. First of all, a threshold value t_noun is defined (equation 2) as the ratio of the number of senses in animacy files to the number of total senses. This threshold value can be obtained by training on a corpus and the value is selected when the accuracy rate reaches the maximum.

$$t\_noun = \frac{the\_number\_of\_senses\_in\_animacy\_files}{the\_total\_senses\_of\_the\_noun} \tag{2}$$

$$t\_verb = \frac{the\_number\_of\_senses\_in\_animacy\_files}{the\_total\_senses\_of\_the\_verb} \tag{3}$$

$$accuracy = \frac{the\_number\_of\_animacy\_entities\_identified\_correctly}{the\_total\_number\_of\_animacy\_entities} \tag{4}$$

Besides the utilization of noun hypernym relation, unique beginners of verbs are taken into consideration as well. These lexicographer files with respect to verb synsets are {cognition}, {communication}, {emotion}, and {social} (table 1). The sense of a verb, for example "read", varies from context to context as well. We can also define a threshold value t_verb as the ratio of the number of senses in animacy files (table 1) to the number of total senses.

Table 1: Example of animate verb.

Unique beginners	Example of verb
{cognition}	Think, analyze, judge …
{communication}	Tell, ask, teach …
{emotion}	Feel, love, fear …
{social}	Participate, make, establish …

The training data from the Brown corpus consists of 10,134 words, 2,155 noun phrases, and 517 animacy entities. It shows that 24% of the noun phrases in the corpus refer to animacy entities whereas 76% of them refer to inanimacy ones. Threshold values can be obtained by training on the corpus and select the value when the accuracy rate (equation 4) reaches the maximum. Therefore t_noun and t_verb are achieved to be 0.8 and 0.9 respectively according to the observation in figure 4.

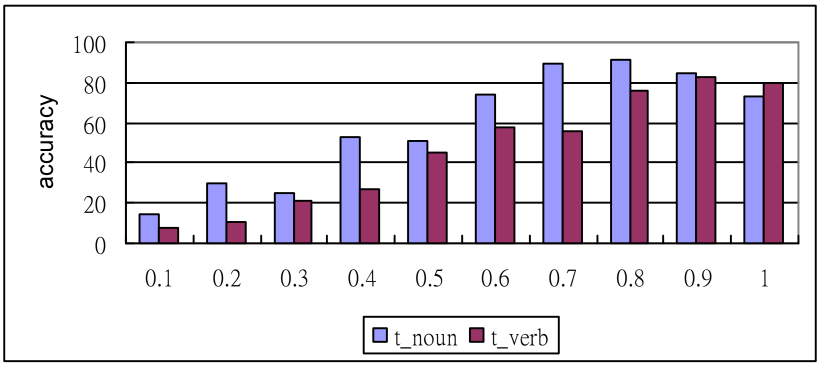
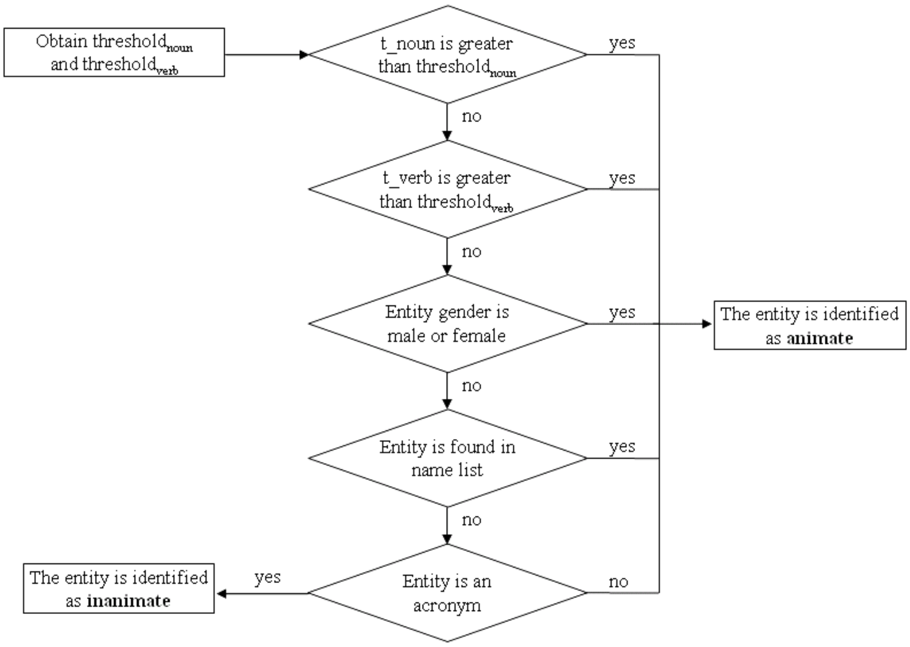


Figure 4: Thresholds of Animacy Entities.

The process of determining whether a noun phrase belong to animacy or not is described below：


Obtain $threshold_{noun}$ and $threshold_{verb}$
t_noun is greater than $threshold_{noun}$
yes
no
t_verb is greater than $threshold_{verb}$
yes
no
Entity gender is male or female
yes
The entity is identified as **animate**
no
Entity is found in name list
yes
no
Entity is an acronym
yes
The entity is identified as **inanimate**
no

2.2.7 Heuristic Rules

I. Syntactic parallelism rule

The syntactic parallelism could be an important clue while other constraints or preferences could not be employed to identify an unique unambiguous antecedent. It denotes the preference that correct antecedent has the same part-of-speech and grammatical function as the anaphor. The grammatical function of nouns can be subject, object or subject complement. The subject is the person, thing, concept or idea that is the topic of the sentence. The object is directly or indirectly affected by the nature of the verb. Words which follow verbs are not always direct or indirect objects. After a particular kind of verb, nouns remain in the subjective case. We call these subjective completions or subject complements.

For example:

The security guard took off **the uniform** after getting off duty.

He put **it** in the bottom of the closet.

The "**He**" (the subject) in the second sentence refers to "**The security guard**" which is also the subject of the first sentence. In the same way, the "**it**" refers to "**the uniform**" which is the object of the first sentence as well. Empirical evidence also shows that anaphors usually match their antecedents in their syntactic functions.

II. Semantic parallelism rule

This preference works with identifying collocation patterns in which anaphora took place. In this way, system can automatically identify semantic roles and employ them to select the most appropriate candidate. Collocation relations specify the relation between words that tend to co-occur in the same lexical contexts. It emphasizes that noun phrases which have the same semantic role as the anaphor are favored.

III. Definiteness rule

Definiteness is a category concerned with the grammaticalization of identifiability and nonidentifiability of referents. A definite noun phrase is a noun phrase that starts with the word "the", for example, "the young lady" is a definite noun phrase. Definite noun phrases which can be identified uniquely are more likely to be the antecedent of anaphors than indefinite ones.

IV. Mention Frequency rule

Iterated items in the context are regarded as the likely candidates for the antecedent of an anaphor. Generally, the high frequent mentioned items denote the focus of the topic as well as the most likely candidate.

V. Sentence recency rule

Recency information is employed by most of the implementations for anaphora resolution. In [Lappin, 94] the recency factor is the one with highest weight among a set of factors that influence the choice of antecedent. The recency factor states that if there are two (or more) candidate antecedents for an anaphor and all of these candidates satisfy the consistency restrictions for the anaphor (i.e. they are qualified candidates) then the most recent one (the one closest to the anaphor) is chosen. In [Mitkov et al., 01], the average distance (in sentences) between the anaphor and the antecedent is 1.3, and the average distance in noun phrases is 4.3 NPs.

VI. Non-prepositional noun phrase rule

A noun phrase not contained in another noun phrase is favored as the possible candidate. This condition can be explained from the perspective of functional ranking: subject > direct object > indirect object. A noun phrase embedded in a prepositional noun phrase is usually an indirect object.

VII. Conjunction constraint rule

Conjunctions are usually used to link words, phrases and clauses. If the candidate is connected with the anaphor by a conjunction, they can hardly have anaphora relation.

For example:

Mr. Brown teaches in a high school. Both **Jane** and **he** enjoy watching the movies in the weekend.

2.3 The Brown Corpus

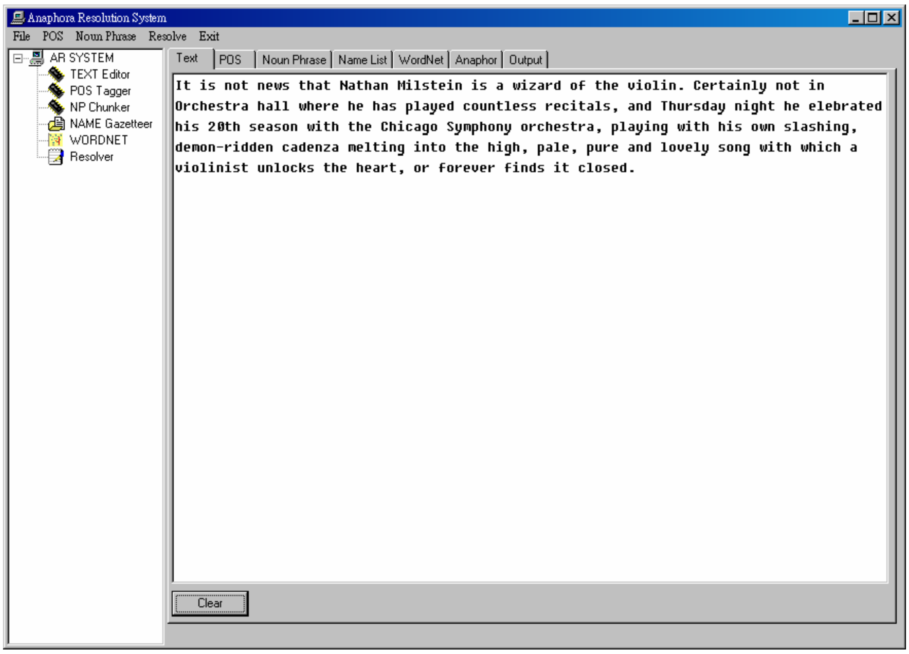
The training and testing text are selected randomly from the Brown corpus. The Corpus is divided into 500 samples of about 2000 words each. The samples represent a wide range of styles and varieties of prose. The main categories are listed in figure 5.

A. Press: Reportage	J. Learned
B. Press: Editorial	K. General Fiction
C. Press: Reviews	L. Mystery and Detective Fiction
D. Religion	M. Science Fiction
E. Skills and Hobbies	N. Adventure and Western Fiction
F. Popular Lore	P. Romance and Love Story
G. Biography, Memoirs, etc.	R. Humor
H. Miscellaneous	

Figure 5: Categories of the Brown corpus.

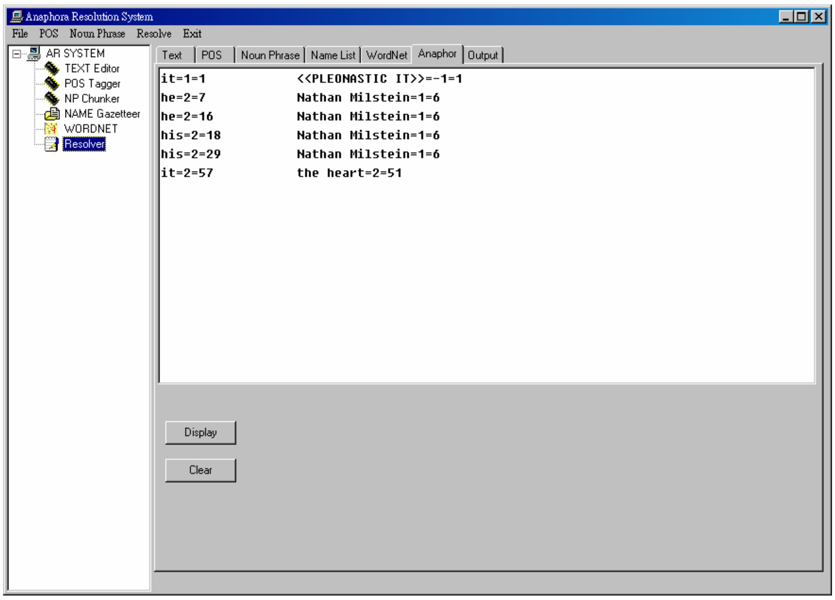
2.4 System functions

The main system window is shown in figure 6. The text editor is used to input raw text without any annotations and shows the analyzed result. The POS tagger component takes the input text and outputs tokens, lemmas, most likely tags and the number of alternative tags. NP chunker makes use of finite state machine (FSM) to recognize strings belong to a specified regular set.



It is not news that Nathan Milstein is a wizard of the violin. Certainly not in Orchestra hall where he has played countless recitals, and Thursday night he elebrated his 20th season with the Chicago Symphony orchestra, playing with his own slashing, demon-ridden cadenza melting into the high, pale, pure and lovely song with which a violinist unlocks the heart, or forever finds it closed.

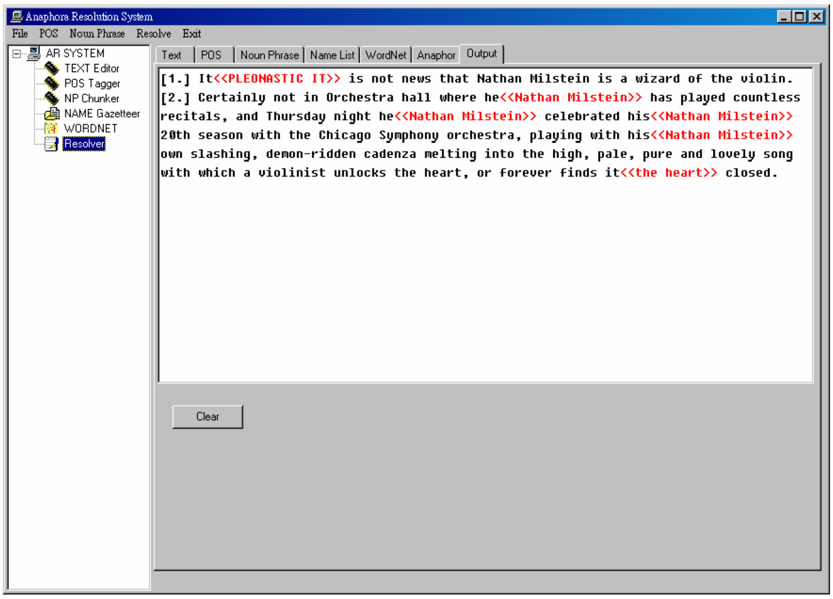
Figure 6: The main system window.



```
it=1=1          <<PLEONASTIC IT>>=-1=1
he=2=7          Nathan Milstein=1=6
he=2=16         Nathan Milstein=1=6
his=2=18        Nathan Milstein=1=6
his=2=29        Nathan Milstein=1=6
it=2=57         the heart=2=51
```

Figure 7: Anaphora pairs.

After performing the selection procedure, the most appropriate antecedent is chosen to match each anaphor in the text. Figure 7 illustrates the result of anaphora pairs in each line in which sentence number and word number are attached to the end of the entities. For example, the "it" in the first word of the first sentence denotes a pleonastic-it and the other "it" in the 57th word of the second sentence refers to "the heart". Figure 8 shows the original text input with antecedent annotation followed each anaphor in the text. All the annotations are highlighted to make it easy to carry out the subsequent testing purposes.



```
[1.] It<<PLEONASTIC IT>> is not news that Nathan Milstein is a wizard of the violin.
[2.] Certainly not in Orchestra hall where he<<Nathan Milstein>> has played countless
recitals, and Thursday night he<<Nathan Milstein>> celebrated his<<Nathan Milstein>>
20th season with the Chicago Symphony orchestra, playing with his<<Nathan Milstein>>
own slashing, demon-ridden cadenza melting into the high, pale, pure and lovely song
with which a violinist unlocks the heart, or forever finds it<<the heart>> closed.
```

Figure 8: Anaphor with antecedent annotation.

3. Experimental Results and Analysis

The proposed system is developed in the following environment (table 2).

Table 2: System environment.

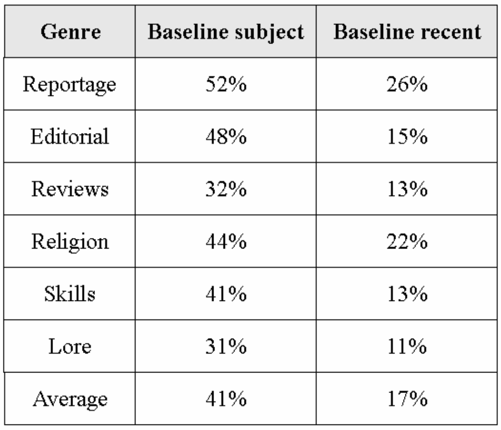
Operating System	Microsoft Windows 2000 Advanced Server
Main Processor	AMD Athlon K7 866MHZ
Main Memory	256 MB SDRAM
Graphic Card	NVIDIA Geforce2 Mx 32M
Programming language	Borland C++ Builder 5.0

The evaluation task is based on random texts selected from the Brown corpus of different genres. There are 14,124 words, 2,970 noun phrases and 530 anaphors in the testing data. Two baseline models are set up to compare the effectiveness with our proposed anaphora resolution (AR) system. The first baseline model (called baseline subject) performs the number and gender agreement between candidates and anaphors, and then chooses the most recent subject as the antecedent from the candidate set. The second baseline model (called baseline recent) performs a similar procedure but it selects the most recent noun phrase as the antecedent which matches the number and gender agreement with the anaphor. The measurement can be calculated as follows:

$$Success\ Rate = \frac{number\ of\ correctly\ resolved\ anaphors}{number\ of\ all\ anaphors} \qquad (5)$$

In the result of our experiment baseline subject (table 3), there are 41% of antecedents can be identified by finding the most recent subject, however, only 17% of antecedents can be resolved by means of selecting the most recent noun phrase with the same gender and number agreement to anaphors.

Table 3: Success rate of baseline models.



Genre	Baseline subject	Baseline recent
Reportage	52%	26%
Editorial	48%	15%
Reviews	32%	13%
Religion	44%	22%
Skills	41%	13%
Lore	31%	11%
Average	41%	17%

Figure 9 presents the distribution of sentence distance between antecedents and anaphors. The value 0 denotes intra-sentential anaphora and other values mean inter-sentential anaphora. Figure 10 shows the average word distance distribution with respect to each genre. The identification of pleonastic-it can be achieved to 89% accuracy (table 4).

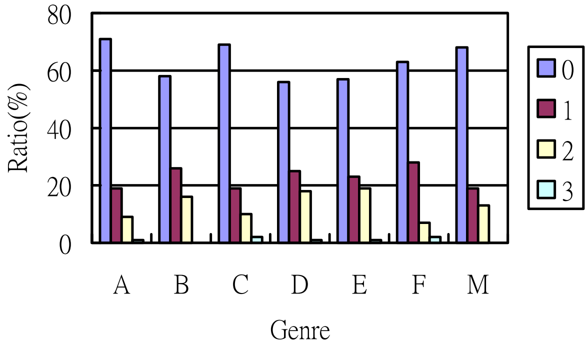


Figure 9: Referential sentence distance distribution.

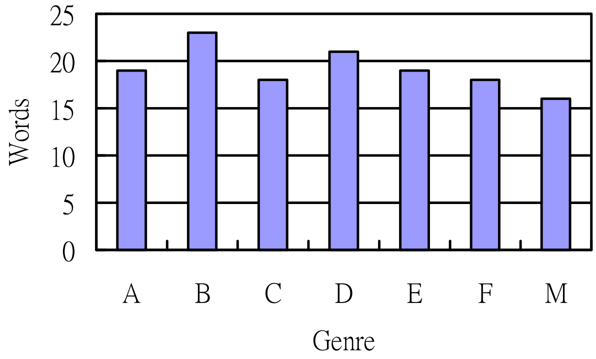


Figure 10: Referential word distance distribution.

Table 4: Pleonastic-it identification.

	Number of Anaphora	Anaphoric expression	Number of Pleonastic-it	Ratio of Pleonastic-it	Accuracy of identification
Total	530	483	47	9%	89%

The evaluation result of our system which applies animacy agreement and heuristic rules for resolution is listed in table 5. It also contains the results for each individual genre of testing data and the overall success rate reaches 77%.

Table 5: Success rate of AR system.

Genre	Words	Lines	NPs	Anims	Anaphors	Success Rate
Reportage	1972	90	488	110	52	80%
Editorial	1967	95	458	54	54	80%
Reviews	2104	113	480	121	92	79%
Religion	2002	80	395	75	68	76%
Skills	2027	89	391	67	89	78%
Lore	2018	75	434	51	69	69%
Fiction	2034	120	324	53	106	79%
Total	14124	662	2970	531	530	77%

4. Conclusion and Future Work

In this paper, the WordNet ontology and heuristic rules are adopted to the anaphora resolution. The recognition of animacy entities and gender features in the discourses is helpful to the promotion of resolution accuracy. The proposed system is able to deal with intra-sentential and inter-sentential anaphora in English text and includes an appropriate treatment of pleonastic pronouns. From experiment results, our proposed method is comparable with prior works using fully parsing of the text. In contrast to most anaphora resolution approaches, our system benefits from the recognition of animacy occurrence and operates in fully automatic mode to achieve optimal performance. With the growing interest in natural language processing and its various applications, anaphora resolution is worth considering for further message understanding and the consistency of discourses.

Our future work will be directed into following studies:

1. Extending the set of anaphor being processed:
   This analysis aims at identifying instances (such as definite anaphor) that could be useful in anaphora resolution.
2. Resolving nominal coreference:
   The language resource WordNet can be utilized to identify the coreference relation on the basis of synonymy/hypernym/hyponym relation.

References

Aarts Jan, Henk Barkema and Nelleke Oostdijk (1997), “The TOSCA-ICLE Tagset: Tagging Manual”, TOSCA Research Group for Corpus Linguistics.

Baldwin, Breck (1997), “CogNIAC: high precision coreference with limited knowledge and linguistic resources”, Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution, pp. 38-45.

Bontcheva, Kalina, Marin Dimitrov, Diana Maynard and Valentin Tablan (2002), “Shallow Methods for Named Entity Coreference Resolution”, Proceedings of TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES (TALN), pp. 24-32.

Cardie, Claire and Kiri Wagstaff (1999), “Noun Phrase Coreference as Clustering”, Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Chen, Kuang-hua and Hsin-Hsi Chen (1994), “Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation”, Proceedings of the 32nd ACL Annual Meeting, 1994, pp. 234-241.

Dagan, Ido and Alon Itai (1990), “Automatic processing of large corpora for the resolution of anaphora references”, Proceedings of the 13th International Conference on Computational Linguistics (COLING'90), Vol. III, 1-3, Helsinki, Finland.

Denber, Michel (1998), “Automatic resolution of anaphora in English”, Technical report, Eastman Kodak Co.

Evans, Richard and Constantin Orasan (2000), “Improving anaphora resolution by identifying animate entities in texts”, In Proceedings of DAARC-2000.

Ge, Niyu, John Hale and Eugene Charniak (1998), “A Statistical Approach to Anaphora Resolution”, Proceedings of the Sixth Workshop on Very Large Corpora (COLING-ACL98), pp.161-170.

Kennedy, Christopher and Branimir Boguraev (1996), “Anaphora for everyone: Pronominal anaphora resolution without a parser”, Proceedings of the 16th International Conference on Computational Linguistics, pp.113-118.

Lappin, Shalom and Herbert Leass (1994), “An Algorithm for Pronominal Anaphora Resolution”, Computational Linguistics, Volume 20, Part 4, pp. 535-561.

Miller, George (1993), “Nouns in WordNet: A Lexical Inheritance System”, Journal of Lexicography, pp. 245-264.

Mitkov, Ruslan (1998), “Robust pronoun resolution with limited knowledge”, Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference Montreal, Canada. pp. 869-875.

Mitkov, Ruslan (1999), “Anaphora Resolution: The State of the Art”, Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution)

Mitkov, Ruslan and Catalina Barbu (2001), “Evaluation tool for rule-based anaphora resolution methods”, Proceedings of ACL'01, Toulouse, 2001.

Mitkov, Ruslan, Richard Evans and Constantin Orasan (2002), “A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method”, In Proceedings of CICLing- 2000, Mexico City, Mexico.

Wang, Ning, Chunfa Yuan, K.F. Wang and Wenjie Li (2002), “Anaphora Resolution in Chinese Financial News for Information Extraction”, Proceedings of 4th World Congress on Intelligent Control and Automation, June 2002, Shanghai, pp.2422-2426.

文件自我擴展於自動分類之應用

Application of Document Self-Expansion to Text Categorization

曾元顯 ， 莊大衛
輔仁大學 圖書資訊學系
台北縣新莊市中正路 510 號 242
TEL: 02-29031111 ext 2333, FAX: 02-29017405
tseng@lins.fju.edu.tw

摘要：

近幾年自動分類的研究顯示，訓練文件越多，分類效果越好。然而，訓練文件的獲得需要花費相當的人力與時間，此一成本常造成使用單位導入自動分類流程的困擾。針對此問題，本文提出一種文件自我擴展方法，在沒有利用任何額外資源的情況下，全自動的增加訓練文件，以期達到降低成本、提高成效的目的。經由兩種分類測試集以及兩種分類器的實驗驗證，顯示此方法在原始訓練文件數越少時，其改進的效果越明顯。而且此改進方法，乃策略層面上的技巧，與分類器無關，亦即任何一種分類器都可以運用上面的技巧來增強其分類效果。

關鍵詞：文件分類，機器學習、文件擴展、中文、資訊檢索

一、前言

「文件主題分類」或簡稱「文件分類」（document classification or text categorization）是指依文件的「內容主旨」給定「類別」（class or category）的意思。文件分類的目的，在對文件進行分門別類的加值處理，使得文件易於管理、利用。分類後的文件，可提供使用者依主題查找文件而不受文件用詞的限制。另外，文件分類後，還可顯示館藏文件的主題分佈與範圍，對館藏文件的後續徵集與使用者的文件使用情形，提供重要的決策參考。

近年來，拜資訊技術普及運用之賜，各個企業與機構的數位文件不斷累積，數量大到難以有效的管理與利用，文件分類的需求也就遽然而生。為此，如何利用自動化的技術，快速有效的協助人工分類，來應付大量暴增的分類需求，是現今資訊服務與知識管理的重要課題。

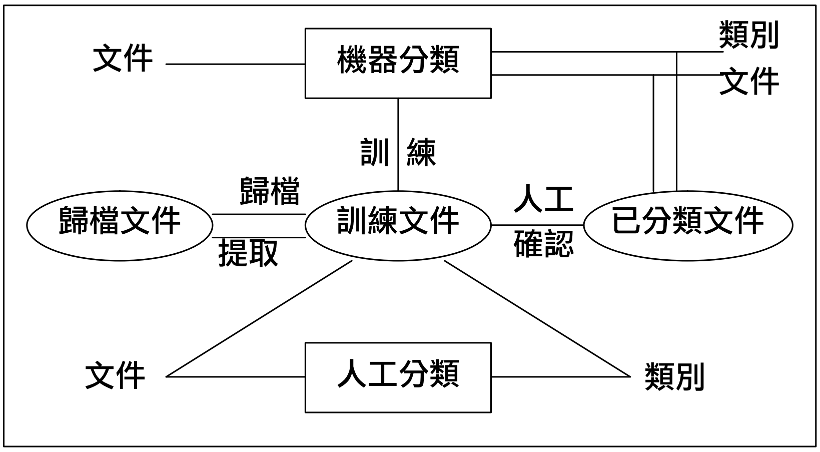
文件分類，需要瞭解文件的主題大意，才能給定類別，因此是相當高階的知識處理工作。要將文件分類自動化，必須先整理出分類時的規則，電腦才能據以執行。然而，有效的分類規則通常難以用人工分析歸納獲得。因此，機器在做

自動分類之前，還必須加以訓練，使其自動學習出人工分類的經驗與知識。

所謂訓練，就是讓機器去分析一堆「訓練文件」，如圖一所示。訓練文件記錄了人工進行文件分類的知識，這種知識相當隱晦，只是一堆（文件=>類別）的對應記錄。機器在反覆的閱讀文件以及其標示的類別後，自動歸納出一些對應規則，使其下次看到類似的文件時，可以給出適當的類別。

機器分類雖然速度快、節省大量人力，缺點則是需要事先準備相當數量的訓練文件，機器才能做出有效的分類。近幾年研究自動分類的經驗顯示，訓練文件越多，分類效果越好。因此，各個機構在導入自動分類時，需要事先準備一定數量的訓練文件。然而，訓練文件的獲得需要花費相當的人力與時間，此一成本常造成了導入自動分類流程的困擾。

此外，即便準備好了訓練文件，可能由於各個類別在整個事件機率分佈上的自然現象，個別類別的訓練篇數分佈常有極不平均的現象，亦即訓練篇數多的類別只有少數幾類，而訓練篇數少的類別則佔大多數的類別。以學術界常用做分類研究的 Reuter-21578 測試集（test collection）為例，共 90 個類別、7770 篇訓練文件，最大的 10 類，就佔了 75% 的訓練文件量、平均每類有 719 篇訓練文件，而最小的 20 類，只佔 0.5% 的訓練文件量，平均每類只有 2 篇訓練文件。這種情形在很多真實生活的測試集中常常見到，不是 Reuters 文件獨有的現象 [1]。



圖一：自動分類流程圖。

綜上所述，訓練文件數不足，乃導入自動分類時常碰到的現象。為了維持有效的分類，同時又要降低訓練文件的獲得成本，一個直覺的想法，是以自動的方法來增加訓練文件，使得即便只有少量的訓練文件時，自動分類還能達到一定的效果。這個想法的好處是它跟任何方法都無關，因此可適用於任何既有的分類方法上。

本文便是在少量訓練文件的環境下，探討如何進行有效自動分類的問題。下一節將簡略的分析過去的相關研究。第三節則介紹本文採用的「文件自我擴展」的方法，來增加訓練文件。第四節描述驗證此方法的實驗資料與環境。第五節報告實驗的結果與心得。最後一節總結本文的結論並提出未來可能的研究方向。

二、相關研究

過去數年，國內外有關文件自動分類的研究相當豐富 [2-5]。很多研究嚐試提出不同的方法，讓自動分類達到更高的成效。然而，針對訓練文件量少的情況，來提升自動分類成效的研究，則相對稀少。比較接近的研究題目有 expectation maximization（EM）[6-7] 與 co-training [8]。EM 方法是將人工尚未分類的文件以機器自動分類完後，就視其為已分類文件，而拿來訓練。這過程反覆的進行一直到分類器收斂為止。如此，在沒有人工介入的情形下，用少量人工準備的訓練資料，就可以訓練出初步的分類器，而這個分類器可以用來產生更多的「訓練文件」，來訓練分類器本身。當然這些機器產生的「訓練文件」，其分類錯誤的情形，可能較人工準備的真正訓練文件為高，依此訓練出來的分類器，有可能會不甚準確。然而即便人工準備的訓練文件也不能保證百分之百正確的（不同的人對同一篇文件會給出不同的類別，此種不一致的現象並不少見），因此，只要機器產生的「訓練文件」品質不太差，這種自我訓練大都可以增進分類器的準確度。

Co-training 的方法則是假設文件的特徵可以分成兩組獨立的集合，每一組集合可以訓練出一個分類器。每個分類器都以人工準備好的訓練資料以及個別的特徵集合訓練出初步的分類器。對於尚未分類的文件，每一個分類器都對每一個類別分出一些文件，然後將這些自動分好的文件視為「訓練文件」再去訓練這兩個分類器，如此不斷重複，直到所有未分類的文件都給定類別為止。這個作法是讓某個分類器做出來的訓練文件用來訓練另一個分類器，如此反覆互相訓練，最後彼此的分類準確度可能就越來越好。Co-training 在漸進式地互相自我訓練出兩個分類器後，真正進行文件分類時，再將這兩個分類器的結果融合，作為該文件的分類結果。

上述這兩種方法都可以從少量的訓練文件開始，利用大量多餘的未分類文件，得到不錯的分類成效。CMU 大學的 Nigam 與 Ghani 兩人的實驗中，曾利用 12 篇有標示類別的文件以及 776 篇未標示類別的文件，對 263 篇網頁文件做「課程網頁」與「非課程網頁」的分類，結果 co-training 的錯誤率為 5.4%，EM 的錯誤率為 4.3%，而如果以傳統的方法，且利用到 12+776= 788 篇有標示類別的文件做訓練，則錯誤率為 3.3% [8]。顯見 co-training 與 EM 方法，真的可以用少量訓練文件，就可達到相當好的成效。可惜 EM 與 Co-training 的計算量都很大，每次反覆訓練一次，等於又做了一次傳統分類方法的訓練。

上述兩種方法的另一個缺點，是當只有少量的訓練文件，而沒有大量多餘的未分類文件可利用時，就無法適用。這個問題會發生在前述類別分配不平均的大量小類別上。也就是說，即使想要以人工蒐集、準備資料，也會因文件出現實例太少，只得出少量的訓練文件，而沒有多餘的未分類文件可用。因而，會有無法運用 EM 或 Co-training 的情形。

三、文件自我擴展

文件自動分類的研究，已觀察到：「訓練文件越多，分類成效越好」的現象。因此在導入自動分類機制的時候，導入單位常常會碰到一個難題：要準備多少訓練文件才夠？準備得太少，效果不好；準備得太多，要投入很多人力、時間成本。如果只需要準備少量訓練文件，就可以得到宛如有很多訓練文件才能獲得的分類成效，豈不兩全其美。

為解決此一問題，本文採用一個策略，就是以自動化的方式，來獲得更多的訓練文件，以達到降低成本、提高成效的目的。跟前述相關研究不同的是，此方法只「擴展」既有的訓練文件，沒有利用到其他的資源，包括未分類文件，因此可跟其他方法一起運用，而不相衝突。

擴展（expansion）的概念，在資訊檢索領域裡常常運用。例如，對查詢而言，有「查詢擴展」（query expansion）的方法 [9]，運用在主題檢索上。其作法是增加一些查詢詞彙或修改原查詢詞彙的權重，來擴增原查詢條件，以期能獲得更佳的查詢結果。對文件而言，也有「文件擴展」的方法，運用在語音文件（如口語播報新聞）的檢索 [10]。其作法是將語音辨識成文字，再以以原查詢條件查詢乾淨的平行文件（與語音文件內容近似的文字文件，如語音新聞文字稿或同一天的新聞文字），以此查詢結果作相關回饋或查詢擴展，再運用到語音辨識過的文件查詢上，以便降低語音辨識錯誤的影響。本文提出的方法，類似資訊檢索的文件擴展法，但不需要額外的平行文件，只從原文件本身擴展，因此稱為「文件自我擴展」法。

這裡的文件自我擴展作法，是對每一個類別，從其現有的訓練文件中，擷取每篇文件的部分片段，組成新的文件，以增加該類別的訓練文件數。理想上，在「擷取每篇文件的部分片段」方面，應該要擷取可以彰顯文件主題的片段，例如利用自動摘要技術擷取文件的重要片段；在「組成新的文件」時，簡單的作法，是像遺傳演算法的基因重組那樣，將同類中數篇文件的標題或摘要，拿來交叉組合，做成新的文件。雖然這樣組出來的新文件，對人而言，也許語句不連貫，沒有實質的義意，但重要的類別用詞，就會重新分佈，而可能有助於分類器的學習、訓練。最簡單的效果，就是重要的詞彙在該類別的不同文件中重複出現了，而不重要的詞彙，則因為較少被選出來而降低其在分類中能夠扮演的角色。

基於上述的想法，本文提出兩種文件擴展法：一是摘要擴展法，另一是詞彙擴展法。

摘要擴展法的構想，是將每篇訓練文件以自動摘要法將其句子按照重要性排序，排序在前面的句子才視為該篇文件的摘要，然後依此摘要再組成新文件。在此，自動摘要法可以選擇只取文件的標題，那麼此方法將簡化成「標題擴展法」。當然，也可以用關鍵詞彙出現的次數、句子本身出現的位置，或句中出現特殊詞彙（如：「因此」、「所以」、「結論是」）的資訊，或以機器學習的方式，來一起決定句子的重要性。

在後面的實驗驗證裡，我們選擇較簡單而運用性較廣的方法，即只計數句子中包含關鍵詞彙的出現次數，來決定該句子的重要性。例如，若某個句子包含A、B、C 三個關鍵詞，且他們在整篇文件中出現的次數分別為 2、4、3,，那麼該句子的重要性為 2+4+3=9。文件中的每個句子都以此計算其重要性，再由大到小排序。

上述所謂關鍵詞彙，是以 Tseng 的演算法求出的最大重複字串（maximally repeated string）[11]，做為該文件的關鍵詞彙。此方法假設文件的主題詞彙會重複出現，但並非所有的重複字串都是有用的關鍵詞，它們必須是最長的，或是出現頻率最高的，因此稱為最大重複字串。例如前兩句中「最大重複字串」出現了二次，而「重複字串」出現了三次，那麼這兩個詞都會被擷取出來。但「大重複字」此字串也出現二次，但因它是「最大重複字串」的完全子字串，所以不會被擷取出來成為關鍵詞。

一旦每篇文件的句子按上述方式排列後，便隨機的選取同一類別中任一文件的前 S 個句子，來累積出新文件，一直到該新文件的長度為所有文件的平均長度對為止。在後面的實驗驗證中，我們只取文件的第一個句子，期使其累積出跟舊文件差異較大的新文件。

在摘要擴展法裡，可能會引入不相干的詞彙在新文件裡，因此，在詞彙擴展法的構想裡，便希望只擷取出跟該類別有關的特徵詞彙來擴展。Yang 等人曾比較了五種特徵詞選取方式，其實驗結果顯示，在五種方法裡，Chi-square 與 Information Gain 同樣為最有效的特徵詞選取方法 [12]。若以表一中詞彙 T 在類別 C 的出現篇數分佈表示，Chi-square 計算某個詞 T 與某類別 C 的相關性如下：

$$\chi^2(T,C) = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP})^2}{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})(\mathrm{FN}+\mathrm{TN})}$$

對某一類別用 Chi-square 選詞，就是將所有的詞彙依照 Chi-square 值做排序，然後選出其中 Chi-square 值最大的前 N 個詞。

表一：詞彙 T 在類別 C 中的出現篇數分佈表

	詞彙 T		
類別 C		出現篇數	沒出現篇數
	是	TP	FN
	否	FP	TN

然而，Ng 等人的研究觀察顯示 [13]，Chi-square 會同時選出正相關與負相關的詞彙，因為正相關與負相關的詞都因 Chi-sqaure 的二次方計算，使得其值都變成正數，造成不出現在類別 C 中的詞彙，也會被選為類別 C 的特徵詞。這對文件分類是沒有幫助的。因為大部分的分類方法，都是依賴文件中出現某個詞，來計算其權重，而將文件分為某個類別，而不是依賴文件中沒有出現某個詞，來將文件分為該類別。因此，Ng 等人提倡改用相關係數（即單邊的 Chi-square）

來選詞：

$$Co(T,C) = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})(\mathrm{FN}+\mathrm{TN})}}$$

如此，與類別負相關的詞，會因為變成負數，被降低排序，而比較不可能被選為類別的特徵詞。以某一只有「營建類」與「非營建類」兩類的分類文件集為例，表二顯示其 Chi-square 與相關係數選出的前六個詞。Chi-square 選出的詞彙中，對「營建類」而言，「設備」、「公告」兩詞事實上與「營建類」為負相關，即此兩個詞在營建類的文件中極少出現，反而在「非營建類」的文件中極常出現。同理「工程」、「改善」與「非營建類」為負相關，且其相關係數分別為 -0.7880 與 -0.3169，平方後，恰為 Chi-square 值：0.6210 與 0.1004。表二顯示，相關係數選出的正相關詞彙，較符合分類需要的特徵詞彙。

表二：Chi-square 與相關係數選出的詞彙比較表

Chi-square 選詞		相關係數選詞	
營建類	非營建類	營建類	非營建類
工程 0.6210	工程 **0.6210**	工程 0.7880	設備 0.2854
改善 0.1004	改善 **0.1004**	改善 0.3169	電腦 0.2231
設備 **0.0815**	設備 0.0815	路面 0.2009	採購 0.2231
公告 **0.0425**	電腦 0.0498	道路 0.1764	公告 0.2062
路面 0.0404	採購 0.0498	新建 0.1629	系統 0.1764
道路 0.0311	公告 0.0425	土城市 0.1563	購置 0.1484

當選出的類別詞彙很少時，相關係數與 Chi-square 選出來的詞彙會有較大的差異。但當選出的詞彙數較多時，此兩種方法得到的特徵詞彙，差異就縮小了。這可以解釋為何 Yang 等人利用 Chi-sqaure 選詞，還可以得到不錯的結果。但這裡我們要用類別特徵詞來增加文件數，因為擴增的文件不會太多，因此選擇以相關係數來選詞較為妥當。

在詞彙擴展法裡，類別的特徵詞以相關係數計算、排序後，取前 K 個詞，以每個詞就視為一份新文件的方式，來增加該類別的訓練文件。

四、實驗設計

為了瞭解上述想法的效果，本文以中文文件的分類來驗證。過去的分類研究顯示，不同的分類法對不同的測試集有不同的表現。單獨以某種分類法在某種測試集上做實驗，容易產生偏向（bias）的實驗結論。因此，本文特別以兩種分類法在兩種測試集上進行交叉驗證。

這兩種測試集都於 2001 年時得自 PC home Online 的線上文件。一是 PC home 蒐集的新聞，共 12 類，為方便爾後的討論，稱其為 News 測試集，其每篇

文件的平均長度為 9.87 個句子。表三顯示其類別名稱、訓練文件與測試文件的篇數。另一測試集是 PC home 製作的網頁分類描述，共 26 類，全都是「網路與電腦」類別底下的細類，稱其為 WebDes 測試集，其每篇文件的平均長度為 2.10 個句子。表四顯示其類別名稱、訓練文件與測試文件篇數。

News 測試集全部的文件數有 914 篇，最大類與最小類的篇數相差約 30 倍。WebDes 測試集全部文件數有 1686 篇，最大類與最小類的篇數相差約 90 倍。此兩測試集的每一篇文件都是單一分類，亦即沒有任何一篇文件分在兩個或兩個以上的類別，且每一類的訓練篇數與測試篇數大多維持在 7：3 的比例，只有當該類實例太少時，才無法維持此比例。

表三：News 測試集類別名稱與文件篇數

編號	類別	訓練	測試	合計	編號	類別	訓練	測試	合計
1	產業	232	99	331	7	地方	29	12	41
2	財經	117	50	167	8	科技	18	7	25
3	政治	78	33	111	9	體育	12	4	16
4	社會	53	22	75	10	醫藥	10	4	14
5	生活	40	17	57	11	文教	10	4	14
6	娛樂	38	15	53	12	休閒	7	3	10

表四：WebDes 測試集類別名稱與文件篇數

編號	類別	訓練	測試	合計	編號	類別	訓練	測試	合計
1	網頁設計工作室	262	112	374	14	搜尋引擎連結	25	10	35
2	網頁設計教學	194	82	276	15	網站宣傳	24	10	34
3	電子賀卡	140	59	199	16	搜尋引擎	17	6	23
4	國內網站網頁搜尋	66	27	93	17	網路文化	16	6	22
5	駭客	53	22	75	18	Proxy	14	5	19
6	主題搜尋	53	22	75	19	Plug-in	12	5	17
7	ISP	49	21	70	20	固接專線	11	4	15
8	網域註冊	45	18	63	21	瀏覽器	11	4	15
9	網路資訊討論	40	17	57	22	電子商務	10	4	14
10	國外網站網頁搜尋	36	15	51	23	檔案搜尋	6	2	8
11	網路調查	35	14	49	24	BBS 文章搜尋	5	2	7
12	網路安全	33	14	47	25	Intranet	4	2	6
13	網站評鑑	27	11	38	26	電子郵件搜尋	2	2	4

在分類方法方面，近年來常被驗證效果最好的分類法為：SVM（Support Vector Machine）與 KNN（K-Nearest Neighbor），本文選擇此兩方法來實驗。我們選擇 Thorsten Joachims 製作的 SVMlight 作為 SVM 分類器 [14-15]。經過一些

測試，我們以 SVMlight 的預設環境做分類（線性分類），因為這樣效果最好，並根據 SVMlight 的使用說明以其類別輸出值的正負號做為類別分類的依據。這是因為 SVM 是二元分類器（binary classifier），因此有 C 個類別要分類時，就要做出 C 個 SVM 分類器，當某一類別的分類器其輸出值為正時，就將文件分為該類別。但我們發現很多文件都沒有任何類別分出來（所有的類別其輸出值均為負數），因此我們改變分類方法，即取類別輸出值最大者為分類的類別。在我們的實驗中，這樣改變之後，都能夠增進 SVMlight 的分類成效。

至於 KNN 分類器方面，我們實作了一套 KNN 分類系統。由於 KNN 分類器的成效非常依賴於文件相似度的正確計算，我們特別以下面公式計算：

$$Sim(d_i, q_j) = \frac{\sum_{k=1}^{T} d_{i,k} q_{j,k}}{(bytesize_{d_i})^{0.375}\sqrt{\sum_{k=1}^{T} q_{j,k}^2}}$$

其中 q_j 是輸入文件，d_i 是訓練文件，$q_{j,k}$（$d_{i,k}$）是詞彙 k 的權重，以詞頻（term frequency）及反相篇數（inverse document frequency）來加權計算，而 $bytesize_{dj}$ 為文件的長度，以位元（byte）為單位。此相似度公式乃 Singhal 等人為改進 Cosine 相似度公式的缺點而提出的，在 Singhal 等人以及某些中文 OCR 文件的（主題）檢索實驗中，bytesize 的確比 Cosine 的成效更好 [16-18]。

KNN 分類法中每一個類別的分數，以下面公式計算：

$$y(q, c_j) = \sum_{d_j \in KNN} Sim(d_i, q)\, y(d_i, c_j)$$

其中 y(d,c)為 1 或 0 的值，代表訓練文件 d 是否為類別 c，而 KNN 表示跟文件 q 最相近的 K 篇訓練文件的集合。在後面的實驗中，K 都取 20，並以 y 值中最大者的類別 c，做為文件 q 的類別。

文件分類的成效，受很多因素影響 [2-3]，其中之一就是用來分類的特徵詞彙個數與選擇方式。我們試驗了數種選擇方式以及詞彙個數，把效果最好的用在後面的實驗中。對 SVM 分類器，在兩個測試集中，我們都取文件篇數大於 1 且 Chi-square 值大於 0 者為特徵詞。對 KNN 分類器，在 News 測試集中，所有的詞都來拿分類，在 WebDes 測試集中，則只有文件篇數大於 1 的詞彙，才用來分類。至於這些文件詞彙，是以 Tseng 描述的方法從文件中所取出來的詞彙 [19]，包括字典裡的詞、不在字典裡的最大重複詞，以及少數無法斷字的單字詞。

為了測試少量訓練文件時，本文方法的效果，我們對這兩個測試集的訓練文件做 5%、10%、20%、40%以及 100%的縮減取樣。亦即，對每一個類別，將其訓練文件分成 100/p 個等份，其中 p 代表前述的百分比，然後取其第一等份的文件作為縮減後的訓練資料。若原訓練文件本就不多，則第一等份至少必須包含 1 篇訓練文件。因此，縮減後每一類都還有訓練文件，而測試文件則保持原來的不變動。

在前述文件擴展的方法裡，對每一類別擴增多少文件數，乃一實驗參數。在此，我們試驗三種文件擴增的數量，分別求其成效，加總平均後，再與原來沒

有做文件擴展的成效作比較，以得到比較穩定的結果。

此擴展的新文件數量與原訓練文件數有關。對每一個縮減的測試集，此三個數量為 $np^{0.5}$，其中 p 代表前述的縮減百分比，而 n 在 News 測試集中分別為 10、20、30，在 WebDes 測試集中，分別為 40、60、80。這是因為 News 與 WebDes 訓練文件數不同，因此選用的擴增文件數也不同。例如，對縮減成 5%的 News 訓練文件，用來實驗的三種擴增文件數分別為 2（$=10*0.05^{0.5}$）、4、6，對 WebDes 而言，其擴增文件數分別為 8、13、17。

根據上述方式決定擴增文件數 E 之後，只有當該類的文件數不足 E 時，才擴展其文件數達 E 篇文件。

在成效評估方面，不同的度量方式有各自不同的強調對象，因而容易導致偏向（bias）的結論。本文以 MicroF 以及 MacroF 值同時呈現分類的效果，其計算方式如下：

$$MicroF = \frac{2 \times \sum_{i=1}^{C} TP_i}{2 \times \sum_{i=1}^{C} TP_i + \sum_{i=1}^{C} FP_i + \sum_{i=1}^{C} FN_i}$$

$$MacroF = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i}$$

其中 C 是類別總數，i 代表某一類別，而 TP_i（True Positive）、FP_i（False Positive）、FN_i（False Negative）類似表一的意義，分別代表：是類別 i 而且也正確分為類別 i 的篇數、不是 i 類卻分為 i 類的篇數、是 i 類卻沒有分為 i 類的篇數。此 F 值乃從精確率（P）與召回率（R）的常見公式：$F = 2PR/(P+R) = 2TP/(2TP+FP+FN)$ 演化而來，其中 $P = TP/(TP+FP)$、$R = TP/(TP+FN)$。

由於 MicroF 是全部文件一起累加統計，不分類別，因此容易受到大類別（佔大多數文件）表現好壞的影響。相對的，MacroF 考慮每個類別的成效後再做平均，因此容易受到大量的小類別影響。將兩種平均數據都報告出來，可以瞭解大多數文件的分類效果（MicroF），以及大多數類別的分類效果（MacroF）。

五、實驗結果

實驗結果顯示於表五與表六。第一欄顯示縮減後的訓練文件量，第二欄與第五欄為運用 KNN 與 SVM 對測試文件做分類得到的數值，其他欄為運用 KNN 或 SVM 再加上文件擴展而獲得的分類成效值，其中 s 與 t 分別代表摘要擴展法與詞彙擴展法。表中粗體的數值表示文件擴展法比原始方法效果較好的情形。從表中可知：

1. 訓練文件越多，效果越好。
2. 訓練文件越少時，文件擴展法的改進成效越明顯。
3. 文件擴展法對 MacroF 的改進效果，比 MicroF 高。
4. 摘要擴展法與詞彙擴展法的效果各有優劣，但詞彙擴展法計算量較低，

因為它只增加少數幾個詞到原始的訓練文件中而已（每一篇新增的文件裡只包含一個詞）。

5. 訓練文件夠多時，再怎們運用改進策略，成效有限。必須根本的改變分類方法，才有可能大幅度地提昇成效。
6. 不同的分類方法在不同的分類問題上，有不同的表現。例如：在新聞文件的例子中，KNN 比 SVM 好，但在網頁描述的文件上則是 SVM 比 KNN 好（MicroF 雖相同，但 MacroF 方面 SVM 比 KNN 好）。
7. 相同的分類方法，在不同的分類問題上，其成效不盡相同。例如，網頁描述文件的分類數據顯示，SVM 的 MicroF 有 0.78 的成效，但在新聞文件中，只有 0.71 的成效。
8. 單獨一種分類方法，在單獨一種分類文件集上獲得的成效改進，難以保證其在另一種分類文件集上，也會有相同好的效果。

表五(a)：News 測試集的 MicroF 值

Sample	KNN	KNNs	KNNt	SVM	SVMs	SVMt
5%	0.47	**0.51**	**0.48**	0.40	**0.45**	**0.41**
10%	0.58	**0.64**	**0.60**	0.57	**0.60**	**0.59**
20%	0.70	0.67	0.70	0.63	0.62	**0.68**
40%	0.72	0.72	0.72	0.63	**0.65**	**0.71**
100%	0.79	0.78	0.77	0.71	**0.72**	**0.74**

表五(b)：News 測試集的 MacroF 值

Sample	KNN	KNNs	KNNt	SVM	SVMs	SVMt
5%	0.30	**0.35**	0.28	0.19	**0.29**	**0.27**
10%	0.32	**0.49**	**0.40**	0.31	**0.42**	**0.42**
20%	0.50	**0.54**	**0.52**	0.45	**0.49**	**0.54**
40%	0.62	0.61	**0.65**	0.49	**0.55**	**0.61**
100%	0.73	**0.76**	0.70	0.64	**0.66**	**0.69**

表六(a)：WebDes 測試集的 MicroF 值

Sample	KNN	KNNs	KNNt	SVM	SVMs	SVMt
5%	0.64	**0.69**	**0.67**	0.67	**0.68**	0.67
10%	0.69	**0.70**	**0.70**	0.71	0.70	**0.72**
20%	0.67	**0.73**	**0.74**	0.65	**0.73**	**0.75**
40%	0.75	0.75	**0.76**	0.78	0.77	**0.79**
100%	0.78	0.78	0.78	0.78	0.78	0.78

表六(b)：WebDes 測試集的 MacroF 值

Sample	KNN	KNNs	KNNt	SVM	SVMs	SVMt
5%	0.32	**0.43**	**0.39**	0.35	**0.43**	**0.37**
10%	0.38	**0.46**	**0.45**	0.42	**0.49**	**0.47**
20%	0.45	**0.49**	**0.51**	0.46	**0.52**	**0.54**
40%	0.55	**0.57**	**0.58**	0.61	**0.63**	0.61
100%	0.58	**0.63**	**0.61**	0.67	0.66	0.67

六、結論

文件分類在知識管理、資訊組織與檢索的服務上，是很重要的工作。由於需要大量而密集的知識加工，傳統上文件分類大都由人力進行。但其耗費的時間與成本相當可觀。自動分類系統近年來雖有研究，然而導入自動分類流程仍有相當的障礙，其中之一就是要準備足夠數量的訓練文件。

本文提出文件自我擴展法，在沒有利用任何額外資源的情況下，全自動的增加訓練文件，以期能提升分類的效果。在原始訓練文件數越少時，其改進的效果越明顯。而且此改進方法，乃策略層面上的技巧，與分類器無關，亦即任何一種分類器都可以運用上面的技巧來增強其分類效果。

雖然訓練文件數夠多時，此方法改進的成效不明顯，但這並非超乎預期。就好像利用壓縮法，將文件壓縮一遍，可以得到不錯的壓縮率，但再壓縮第二遍時，就得不到壓縮效果。如同上一節第 5 點所示，訓練文件夠多時，再怎們運用改進策略，其成效都有限。必須根本的改變分類（壓縮）方法，才可能大幅度地提升成效。

在少量訓練文件時就能提升成效是有價值的，這意味著我們能夠更快地提升自動分類系統的效益。如果將此方法與 EM 法結合，相當於我們在少量訓練文件時，即可獲得較佳的初始成效，這比起單獨運用 EM 法似乎效果更好。未來的研究將設計類似的實驗環境，以印證這個論點。

誌謝

本研究由國科會專題研究計畫補助，編號：NSC 92-2213-E-030-017-。

參考文獻

[1] Yiming Yang, "A Study on Thresholding Strategies for Text Categorization," Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 137-145.

[2] Fabrizio Sebastiani, “Machine Learning in Automated Text Categorization,” *ACM Computing Surveys*, 34(1):1-47, 2002.

[3] 曾元顯, "文件主題自動分類成效因素探討", 「中國圖書館學會會報」, 2002 年 6 月, 第 68 期, 頁 62-83.

[4] Fabrizio Sebastiani, Alessandro Sperduti and Nicola Valdambrini, "An Improved Boosting Algorithm and its Application to Text Categorization," Proceedings of the 9th International Conference on Information and Knowledge Management CIKM 2000, Pages 78 - 85.

[5] Yiming Yang and Xin Liu, "A Re-Examination of Text Categorization Methods," Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999, Pages 42 - 49.

[6] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," Machine Learning, 39(2/3):103-134, 2000.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, 39(1):1-38, 1977.

[8] Kamal Nigam and Rayid Ghani, "Analyzing the Effectiveness and Applicability of Co-training," Proceedings of the ninth international conference on information and knowledge management CIKM 2000, McLean, Virginia, United States, pp. 86 – 93.

[9] William B. Frakes and Ricardo Baeza-Yates, *Infomation Retrieval: Data Structure and Algorithms*, Prentice Hall, 1992.

[10] Amit Singhal and Fernando Pereira, “Document Expansion for Speech Retrieval,” Proceedings of the 22th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999, pp.34-41.

[11] 曾元顯, 第一章數位文件關鍵特徵之自動擷取, 數位文件之資訊擷取與檢索, 269 頁, 2000 年 9 月, ISBN 957-99750-3-2 , 全壘打文化事業有限公司出版.

[12] Yiming Yang and J. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the International Conference on Machine Learning (ICML'97), 1997, pp. 412-420.

[13] Hwee Tou Ng, Wei Boon Goh and Kok Leong Low, "Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization," Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1997, Pages 67 - 73.

[14] Thorsten Joachims, SVMlight: Support Vector Machine, version 5, http://svmlight.joachims.org/, 2002/03/07.

[15] Thorsten Joachims, "A Statistical Learning Model of Text Classification for Support Vector Machines," Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 128-136.

[16] Amit Singhal, Gerard Salton, and Chris Buckley, "Length Normalization in Degraded Text Collections," Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 149-162.

[17] Yuen-Hsien Tseng, "Automatic Catalogu ing and Searching for Retrospective Data by Use of OCR Text", Journal of American Society for Information Science and Technology, Vol. 52, No. 5, 2001, pp. 378-390.

[18] Yuen-Hsien Tseng and Douglas W. Oard, "Document Image Retrieval Techniques for Chinese" Proceedings of the Fourth Symposium on Document Image Understanding Technology, Columbia Maryland, April 23-25th, 2001, pp. 151-158.

[19] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, 2002, pp. 1130-1138.

Auto-Discovery of NVEF Word-Pairs in Chinese

Jia-Lin Tsai, Gladys Hsieh and Wen-Lian Hsu
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
{tsaijl,gladys,hsu}@iis.sinica.edu.tw

Abstract

A meaningful noun-verb word-pair in a sentence is called a noun-verb event-frame (NVFE). Previously, we have developed an NVEF word-pair identifier to demonstrate that NVEF knowledge can be used effectively to resolve the Chinese word-sense disambiguation (WSD) problem (with 93.7% accuracy) and the Chinese syllable-to-word (STW) conversion problem (with 99.66% accuracy) on the NVEF related portion.

In this paper, we propose a method for automatically acquiring a large scale NVEF knowledge without human intervention. The automatic discovery of NVEF knowledge includes four major processes: (1) segmentation check; (2) Initial Part-of-speech (POS) sequence generation; (3) NV knowledge generation and (4) automatic NVEF knowledge confirmation.

Our experimental results show that the precision of the automatically acquired NVEF knowledge reaches 98.52% for the test sentences. In fact, it has automatically discovered more than three hundred thousand NVEF word-pairs from the 2001 *United Daily News* (2001 *UDN*) corpus. The acquired NVEF knowledge covers 48% NV-sentences in *Academia Sinica Balanced Corpus* (***ASBC***), where an NV-sentence is one including at least a noun and a verb.

In the future, we will expand the size of NVEF knowledge to cover more than 75% of NV-sentences in ***ASBC***. We will also apply the acquired NVEF knowledge to support other NLP researches, in particular, shallow parsing, syllable/speech understanding and text indexing.

Keywords: noun-verb event frame (NVEF), machine learning, Hownet, WSD, STW

1. Introduction

The most challenging problem in NLP is to program computers to understand natural languages. For a human being, efficient syllable-to-word (STW) conversion and word sense disambiguation (WSD) arise naturally while a sentence is understood. Therefore, in designing a natural language understanding (NLD) system, two basic problems are to derive methods and knowledge for effectively performing the tasks of STW and WSD.

For most languages, a sentence is a grammatical organization of words expressing a complete thought [Chu 1982, Fromkin *et al.* 1998]. Since a word is usually encoded with ploy-senses, to understand language, efficient word sense disambiguation (WSD) becomes a critical problem for any NLD system. According to a study in cognitive science [Choueka *et al.* 1983], people often disambiguate word sense using only a few other words in a given context (frequently only one additional word). Thus, the relationships between one word and others can be effectively used to resolve ambiguity. Furthermore, from [Small *et al.* 1988, Krovetz *et al.* 1992, Resnik *et al.* 2000], most ambiguities occur with nouns and verbs, and the object-event (i.e. noun-verb) distinction is a major ontological division for humans [Carey 1992]. Tsai *et al.* (2002) have shown that the knowledge of noun-verb event frame (NVEF) sense/word-pairs can be used effectively to achieve a WSD accuracy of 93.7% for the NVEF related portion in Chinese, which supports the above claim of [Choueka *et al.* 1983].

The most common relationships between nouns and verbs are subject-predicate (SP) and verb-object (VO) [胡裕樹 *et al.* 1995, Fromkin *et al.* 1998]. In Chinese, such NV relationships could be found in various language units: compounds, phrases or sentences [Li *et al.* 1997]. As our observation, the major NV relationships in compounds/phrases are SP, VO, MH (modifier-head) and VC (verb-complement) constructions; the major NV relationships in sentences are SP and VO constructions. Consider the Chinese sentence: 這輛車行駛順暢(This car moves well). There are two possible NV word-pairs, "車-行駛(car, move)" and "車行-駛(auto shop, move)." It is clear that the permissible (or meaningful) NV word-pair is "車-行駛(car, move)" and it is a SP construction. We call such a permissible NV word-pair a noun-verb *event frame* (*NVEF*) word-pair. And, the collection of the NV word-pair 車-行駛 and its sense-pair **Land-Vehicle|車-VehicleGo|駛** is called a permissible NVEF knowledge.

The most popular input method for Chinese is syllable-based. Since the average number of characters sharing the same syllable is 17, efficient STW conversion becomes an indispensable tool. Tsai *et al.* (2002) have shown that the NVEF knowledge can be used to achieve a STW accuracy rate of 99.66% for converting NVEF related words. Since the creation of NVEF knowledge bears no particular application in mind, and still it can be used to effectively resolve the WSD and STW problems, the NVEF knowledge is potentially application independent for NLP. We shall further investigate the effectiveness of NVEF knowledge in other NLP applications, such as syllable/speech understanding and full/shallow parsing.

We have reported a semi-automatic generation of NVEF knowledge in [Tsai *et al*. 2002]. This method uses the N-V frequencies in sentences groups to generate NVEF candidates to be filtered by human editors. However, it is quite laborious to create a large scale NVEF knowledge. In this paper, we propose a new method to discover NVEF knowledge automatically from running texts, and construct a large scale NVEF knowledge efficiently.

This paper is arranged as follows. In Section 2, we present the details of auto-discovery of NVEF knowledge. Experimental results and analyses are described in Section 3. Conclusion and directions for future researches will be discussed in Section 4.

2. Development of Auto-Discovery of NVEF Knowledge

To develop an auto-discovery system for NVEF knowledge (AUTO-NVEF), we use Hownet 1.0 [Dong] as a system dictionary. This system dictionary provides knowledge of the Chinese word (58,541 words), parts-of-speech (POS) and word senses, in which there are 33,264 nouns, 16,723 verbs and 16,469 senses (including 10,011 noun-senses and 4,462 verb-senses).

2.1 Definition of the NVEF Knowledge

The sense of a word is defined as its DEF (concept definition) in Hownet. Table 1 lists three different senses of the Chinese word "車(Che/car/turn)." In Hownet, the DEF of a word consists of its main feature and secondary features. For example, in the DEF "character|文字,surname|姓,human|人,ProperName|專" of the word "車(Che)," the first item "character|文字" is the main feature, and the remaining three items, "surname|姓," "human|人," and "ProperName|專," are its secondary features. The main feature in Hownet can inherit features in the hypernym-hyponym hierarchy. There are approximately 1,500 features in Hownet. Each of these features is called a *sememe*, which refers to the smallest semantic unit that cannot be further reduced.

Table 1. Three different senses of the Chinese word "車(Che/car/turn)"

C.Word [a]	E.Word [a]	Part-of-speech	Sense (i.e. DEF in Hownet)
車	Che	Noun	character|文字,surname|姓,human|人,ProperName|專
車	car	Noun	LandVehicle|車
車	turn	Verb	cut|切削

[a] C.Word refers to a Chinese word; E.Word refers to an English word

As we mentioned, a permissible (or meaningful) NV word-pair is a noun-verb event-frame word-pair (*NVEF word-pair*), such as 車-行駛(Che/car/turn, move). From Table 2, the only permissible NVEF sense-pair for 車-行駛(car, move) is **LandVehicle|車-VehicleGo|駛**. Such an NVEF sense-pair and its corresponding NVEF word-pairs is called NVEF knowledge. Here, the combination of the NVEF sense-pair **LandVehicle|車-VehicleGo|駛** and the NVEF word-pair **車-行駛** constructs a collection of NVEF knowledge.

To effectively represent the NVEF knowledge, we have proposed an NVEF knowledge representation tree (NVEF KR-tree) to store and display the collected NVEF knowledge. The details of the NVEF KR-tree are described below [Tsai *et al*. 2002].

2.2 Knowledge Representation Tree of NVEF Sense-Pairs and Word-Pairs

A knowledge representation tree (KR-tree) of NVEF sense-pairs is shown in Fig.1.

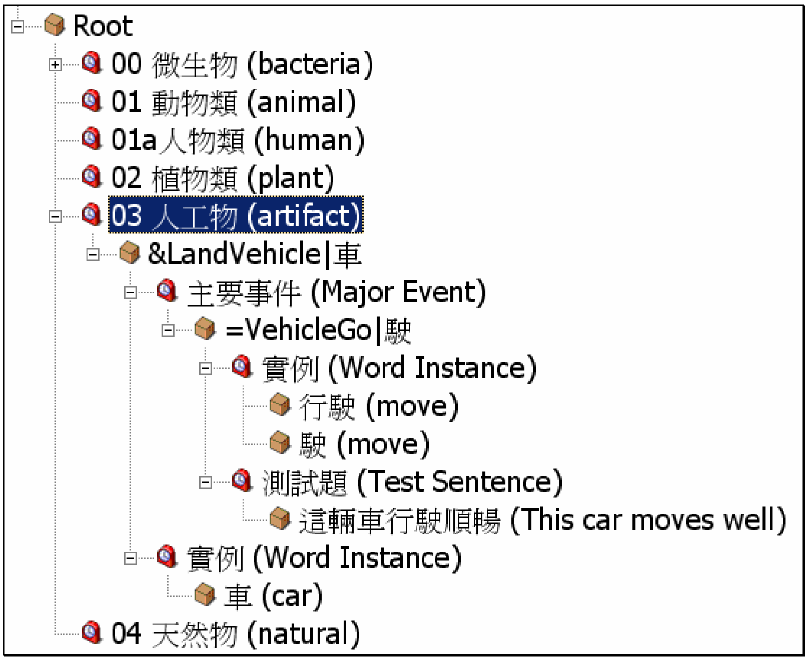


Figure 1. An illustration of the KR-tree using "人工物(artifact)" as an example noun-sense subclass. (The English words in parentheses are provided for explanatory purposes only.)

There are two types of nodes in the KR-tree, namely, ***function nodes*** and ***concept nodes***. Concept nodes refer to words and features in Hownet. Function nodes are used to define the relationships between the parent and children concept nodes. We omit the function node "subclass" so that if a concept node B is the child of another concept node A, then B is a subclass of A. We can classify the noun-sense class (名詞詞義分類) into 15 subclasses according to their main features. These are "微生物(bacteria)," "動物類(animal)," "人物類(human)," "植物類

(plant),” “人工物(artifact),” “天然物(natural),” “事件類(event),” “精神類(mental),” “現象類(phenomena),” “物形類(shape),” “地點類(place),” “位置類(location),” “時間類(time),” “抽象類(abstract)” and “數量類(quantity).” Appendix A provides a sample table of the 15 main features of nouns in each noun-sense subclass.

The three function nodes used in the KR-tree are shown in Figure 1:

(1) **Major-Event** (主要事件): The content of its parent node represents a noun-sense subclass, and the content of its child node represents a verb-sense subclass. A noun-sense subclass and a verb-sense subclass linked by a Major-Event function node is an NVEF subclass sense-pair, such as “&LandVehicle|車” and “=VehcileGo|駛” in Figure 1. To describe various relationships between noun-sense and verb-sense subclasses, we design three subclass sense-symbols, in which “=” means “*exact*,” “&” means “*like*,” and “%” means “*inclusive*.” An example using these symbols is provided below.

Provided that there are three senses S_1, S_2, and S_3 as well as their corresponding words W_1, W_2, and W_3. Let

S_1 = LandVehicle|車,*transport|運送,#human|人,#die|死	W_1=“靈車(hearse)”
S_2 = LandVehicle|車,*transport|運送,#human|人	W_2=“客車(bus)”
S_3 = LandVehicle|車,police|警	W_3=“警車(police car)”

Then, we have that sense/word S_3/W_3 is in the “=LandVehicle|車,police|警” *exact*-subclass; senses/words S_1/W_1 and S_2/W_2 are in the “&LandVehicle|車,*transport|運送” *like*-subclass; and senses/words S_1/W_1, S_2/W_2, and S_3/W_3 are in the “%LandVehicle|車” *inclusive*-subclass.

(2) **Word-Instance** (實例): The content of its children are the words belonging to the sense subclass of its parent node. These words are self-learned by the NVEF sense-pair identifier according to the sentences under the Test-Sentence nodes.

(3) **Test-Sentence** (測試題): The content of its children is several selected test sentences in support of its corresponding NVEF subclass sense-pair.

2.3 Auto-Discovery of NVEF Knowledge

The task of AUTO-NVEF is to automatically find out meaningful NVEF sense/word-pairs (NVEF knowledge) from Chinese sentences. Figure 1 is the flow chart of AUTO-NVEF. There are four major processes in AUTO-NVEF. The details of these major processes are described as follows (see Figure 2 and Table 2).

Process 1. Segmentation check: In this stage, the Chinese sentence will be segmented by two strategies: *right-to-left longest word first* (RL-LWF), and *left-to-right longest word first* (LR-LWF). If both RL-LWF and LR-LWF segmentations are equal (in short form, RL-LWF=LR-LWF) and the word number of the segmentation is greater than one, this segmentation result will be sent to ***process 2***; otherwise, a *NULL* segmentation will be sent. Table 3 is a comparison of word-segmentation accuracies for RL-LWF, LR-LWF and RL-LWF=LR-LWF strategies with CKIP lexicon [CKIP 1995]. The word-segmentation accuracy is the ratio of fully correct segmented sentences to all sentences of *Academia Sinica Balancing Corpus* (***ASBC***) [CKIP 1995]. A fully correct segmented sentence means the segmented result exactly matches its corresponding segmentation ***ASBC***. Table 3 shows that the technique of RL-LWF=LR-LWF achieves the best word-segmentation accuracy.

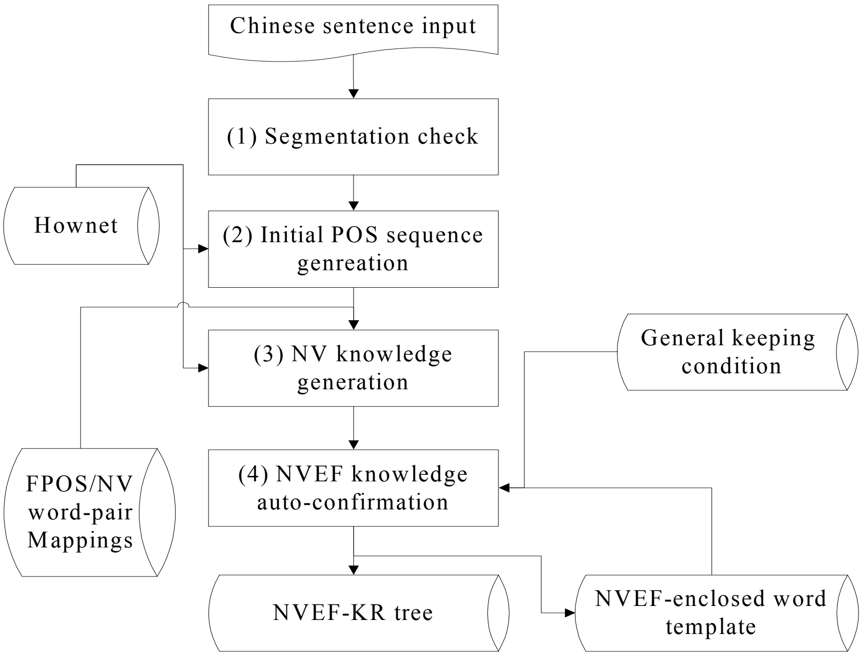

Chinese sentence input
(1) Segmentation check
Hownet
(2) Initial POS sequence genreation
(3) NV knowledge generation
General keeping condition
FPOS/NV word-pair Mappings
(4) NVEF knowledge auto-confirmation
NVEF-KR tree
NVEF-enclosed word template

Figure 2. The flow chart of AUTO-NVEF

Table 2. An illustration of AUTO-NVEF for the Chinese sentence "音樂會現場湧入許多觀眾 (There are many audiences entering the locale of concert)." (The English words in parentheses are included for explanatory purpose only.)

Process	Output
(1)	音樂會(concert)/現場(locale)/湧入(enter)/許多(many)/觀眾(audience)
(2)	$N_1N_2V_3ADJ_4N_5$, where N_1 =[音樂會]; N_2 =[現場]; V_3=[湧入]; ADJ_4=[許多]; N_5=[觀眾]

(3)	NV_1 = "現場/place|地方,#fact|事情/N"
	- "湧入(yong3 ru4)/GoInto|進入/V"
	NV_2 = "觀眾/human|人,*look|看,#entertainment|藝,#sport|體育,*recreation|娛樂/N"
	- "湧入(yong3 ru4)/GoInto|進入/V"

(4)	NV_1 is NVEF knowledge by keeping-condition; learned NVEF template is [音樂會 NV 許多] NV_2 is NVEF knowledge by keeping-condition; learned NVEF template is [現場 V 許多 N]

Table 3. A comparison of word-segmentation accuracies for RL-LWF, LR-LWF and RL-LWF = LR-LWF strategies (the test sentences are ***ASBC*** and the dictionary is CKIP lexicon)

	RL-LWF	LR-LWF	RL-LWF = LF-LWF
Accuracy	82.5%	81.7%	86.86%
Recall	100%	100%	89.33%

Process 2. Initial POS sequence generation: If the output of ***process 1*** is not a *NULL* segmentation, this process will be triggered. This stage is comprised of the following steps.

1) For the segmentation result $w_1/w_2/…/w_{n-1}/w_n$ from ***process 1***, our algorithm compute the POS of w_i, where i = 2 to n, as follows. It first computes the following two sets: a) the *following POS/frequency set* of w_{i-1} by ***ASBC*** tagging corpus and b) the *Hownet POS set* of w_i. Then, it computes the POS intersection of the two sets. Finally, it selects the POS with the largest frequency in the POS intersection to be the POS of w_i. If there are more than one POS with the largest frequency, the POS of w_i will be set to *NULL* POS.
2) Similarly, the POS of w_1 will be determined by the POS with the largest frequency in the POS intersection of the *preceding POS/frequency set* of w_2 and the *Hownet POS set* of w_1.
3) By combining the determined POSs of w_i, where i =1 to n, the ***initial POS sequence*** (***IPOS***) will be generated. Take the Chinese segmentation 生/了 as an example. The following POS/frequency set of the Chinese word 生(bear) is {N/103, PREP/42, STRU/36, V/35, ADV/16, CONJ/10, ECHO/9, ADJ/1}. The Hownet POS set of the Chinese word 了 is {V, STRU}. According to these sets, we have POS intersection {STRU/36, V/35}. Since the POS with the largest frequency in this intersection is **STRU**, the POS of 了 will be set to **STRU**. Similarly, according to the intersection {V/16124, N/1321, ADJ/4} of the preceding POS/frequency set {V/16124, N/1321, PREP/1232, ECHO/121, ADV/58, STRU/26, CONJ/4, ADJ/4} of 了 and the Hownet POS set {V, N, ADJ} of 生, the POS of 生 will be set to **V**. Table 4 is a mapping list of CKIP POS tag and Hownet POS tag.

Table 4. A mapping list of CKIP POS tag and Hownet POS tag

	Noun	Verb	Adjective	Adverb	Preposition	Conjunction	Expletive	Structural Particle
CKIP	N	V	A	D	P	C	T	De
Hownet	N	V	ADJ	ADV	PP	CONJ	ECHO	STRU

Process 3. NV knowledge generation: If the output of ***process 2*** does not include any *NULL* POS, this process will be triggered. The steps of this process are given as follows.

1) Compute the ***final POS sequence*** (***FPOS***). For the portion of contiguous noun sequence (such as N_1N_2) of the ***IPOS***, the last noun (such as N_2) will be kept and the other nouns will

be dropped from the ***IPOS***. This is because the last noun of a contiguous noun sequence (such as 航空/公司) in Chinese is usually the head of such a sequence. This step translates an ***IPOS*** into a ***FPOS***. Take the Chinese sentence 音樂會(N_1)現場(N_2)湧入(V_3)許多(ADJ_4)觀眾(N_5) as an example. Its ***IPOS*** ($N_1N_2V_3ADJ_4N_5$) will be translated into ***FPOS*** ($N_1V_2ADJ_3N_4$).

2) According to the ***FPOS***, the NV word-pairs will be generated. In this case, since the auto-generated NV word-pairs for the ***FPOS*** $N_1V_2ADJ_3N_4$ are N_1V_2 and N_4V_2, the NV word-pairs **現場(N)湧入(V)** and **湧入(V)觀眾(N)** will be generated. Appendix. B lists three sample mappings of the ***FPOSs*** and their corresponding NV word-pairs. In this study, we create about one hundred mappings of ***FPOSs*** and their corresponding NV word-pairs.

3) According to Hownet, it computes all NV sense-pairs for the generated NV word-pairs. For the above case, we have two collections of NV knowledge (see Table 2):

NV_1 = "現場(locale)/place|地方,#fact|事情/N" – "湧入(enter)/GoInto|進入/V", and

NV_2 = "觀眾(audience)/human|人,*look|看,#entertainment|藝,#sport|體育,*recreation|娛樂/N" – "湧入(enter)/GoInto|進入/V".

Process 4. NVEF knowledge auto-confirmation: In this stage, it automatically confirms whether the generated NV knowledge is NVEF knowledge. The two auto-confirmation procedures are given as follows.

(a) **General keeping (GK) condition check**: Each GK condition is constructed by a noun-sense class defined in [Tsai *et al*. 2002] (see Appendix A) and a verb main DEF in Hownet 1.0 [Dong]. For example, the pair of noun-sense class "人物類(human)" and verb main DEF "GoInto|進入" is a GK condition. In [Tsai *et al*. 2002], we created 5,680 GK conditions from the manually confirmed NVEF knowledge. If the noun-sense class and the verb main DEF of the generated NV knowledge fits one of GK conditions, it will be automatically confirmed as a collection of NVEF knowledge and sent to NVEF KR-tree. Appendix. C gives ten GK conditions used in this study.

(b) **NVEF enclosed-word template (NVEF-EW template) check**: If the generated NV knowledge cannot be auto-confirmed as NVEF knowledge in procedure (a), this procedure will be triggered. A NVEF-EW template is composed of all left words and right words of a NVEF word-pair in a Chinese sentence. For example, the NVEF-EW template of the NVEF word-pair "汽車-行駛(car, move)" in the Chinese sentence 這(this)/汽車(car)/似乎(seem)/行駛(move)/順暢(well) is ***這 N 似乎 V 順暢***. In this study, all the NVEF-EW templates are generated from the following resources: i) the collection of manually confirmed NVEF knowledge in [Tsai *et al*. 2002], ii) the automatically confirmed NVEF knowledge and iii) the NVEF-EW templates provided by human editor. In this procedure, if the NVEF-EW template of the generated NV word-pair for the Chinese sentence input matches one of the NVEF-EW templates, it will be automatically confirmed as a collection of NVEF knowledge.

3. Experiments

To evaluate the performance of the proposed auto-discovery of NVEF knowledge, we define the NVEF accuracy and NVEF-identified sentence coverage by Equations (1) and (2):

NVEF accuracy =
# of permissible NVEF knowledge / # of total generated NVEF knowledge. (1)

NVEF-identified sentence coverage =
# of NVEF-identified sentences / # of total NV sentences. (2)

In Equation (1), a permissible NVEF knowledge means the generated NVEF knowledge is manually confirmed as a collection of NVEF knowledge. In Equation (2), if the Chinese sentence contains greater or equal to one NVEF word-pair on our NVEF KR-tree by the NVEF word-pair identifier [Tsai *et al*. 2002], this sentence is called an **NVEF-identified sentence**. If the Chinese sentence contains at least one noun and verb, this sentence is called an **NV sentence**. As our computation, there are about 75% of Chinese sentences in Sinica corpus are NV sentences.

Chinese sentence	高度壓力使有些人[食量]<減少> (High pressure makes some people that their [eating capacity] <decreased>.)		
名詞詞義 (Noun sense)	attribute|屬性,ability|能力,&eat|吃	動詞詞義 (Verb sense)	subtract|削減
名詞 (Noun)	食量 (eating capacity)	動詞 (Verb)	減少 (decrease)

Figure 3. The confirmation UI of NVEF knowledge taking the generated NVEF knowledge for the Chinese sentence 高度壓力使有些人食量減少 (High pressure makes some people that their eating-capacity decreased as an example. (The English words in parentheses, symbols [] used to mark a noun and <> used to mark a verb are there for explanatory purposes only)

3.1 User Interface (UI) for Manually Confirming NVEF Knowledge

An evaluation UI for the generated NVEF knowledge is developed as shown in Figure 3. By this UI, evaluators (native Chinese speakers) can review the generated NVEF knowledge and determine whether it is a permissible NVEF knowledge. Take the Chinese sentence 高度壓力使有些人食量減少(High pressure makes some people that their eating capacity decreased) as an example. For this case, AUTO-NVEF will generate a collection of NVEF knowledge including the NVEF sense-pair [attribute|屬性,ability|能力,&eat|吃]-[subtract|削減] and the NVEF word-pair [食量(eating capacity)]-[減少(decrease)]. According to the confirmation principles of permissible NVEF knowledge, evaluators will confirm this generated NVEF knowledge as a permissible NVEF knowledge. The confirmation principles of permissible NVEF knowledge are given as follows.

3.2 Confirmation Principles of permissible NVEF Knowledge

An auto-generated NVEF knowledge should be confirmed as a collection of permissible NVEF knowledge if it fits all three principles below.

Principle 1. Do the NV word-pair make correct POS tags for the given Chinese sentence?
Principle 2. Do the NV sense-pair and the NV word-pair make sense?
Principle 3. Do most NV word-pair instances for the NV sense-pair satisfy Principles 1 and 2?

3.3 Experimental Results

To evaluate the acquired NVEF knowledge, we divide the 2001 *United Daily News* (2001 UDN) corpus into two distinct sub-corpora. (The UDN 2001 corpus contains 4,539,624 Chinese sentences that were extracted from the *United Daily News* Web site [On-Line United Daily News] from January 17, 2001 to December 30, 2001.)

(1) **Training corpus.** This is the collection of Chinese sentences extracted from the 2001 *UDN* corpus from January 17, 2001 to September 30, 2001. According to the training corpus, we create thirty thousand manually confirmed NVEF word-pairs, which are used to derive the 5,680 general keeping conditions.

(2) **Testing corpus.** This is the collection of Chinese sentences extracted from the 2001 *UDN* corpus from October 1, 2001 to December 31, 2001.

(3) **Test sentences set**. From the testing corpus, we randomly select three days' sentences (October 27, 2001, November 23, 2001 and December 17, 2001) to be our test sentences set.

All of the acquired NVEF knowledge by AUTO-NVEF on the test sentences are manually confirmed by three evaluators. Table 5 is the experimental results of AUTO-NVEF. From Table 5, it shows that AUTO-NVEF can achieve a NVEF accuracy of 98.52%.

Table 5. Experimental results of AUTO-NVEF

Date of test news	NVEF accuracy	Evaluator
October 27, 2001	99.10% (1,095/1,105)	A
November 23, 2001	97.76% (1,090/1,115)	B
December 17, 2001	98.63% (2,156/2,186)	C
Total Average	98.52% (4,341/4,406)	

When we apply AUTO-NVEF to the entire 2001 UDN corpus, it auto-generates 167,203 NVEF sense-pairs (8.6M) and 317,820 NVEF word-pairs (10.1M) on the NVEF KR-tree. Within this data, 47% is generated through the general keeping conditions check and the other 53% is generated by the NVEF-enclosed word templates check.

Table 6. An illustration of four types of NVEF knowledge and their coverage (The English words in parentheses, symbols [] and <> are there for explanatory purposes only)

NV pair Type	**Sentence**	**Noun / DEF**	**Verb / DEF**	**Coverage**
N:V	[工程]<完成> (The construction is now completed)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	24.15%
N-V	全部[工程]預定年底<完成> (All of constructions will be completed by the end of year)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	43.83%
V:N	<完成>[工程] (to complete a construction)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	19.61%
V-N	建商承諾在年底前<完成>鐵路[工程] (The building contractor promise to complete railway construction before the end of this year)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	12.41%

3.3.1 Coverage for the Four Types of NVEF Knowledge

According to the noun and verb positions of NVEF word-pairs in Chinese sentences, the NVEF knowledge can be classified into four types: **N:V**, **N-V**, **V:N**, and **V-N**, where the symbols ":" stands for "next to" and "-" stands for "near by." Table 6 shows examples and the coverage of the four types of NVEF knowledge, in which the ratios (coverage) of the collections of **N:V**, **N-V**, **V:N** and **V-N** are 12.41%, 43.83%, 19.61% and 24.15%, respectively, by applying AUTO-NVEF to 2001 *UDN* corpus. It seems that the percentage of **SP** construction is a little more than that of **VO** construction in the training corpus.

3.3.2 Error Analysis - The Non-Permissible NVEF Knowledge Generated by AUTO-NVEF

One hundred collections of the generated non-permissible NVEF (NP-NVEF) knowledge are analyzed. We classify these into eleven error types as shown in Table 7, which lists the NP-NVEF confirmation principles and the ratios for the eleven error types. The first three types consist of 52% of the cases that do not satisfy the NVEF confirmation principles 1, 2 and 3 in Section 3.2. The fourth type is rare with 1% of the cases. Types 5 to 7 consists of 11% of the cases and are caused from incorrect Hownet lexicon, such as the incorrect word-sense ***exist|存在*** for the Chinese word 盈盈 (an adjective, normally used to describe a beauty's smile). Types 8 to 11 are referred to as the *four NLP errors* (36% of NP-NVEF cases): Type 8 is the problem of different word-senses used in Ancient and Modern Chinese; type 9 is caused by errors in WSD; type 10 is caused by the unknown word problem; and type 11 is caused by incorrect word segmentation.

Table 7. The eleven error types and their confirming principles of non-permissible NVEF knowledge generated by AUTO-NVEF

Type	**Confirming principle of Non-Permissible NVEF Knowledge**	Percentage
1[*]	NV Word-pair cannot make a reasonable and legitimate POS tagging for the Chinese sentence.	33% (33/100)
2[*]	NV sense-par (DEF) and the NV word-pair cannot make sense for each other	17% (17/100)
3[*]	In this NV pair, one of word sense cannot inherit its parent category.	2% (2/100)
4[**]	The NV pair cannot be the proper combination in the sentence although this pair fits principles (a), (b), and (c).	1% (1/100)
5	Incorrect word POS in Hownet	1% (1/100)
6	Incorrect word sense in Hownet	3% (3/100)
7	No proper definition in Hownet Ex:暫居(temporary residence)，it has two meanings, one is <reside|住下>（緊急暫居服務(Emergent temporary residence service)）and another one is <situated|處,Timeshort|暫>（SARS 帶來暫時性的經濟震盪(SARS will produce only a temporary economic shock))·	7% (7/100)
8	Lack of different meaning usage for Old Chinese and Modern Chinese	3% (3/100)
9	Failure of word sense disambiguation (1) General sense Polysemous word (2) Domain sense Person name, Appellation, Organization named as common word Ex: 公牛隊(**Chicago Bulls)** ⇨公牛(**bull**) <livestock|牲畜>；太陽隊 (**Phoenix Suns)** ⇨太陽(**Sun**) <celestial|天體>；花木蘭(**Mulan**)⇨木蘭(**magnolia**)< FlowerGrass|花草>	27% (27/100)
10	Unknown word problem	4% (4/100)
11	Error of word segmentation	2% (2/100)

[*] Types 1 to 3 are contrast to the confirming principles of permissible NVEF knowledge mentioned in section 3.2, respectively.

[**] Type 4 contents principles (a), (b), and (c) in section 3.2 but there is no proper combination in that sentence.

Table 8. Examples of the eleven types of non-permissible NVEF knowledge. (The English words in parentheses, symbols [] and <> are there for explanatory purposes only.)

NP type	**Sentence** (English explanation)	**Noun** (English explanation) **DEF**	**Verb** (English explanation) **DEF**
1	警方維護地方[治安]<辛勞> (Police work hard to safeguard the locality security.)	治安 (public security) attributel屬性,circumstancesl境況,safel安,politicsl政,&organizationl組織	辛勞 (work hard) endeavourl賣力
2	<模糊>的[白宮]景象 (White House looked vague in the heavy fog.)	白宮 (White House) housel房屋,institutionl機構,#politicsl政,(USl美國)	模糊 (vague) PolysemousWordl多義詞,CauseToDol使動,mixl混合
3	<生活>條件[不足] (Lack of living condtions)	不足 (lackness) attributel屬性,fullnessl空滿,incompletel缺,&entityl實體	生活 (life) alivel活著
4	網路帶給[企業]許多<便利> (Internet brings numerous benefits to industries.)	企業 (Industry) InstitutePlacel場所,*producel製造,*selll賣,industriall工,commerciall商	便利 (benefit) benefitl便利
5	<盈盈>[笑靨] (smile radiantly)	笑靨 (a smiling face) part|部件,%human|人,skin|皮	盈盈 (an adjective, normally to describe a beauty's smile) exist|存在
6	保費較貴的<壽險>[保單] (higher fare life insurance policy)	保單 (insurance policy) bill|票據,*guarantee|保證	壽險 (life insurance) guarantee|保證,scope=die|死,commercial|商
7	債券型基金吸金[存款]<失血> Bond foundation makes profit but savings is loss	存款 (bank savings) moneyl貨幣,$SetAsidel留存	失血 (bleed or loss(only use in finance diction)) bleedl出血
8	華南[銀行] 中山<分行> (Hwa-Nan Bank Jung-San Branch)	銀行 (bank) InstitutePlace|場所,@Set Aside|留存,@TakeBack|取回,@lend|借出,#wealth|錢財,commercial|商	分行 (branch) separate|分離
9	[根據]<調查> (according to the investigation)	根據 (evidence) information|信息	調查 (investigate) investigate|調查
10	<零售>[通路] (retail sell routes)	通路 (route) facilities|設施,route|路	零售 (retail sell) sell|賣
11	從今日<起到> 5[月底] (from today to the end of May)	月底 (the end of month) time|時間,ending|末,month|月	起到 (to elaborate) do|做

Table 8 gives the examples for the eleven types of NP-NVEF knowledge. From Tables 8 and 9, 11% of NP-NVEF cases can be resolved by correcting the error lexicon in original Hownet. For the four NLP errors, these cases could be improved with the support of other techniques such as WSD ([Resnik *et al.* 2000, Yang *et al.* 2002]), unknown word identification ([Chang *et al.* 1997, Lai *et al.* 2000, Chen *et al.* 2002, Sun *et al.* 2002 and Tsai *et al.* 2003]) and word segmentation ([Sproat *et al.* 1996, Teahan *et al.* 2000]).

4. Conclusion and Directions for Future Research

In this paper, we present an auto-discovery system of NVEF knowledge that can be used to automatically generate a large scale NVEF knowledge for NLP. The experimental results shows that AUTO-NVEF achieves a NVEF accuracy of 98.52%. By applying AUTO-NVEF to the 2001 *UDN* corpus, we create 167,203 NVEF sense-pairs (8.6M) and 317,820 NVEF word-pairs (10.1M) on the NVEF-KR tree. Using this collection of NVEF knowledge, we have designed an NVEF word-pair identifier [Tsai *et al*. 2002] to achieve a WSD accuracy of 93.7% and a STW accuracy of 99.66% for the NVEF related portion in Chinese sentences. The acquired NVEF knowledge can cover 48% and 50% of NV-sentences in ***ASBC*** and in 2001 ***UDN*** corpus, respectively.

Our database for the NVEF knowledge has not been completed. Currently, there are 66.34% (=6,641/10,011) of the noun-senses in Hownet have been considered in the NVEF knowledge construction. The remaining 33.66% of the noun-senses in Hownet not dealt with yet are caused by two problems: (1) those words with ploy-noun-senses or poly-verb-senses, which are difficult to be resolved by WSD, especially those single-character words; and (2) corpus sparseness. We will continue expanding our NVEF knowledge through other corpora. The mechanism of AUTO-NVEF will be extended to auto-generate other meaningful co-occurrence semantic restrictions, in particular, noun-noun association frame (NNAF) pairs, noun-adjective grammar frame (NAGF) pairs and verb-adverb grammar frame (VDGF) pairs. As of our knowledge, the NVEF/NNAF/NAGF/VDGF pairs are the four most important co-occurrence semantic restrictions for language understanding.

Since the creation of NVEF knowledge bears no particular application in mind, and still it can be used to effectively resolve the WSD and STW problems, the NVEF knowledge is potentially application independent for NLP. We shall further investigate the effectiveness of NVEF knowledge in other NLP applications, such as syllable/speech understanding and full/shallow parsing.

5. Acknowledgements

We are grateful to our colleagues in the Intelligent Agent Systems Lab. (IASL), Li-Yeng Chiu, Mark Shia, Gladys Hsieh, Masia Yu, Yi-Fan Chang, Jeng-Woei Su and Win-wei Mai, who helped us create and verify all the necessary NVEF knowledge and tools for this study. We would also like to thank Prof. Zhen-Dong Dong for providing us with the Hownet dictionary.

Reference

Carey, S., “The origin and evolution of everyday concepts (In R. N. Giere, ed.),” *Cognitive Models of Science*, Minneapolis: University of Minnesota Press, 1992.

Chang, J. S. and K. Y. Su, “An Unsupervised Iterative Method for Chinese New Lexicon Extraction,” *International Journal of Computational Linguistics & Chinese language Processing*, 1997Choueka, Y. and S. Lusignan, "A Connectionist Scheme for Modeling Word Sense Disambiguation," *Cognition and Brain Theory*, 6 (1) 1983, pp.89-120.

Chen, K. J. and W. Y. Ma, "Unknown Word Extraction for Chinese Documents," *Proceedings of 19th COLING 2002*, Taipei, pp.169-175

Chu, S. C. R., *Chinese Grammar and English Grammar: a Comparative Study*, The Commerical Press, Ltd. The Republic of China, 1982

CKIP. *Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica, 1995. http://godel.iis.sinica.edu.tw/CKIP/r_content.html

Dong, Z. and Q. Dong, *Hownet*, http://www.keenage.com/

Fromkin, V. and R. Rodman, *An Introduction to Language*, Sixth Edition, Holt, Rinehart and Winston, 1998

Krovetz, R. and W. B. Croft, "Lexical Ambiguity and Information Retrieval," *ACM Transactions on Information Systems*, 10 (2), 1992, pp.115-141.

Lai, Y. S. and Wu, C. H., "Unknown Word and Phrase Extraction Using a Phrase-Like-Unit-based Likelihood Ratio," *International Journal of Computer Processing Oriental Language*, 13(1), pp.83-95

Li, N. C. and S. A. Thompson, Mandarin Chinese: a Functional Reference Grammar, The Crane Publishing Co., Ltd. Taipei, Taiwan, 1997

On-Line United Daily News, http://udnnews.com/NEWS/

Resnik, P. and D. Yarowsky, "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation," *Natural Language Engineering*, 5 (3), 2000, pp.113-133.

Small, S., and G. Cottrell, and M. E. Tannenhaus, *Lexical Ambiguity Resolution*, Morgan Kaufmann, Palo Alto, Calif., 1988.

Sun, J., J. Gao, L. Zhang, M. Zhou and C. Huang, "Chinese Named Entity Identification Using Class-based Language Model," *Proceedings of 19th COLING 2002*, Taipei, pp.967-973

Sproat, R. and C. Shih, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22(3), 1996, pp.377-404

Teahan, W.J., Wen, Y., McNab, R.J., Witten, I.H., "A compression-based algorithm for chinese word segmentation," *Computational Linguistics*, 26, 2000, pp.375-393

Tsai, J. L, W. L. Hsu and J. W. Su, "Word sense disambiguation and sense-based NV event-frame identifier," *Computational Linguistics and Chinese Language Processing,* Vol. 7, No. 1, February 2002, pp.29-46

Tsai, J. L, W. L. Hsu, "Applying NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem," *Proceedings of 19th COLING 2002*, Taipei, pp.1016-1022

Tsai, J. L, C. L. Sung and W. L. Hsu, "Chinese Word Auto-Confirming Agent," *Proceeding of ROCLING XV*, 2003

Yang, X. and Li T., “A study of Semantic Disambiguation Based on HowNet,” *Computational Linguistics and Chinese Language Processing,* Vol. 7, No. 1, February 2002, pp.47-78

陳克健，洪偉美，“中文裏「動—名」述賓結構與「動—名」偏正結構的分析，” *Communication of COLIPS*, 6(2), 1996, pp.73-79

胡裕樹，范曉，*動詞研究*，河南大學出版社，1995

Appendix A. A Sample Table of the Main Features of Nouns and their corresponding Noun-Sense Classes

An example Main Feature	Noun-sense Class
bacteria|微生物	微生物
AnimalHuman|動物	動物類
human|人	人物類
plant|植物	植物類
artifact|人工物	人工物
natural|天然物	天然物
fact|事情	事件類
mental|精神	精神類
phenomena|現象	現象類
shape|物形	物形類
InstitutePlace|場所	地點類
location|位置	位置類
attribute|屬性	抽象類
quantity|數量	數量類

Appendix B. Example Mappings of FPOS and NV Word-Pairs

FPOS	NV word-pairs	Example, [] stands for noun and <> stands for verb
$N_1V_2ADJ_3N_4$	N_1V_2 & N_4V_2	[學生]<購買>許多[筆記本]
N_1V_2	N_1V_2	[雜草]<枯萎>
N_1 ADJ_2 ADV_3V_4	N_1V_4	[意願]遲未<回升>

Appendix C. Ten Examples of General-Keeping (GK) Conditions

Noun-sense class	Verb DEF	Example, [] stands for noun and <> stands for verb
微生物(bacteria)	own|有	已經使[細菌]<具有>高度抗藥性
位置類(location)	arrive|到達	若正好<蒞臨>[西班牙]
植物類(plant)	decline|衰敗	田中[雜草]<枯萎>
人工物(artifact)	buy|買	民眾不需要急著<購買>[米酒]
天然物(natural)	LeaveFor|前往	立刻驅船<前往>蘭嶼[海域]試竿
事件類(event)	alter|改變	批評這會<扭曲>[貿易]
精神類(mental)	BecomeMore|增多	民間投資[意願]遲未<回升>
現象類(phenomena)	announce|發表	做任何<公開>[承諾]
物形類(Shape)	be|是,all|全	由於從腰部以下<都是>合身[線條]
地點類(place)	from|相距	<距離>[小學]七百公尺

Reliable and Cost-Effective PoS-Tagging

Yu-Fang Tsai **Keh-Jiann Chen**
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan 115
eddie,kchen@iis.sinica.edu.tw

Abstract

In order to achieve fast and high quality Part-of-speech (PoS) tagging, algorithms should be high accuracy and require less manually proofreading. To evaluate a tagging system, we proposed a new criterion of reliability, which is a kind of cost-effective criterion, instead of the conventional criterion of accuracy. The most cost-effective tagging algorithm is judged according to amount of manual editing and achieved final accuracy. The reliability of a tagging algorithm is defined to be the estimated best accuracy of the tagging under a fixed amount of proofreading.

We compared the tagging accuracies and reliabilities among different tagging algorithms, such as Markov bi-gram model, Bayesian classifier, and context-rule classifier. According to our experiments, for the best cost-effective tagging algorithm, in average, 20% of samples of ambivalence words need to be rechecked to achieve an estimated final accuracy of 99%. The tradeoffs between amount of proofreading and final accuracy for different algorithms are also compared. It concludes that an algorithm with highest accuracy may not always be the most reliable algorithm.

1 Introduction

Part-of-speech tagging for a large corpus is a labor intensive and time-consuming task. Most of time and labors were spent on proofreading and never achieved 100% accuracy, as exemplified by many public available corpora. Since manual proofreading is inevitable, how do we derive the most cost-effective tagging algorithm? To reduce efforts of manual editing, a new concept of reliable tagging was proposed. The idea is as follows. An evaluation score, as an indicator of tagging confidence, is made for each tagging decision. If a high confidence value is achieved, it indicates that this tagging decision is very likely correct. On the other hand, a low confidence value means the tagging result might require manual checking. If a tagging algorithm can provide a very reliable confidence evaluation, it means that most of high confidence tagging results need not manually checked. As a result, time and manual efforts for tagging processes can be reduced drastically. The reliability of a tagging algorithm is defined as follows.

Reliability = The estimated final accuracy achieved by the tagging model under the constraint that only a fixed amount target words with the lowest confidence value is manually proofread.

It is slightly different from the notion of tagging accuracy. It is possible that a higher accuracy algorithm might require more manual proofreading than a reliable algorithm with lower accuracy.

The tagging accuracies were compared among different tagging algorithms, such as Markov PoS bi-gram model, Bayesian classifier, and context-rule classifier. In addition, confidence measures of the tagging will be defined. In this paper, the above three algorithms are designed and the most cost-effective algorithm is also determined.

2 Reliability vs. Accuracy

The reported accuracies of automatic tagging algorithms are about 95% to 96% (Chang et al., 1993; Lua, 1996; Liu et al., 1995). If we can pinpoint the errors, only 4~5% of the target corpus has to be revised to achieve 100% accuracy. However, since the occurrences of errors are unknown, conventionally the whole corpus has to be reexamined. It is most tedious and time consuming, since a practically useful tagged corpus is at least in the size of several million words. In order to reduce the manual editing and speed up the construction process of a large tagged corpus, only potential errors of tagging will be rechecked manually (Kveton et al., 2002; Nakagawa et al., 2002). The problem is how we find the potential errors. Suppose that a probabilistic-based tagging method will assign a probability to each PoS of a target word by investigating the context of this target word *w*. The hypothesis is that if the probability $P(c_1 \mid w, context)$ of the top choice candidate c_1 is much higher than the probability $P(c_2 \mid w, context)$ of the second choice candidate c_2, then the confidence value assigned for c_1 is also higher. (Hereafter, for simplification, if without confusing, we will use $P(c)$ to stand for $P(c \mid w, context)$.) Likewise, if the probability $P(c_1)$ is closer to the probability $P(c_2)$, then the confidence value assigned for c_1 is also lower. We try to prove the above hypothesis by empirical methods. For each different tagging method, we define its confidence measure according to the above hypothesis and to see whether or not tagging errors are generally occurred at the words with low tagging confidence. If the hypothesis is true, we can proofread the auto-tagged results only on words with low confidence values. Furthermore, the final accuracy of the tagging after partial proofreading can also be estimated by the accuracy of the tagging algorithm and the amount of errors contained in the proofread data. For instance, a system has a tagging accuracy of 94% and supposes that K% of the target words with the lowest confidence scores covers 80% of errors. After proofreading those K% of words in the tagged words, those 80% errors are fixed. Therefore the reliability score of this tagging system of K% proofread will be 1 - (error rate) * (reduced error rate) = 1 - ((1 - accuracy rate) * 20%) = 1 - ((1 - 94%) * 20%) = 0.988. On the other hand, another tagging system has a higher tagging accuracy of 96%, but its confidence measure is not very reliable, such that the K% of the words with the lowest confidence scores contains only 50% of errors. Then the reliability of this system is 1 - ((1 - 96%) * 50%) = 0.980, which is lower than the first system. That is to say after spending the same amount of effort of manual proofreading, the first system achieves a better results even it has lower tagging accuracy. In other word, a reliable system is more cost-effective.

3 Tagging Algorithms and Confidence Measures

In this study, we are going to test three different tagging algorithms based on same training data and testing data, and to find out the most reliable tagging algorithm. The three tagging algorithms are

的(DE)	重要(VH)	**研究(Nv)**	機構(Na)	之(DE)
相當(Dfa)	重視(VJ)	**研究(Nv)**	開發(Nv)	，(COMMACATEGORY)
內(Ncd)	重點(Na)	**研究(Nv)**	需求(Na)	。(PERIODCATEGORY)
仍(D)	限於(VJ)	**研究(Nv)**	階段(Na)	。(PERIODCATEGORY)
民族(Na)	音樂(Na)	**研究(VE)**	者(Na)	明立國(Nb)
赴(VCL)	香港(Nc)	**研究(VE)**	該(Nes)	地(Na)
亦(D)	值得(VH)	**研究(VE)**	。(PERIODCATEGORY)	
合宜性(Na)	值得(VH)	**研究(VE)**	。(PERIODCATEGORY)	
更(D)	值得(VH)	**研究(Nv)**	。(PERIODCATEGORY)	

Table 1 Sample keyword-in-context file of the words ‘研究’ sorted by its left/right context

Markov bi-gram model, Bayesian classifier, and context-rule classifier. The training data and testing data are extracted from Sinica corpus, a 5 million word balanced Chinese corpus with PoS tagging (Chen et al., 1996). The confidence measure will be defined for each algorithm and the best accuracy will be estimated at the constraint of only a fixed amount of testing data being proofread.

It is easier to proofread and make more consistent tagging results, if proofreading processes were done by checking the keyword-in-context file for each ambivalence word and only the tagging results of ambivalence word need to be proofread. The words with single PoS need not be rechecked their PoS tagging. For instance, in Table 1, the keyword-in-context file of the word ‘研究’ (research), which has PoS of verb type *VE* and noun type *Nv*, is sorted according to its left/right context.

The proofreader can see the other examples as references to determine whether or not each tagging result is correct. If all of the occurrences of ambivalence word have to be rechecked, it is still too much of the work. Therefore only words with low confidence scores will be rechecked.

A general confidence measure was defined as the value of $\frac{P(c_1)}{P(c_1)+P(c_2)}$, where $P(c_1)$ is the probability of the top choice PoS c_1 assigned by the tagging algorithm and $P(c_2)$ is the probability of the second choice PoS c_2 [1]. The common terms used in the following tagging algorithms were also defined as follows:

w_k	The k-th word in a sequence
c_k	The PoS associated with k-th word w_k
$w_1c_1,...,w_nc_n$	A word sequence containing n words with their associated categories respectively

3.1 Markov Bi-gram Model

The most widely used tagging models are part-of-speech n-gram models, in particular bi-gram and tri-gram model. In a bi-gram model, it looks at pair of categories (or words) and uses the conditional probability of $P(c_k \mid c_{k-1})$, and the Markov assumption is that the probability of a PoS occurring depends only on the PoS before it.

Given a word sequence $w_1,...w_n$, the Markov bi-gram model searches for the PoS sequence $c_1,...c_n$ such that argmax $\Pi P(w_k \mid c_k)$ * $P(c_k \mid c_{k-1})$ is achieved. In our experiment, since we are

[1] Log-likelihood ratio of log P(c1)/P(c2) is another alternation of confidence measure. However, for some tagging algorithms, they may not necessary produce real probability estimation for each PoS, such as context-rule model. The scaling control for log-likelihood ratio will be hard for those algorithms. In addition, the range of our confidence score is between 0.5~1.0. Therefore, the above confidence value is adopted.

only focusing on the resolution of ambivalence words only, a twisted Markov bi-gram model was applied. For each ambivalence target word, its PoS with the highest model probability is tagged. The probability of each candidate PoS c_k for a target word w_k is estimated by $P(c_k | c_{k-1})$ * $P(c_{k+1} | c_k) * P(w_k | c_k)$. There are two approaches to estimate the statistical data for $P(c_k | c_{k-1})$ and $P(c_{k+1} | c_k)$. One is to count all the occurrences in the training data, and another one is to count only the occurrences in which each w_k occurs. According to the experiments, to estimate the statistic data using w_k dependent data is better than using all sequences. In other words, the algorithm tags the PoS c_k for w_k, such that c_k maximizes the probability of $P(c_k | w_k, c_{k-1})$ * $P(c_{k+1} | w_k, c_k) * P(w_k | c_k)$ instead of maximizing the probability of $P(c_k | c_{k-1}) * P(c_{k+1} | c_k)$ * $P(w_k | c_k)$.

3.2 Bayesian Classifier

The Bayesian classifier algorithm adopts the Bayes theorem (Manning et al., 1999) that swaps the order of dependence between events. That is, it calculates $P(c_{k-1} | c_k)$ instead of $P(c_k | c_{k-1})$. The probability of each candidate PoS c_k in Bayesian classifier is calculated by $P(c_{k-1} | w_k, c_k) * P(c_{k+1} | w_k, c_k)$ $* P(c_k | w_k)$. The Bayesian classifier tags the PoS c_k for w_k, such that c_k maximizes the probability of $P(c_{k-1} | w_k, c_k) * P(c_{k+1} | w_k, c_k) * P(c_k | w_k)$.

3.3 Context-Rule Model

Dependency features utilized in determining the best PoS-tag in both Markov and Bayesian models are categories of context words. As a matter of fact, for some cases the best PoS-tags might be determined by other context features, such as context words (Brill, 1992). In the context-rule model, broader scope of context information is utilized in determining the best PoS-tag. We extend the scope of the dependency context of a target word into its 2 by 2 context windows. Therefore the context features of a word can be represented by the vector of $[w_{-2}, c_{-2}, w_{-1}, c_{-1}, w_1, c_1, w_2, c_2]$. Each feature vector may be associated with a unique PoS-tag or many ambiguous PoS-tags. Their association probability of a possible PoS c'_0 is *P*($c'_0 \mid w_0$, *feature vector*). If for some (w_0 , c'_0), the value of *P*($c'_0 \mid w_0$, *feature vector*) is not 1, it means that the c_0 of w_0 cannot be uniquely determined by its context vector. Some additional features have to be incorporated to resolve the ambiguity. If for a word w_0, all of its PoS c'_0 such that the value of *P*($c'_0 \mid w_0$, *feature vector*) is zero which means there is no training examples with the same context vector of w_0. If the full scope of the context feature vector is used, data sparseness problem will seriously hurt the system performance. Therefore partial feature vectors are used instead of full feature vectors. The partial feature vectors applied in our context-rule classifier are w_{-1}, w_1, $c_{-2}c_{-1}$, c_1c_2, $c_{-1}c_1$, $w_{-2}c_{-1}$, $w_{-1}c_{-1}$, and c_1w_2.

At the training stage, for each feature vector type, many rule instances will be generated and their probabilities associated with PoS of the target word are also calculated. For instance, with the feature vector types of w_{-1}, w_1, $c_{-2}c_{-1}$, c_1c_2,…, we can extract rule patterns of w_{-1}(先生), w_1(之餘), $c_{-2}c_{-1}$ (*Nb*, *Na*), c_1c_2 (*Ng*, *COMMA*), ... etc, associated with the PoS *VE* of target word from the following sentence while the target word is '研究 research'.

周 Tsou (Nb) 先生 Mr (Na) 研究 research (VE) 之餘 after (Ng) ，(COMMA)

" After Mr. Tsou has done his research,"

By investigating all training data, various different rule patterns (associated with a candidate PoS of a target word) will be generated and their association probabilities of $P(c'_0 \mid w_0, \mathit{feature\ vector})$ are also derived. For instance, If we take those word sequences listed in Table 1 as training data and $c_{-1}c_1$ as feature pattern, and set '研究' as target word, we would train with a result containing a rule pattern = $c_{-1}c_1$ (*VH*, *PERIOD*) and derive the probabilities of *P*(*VE* | '研究', (*VH*, *PERIOD*)) = 2/3 and *P*(*NV* | '研究', (*VH*, *PERIOD*)) = 1/3. The rule patterns and their association probability will be utilized to determine the probability of each candidate PoS of a target word in a testing sentence. Suppose that the target word w_0 has ambiguous categories of $c_1, c_2, \ldots, c_n$, and the context patterns of $pattern_1, pattern_2, \ldots, pattern_m$, then the probability to assign tag c_i to the target word w_0 is defined as follows:

$$P(c_i) \cong \frac{\sum_{y=1}^{m} P(c_i \mid w, pattern_y)}{\sum_{x=1}^{n}\sum_{y=1}^{m} P(c_x \mid w, pattern_y)}$$

In other words, the probabilities of different patterns with the same candidate PoS are accumulated and normalized by the total probability distributed to all candidates as the probability of the candidate PoS. The algorithm will tag the PoS of the highest probability.

Word	Word Sense	Distribution Characteristics
了	an expletive in the Chinese	high frequency
將	get, be about to	average distribution of candidate categories
研究	research	high inconsistence of context information
改變	change	simply two candidate categories
採訪	interview, gather material	low frequency
演出	perform	extremely low frequency

Table 2 Target words used in the experiments

4 Tradeoffs between Amount of Manual Proofreading and the Best Accuracy

There is a tradeoff between amount of manual proofreading and the best accuracy. If the goal of tagging is to achieve an accuracy of 99%, then an estimated threshold value of confidence score to achieve the target accuracy will be given and the tagged word with confidence score less than this designated threshold value will be checked. On the other hand, if the constraint is to finish the tagging process under the constraints of limited time and manual labors, in order to achieve the best accuracy, we will first estimate the amount of partial corpus which can be proofread under the constrained time and labors, and then determine the threshold value of the confidence.

The six ambivalence words with different frequencies, listed in Table 2, were picked as our target words in the experiments. We like to see the tagging accuracy and confidence measure effected by variation of ambivalence and the amount of training data among selected target words. The Sinica corpus is divided into two parts as our training data and testing data. The training data contains 90% of the corpus, while the testing data is the remaining 10%.

Some words' frequencies are too low to have enough training data, such as the target words '採訪 interview' and '演出 perform'. To solve the problem of data sparseness, the Jeffreys-Perks law, or Expected Likehood Estimation (ELE) (Manning et al., 1999), is introduced as the smoothing method for all evaluated tagging algorithms. The probability $P(w_1,...,w_n)$ is defined as $\frac{C(w_1,...,w_n)}{N}$, where $C(w_1,...,w_n)$ is the amount that pattern $w_1,...,w_n$ occurs in the training data, and N is the total amount of all training patterns. To smooth for an unseen event, the probability of $P(w_1,...,w_n)$ is redefined as $\frac{C(w_1,...,w_n)+\lambda}{N+B\lambda}$, where B denotes the amount of all pattern types in training data and λ denotes the default occurrence count for an unseen event. That is to say, we assume a value λ for an unseen event as its occurrence count. If the value of λ is 0, it means that there is no smoothing process for the unseen events. The most widely used value for λ is 0.5, which is also applied in the experiments.

In our experiments, the confidence measure of the ratio of probability gap between top choice candidate and the second choice candidate $\frac{P(c_1)}{P(c_1)+P(c_2)}$ is adopted for all three different models. Figure 1 shows the result pictures of tradeoffs between amount of proofreading and the estimated best accuracies for the three different algorithms. Without any manual proofreading on result tags, the accuracy of context-rule algorithm is about 1.4% higher than the Bayesian classifier and Markov bi-gram model. As the percentage of manual proofreading increases, the accuracy of each algorithm

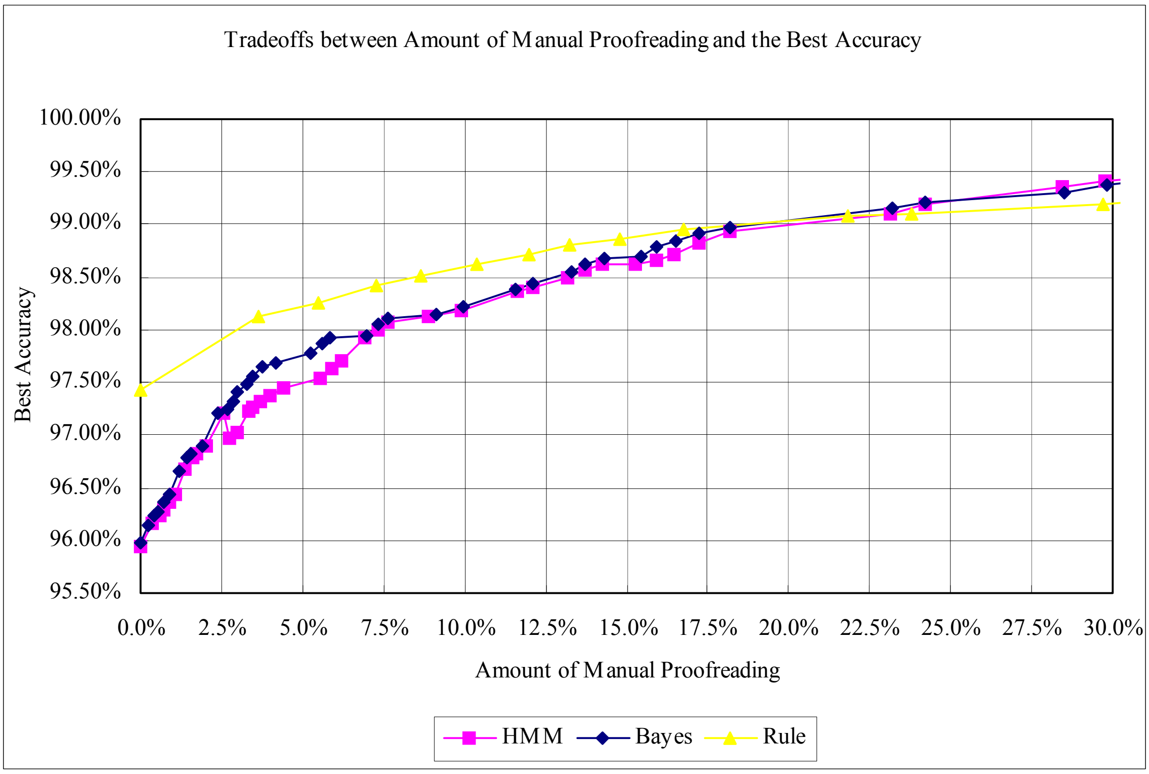


Figure 1 Tradeoffs between amount of manual proofreading and the best accuracy

increases, too. It is obvious to see that the accuracy of context-rule algorithm increases slower than those of other two algorithms while the amount of manual proofreading increases more. The values of best accuracy of three algorithms will meet in a point of 99% approximately, with around 20% of required manual proofreading on result tags. After the meeting point, Bayesian classifier and Markov bi-gram model will have higher value of best accuracy than context-rule classifier when the amount of manual proofreading is over 20% of the tagged results.

The result picture shows that if the required tagging accuracy is over 99% and there are plenty of labors and time available for manual proofreading, the Bayesian classifier and Markov bi-gram model would be better choices, since they have higher best accuracies than the context-rule classifier.

5 Conclusion

In this paper, we proposed a new way of finding the most cost-effective tagging algorithm. The cost-effective is defined in term of a criterion of reliability. The reliability of the system is measured in term of confidence score of ambiguity resolution of each tagging. For the best cost-effective tagging algorithm, in average, 20% of samples of ambivalence words need to be rechecked to achieve an accuracy of 99%. In other word, the manual labor of proofreading is reduced more than 80%.

In future, we like to extend the coverage of confidence checking for all words, including words with single PoS, to detect flexible word uses. The confidence measure for words with single PoS can be made by comparing the tagging probability of this particular PoS with all other categories.

Acknowledgement: The authors would like to thank the anonymous reviews' valuable comments and suggestions. The work is partially supported by the grant of NSC 92-2213-E-001-016.

References

C. H. Chang & C. D. Chen, 1993, "HMM-based Part-of-Speech Tagging for Chinese Corpora," in Proceedings of the Workshop on Very Large Corpora, Columbus, Ohio, pp. 40-47.

C. J. Chen, M. H. Bai, & K. J. Chen, 1997, "Category Guessing for Chinese Unknown Words," in Proceedings of NLPRS97, Phuket, Thailand, pp. 35-40.

Christopher D. Manning & Hinrich Schutze, Foundations of Statistical Natural Language Processing, The MIT Press, 1999, pp. 43-45, pp. 202-204.

E. Brill, “A Simple Rule-Based Part-of-Speech Taggers,” in Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing 1992, pp. 152–155.

K. J. Chen, C. R. Huang, L. P. Chang, & H. L. Hsu, 1996, “Sinica Corpus: Design Methodology for Balanced Corpora,” in Proceedings of PACLIC II, Seoul, Korea, pp. 167-176.

K. T. Lua, 1996, “Part of Speech Tagging of Chinese Sentences Using Genetic Algorithm,” in Proceedings of ICCC96, National University of Singapore, pp. 45-49.

P. Kveton & K. Oliva, 2002, “(Semi-) Automatic Detection of Errors in PoS-Tagged Corpora,” in Proceedings of Coling 2002, Taipei, Tai-wan, pp. 509-515.

S. H. Liu, K. J. Chen, L. P. Chang, & Y. H. Chin, 1995, “Automatic Part-of-Speech Tagging for Chinese Corpora,” on Computer Proceeding of Oriental Languages, Hawaii, Vol. 9, pp.31-48.

T. Nakagawa & Y. Matsumoto, 2002, “Detecting Errors in Corpora Using Support Vector Machines,” in Proceedings of Coling 2002, Taipei, Taiwan, pp.709-715.

Chinese Word Auto-Confirmation Agent

Jia-Lin Tsai, Cheng-Lung Sung and Wen-Lian Hsu
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
{tsaijl,clsung,hsu}@iis.sinica.edu.tw

Abstract

In various Asian languages, including Chinese, there is no space between words in texts. Thus, most Chinese NLP systems must perform word-segmentation (sentence tokenization). However, successful word-segmentation depends on having a sufficiently large lexicon. On the average, about 3% of the words in text are not contained in a lexicon. Therefore, unknown word identification becomes a bottleneck for Chinese NLP systems.

In this paper, we present a Chinese word auto-confirmation (CWAC) agent. CWAC agent uses a hybrid approach that takes advantage of statistical and linguistic approaches. The task of a CWAC agent is to auto-confirm whether an n-gram input (n ≥ 2) is a Chinese word. We design our CWAC agent to satisfy two criteria: (1) a greater than 98% precision rate and a greater than 75% recall rate and (2) domain-independent performance (F-measure). These criteria assure our CWAC agents can work automatically without human intervention. Furthermore, by combining several CWAC agents designed based on different principles, we can construct a multi-CWAC agent through a building-block approach.

Three experiments are conducted in this study. The results demonstrate that, for n-gram frequency ≥ 4 in large corpus, our CWAC agent can satisfy the two criteria and achieve 97.82% precision, 77.11% recall, and 86.24% domain-independent F-measure. No existing systems can achieve such a high precision and domain-independent F-measure.

The proposed method is our first attempt for constructing a CWAC agent. We will continue develop other CWAC agents and integrating them into a multi-CWAC agent system.

Keywords: natural language processing, word segmentation, unknown word, agent

1. Introduction

For a human being, efficient word-segmentation (in Chinese) and word sense disambiguation (WSD) arise naturally while a sentence is understood. However, these problems are still difficult for the computer. One reason is that it is hard to create unseen knowledge in the computer from running texts [Dreyfus 1992]. Here, unseen knowledge refers to contextual meaning and unknown lexicon.

Generally, the task of unknown lexicon identification is to identify (1) unknown word (2) unknown word sense, (3) unknown part-of-speech (POS) of a word and (4) unknown word pronunciation. Unknown word identification (UWI) is the most essential step in dealing with unknown lexicons. However, UWI is still quite difficult for Chinese NLP. From [Lin *et al.* 1993, Chang *et al.* 1997, Lai *et al.* 2000, Chen *et al.* 2002 and Sun *et al.* 2002], the difficulty of Chinese UWI is caused by the following problems:

1. Just as in other Asian languages, Chinese sentences are composed of strings of characters that do not have blank spaces to mark word boundaries.
2. All Chinese characters can either be a morpheme or a word. Take the Chinese character 花 as an example. It can be either a free morpheme or a word.
3. Unknown words, which usually are compound words and proper names, are too numerous to list in a machine-readable dictionary (MRD).

To resolve these issues, statistical, linguistic and hybrid approaches have been developed and investigated. For statistical approaches, researchers use common statistical features, such as maximum entropy [Yu *et al.* 1998, Chieu *et al.* 2002], association strength [Smadja 1993, Dunnin 1993], mutual information [Florian *et al.* 1999, Church 2000], ambiguous matching [Chen & Liu 1992, Sproat *et al.* 1996], and multi-statistical features [Chang *et al.* 1997] for unknown word detection and extraction. For linguistic approaches, three major types of linguistic rules (knowledge): morphology, syntax, and semantics, are used to identify unknown words. Recently, one important trend of UWI follows a hybrid approach so as to take advantage of both merits of statistical and linguistic approaches. Statistical approaches are simple and efficient whereas linguistic approaches are effective in identifying low frequency unknown words [Chang *et al.* 1997, Chen *et al.* 2002].

Auto-detection and auto-confirmation are two basic steps in most UWI systems. Auto-detection is used to detect the possible n-grams candidates from running texts for a better focus, so that in the next auto-confirmation stage, these identification systems need only focus on the set of possible n-grams. In most cases, recall and precision rates are affected by auto-detection and auto-confirmation. Since trade-off would occur between recall and precision, deriving a hybrid approach with precision-recall optimization has become a major challenge [Chang *et al*. 1997, Chen *et al*. 2002].

In this paper, we introduce a Chinese word auto-confirmation (CWAC) agent, which uses a hybrid approach to effectively eliminate human intervention. A CWAC agent is an agent (program) that automatically confirms whether an n-gram input is a Chinese word. We design our CWAC agent to satisfy two criteria: (1) a greater than 98% precision rate and a greater than 75% recall rate and (2) domain-independent performance (F-measure). These criteria assure our CWAC agents can work automatically without human intervention. To our knowledge, no existing system has yet achieved the above criteria.

Furthermore, by combining several CWAC agents designed based on different principles, we can construct a multi-CWAC agent through a building-block approach and service-oriented architecture (such as web services [Graham *et al*. 2002]). Figure 1 illustrates one way of a multi-CWAC agent system combining three CWAC agents. If the number of identified words of a multi-CWAC agent is greater than that of its any single CWAC agent, we believe a multi-CWAC agent could be able to maintain the 98% precision rate and increase its recall rate by merely integrating with more CWAC agents.

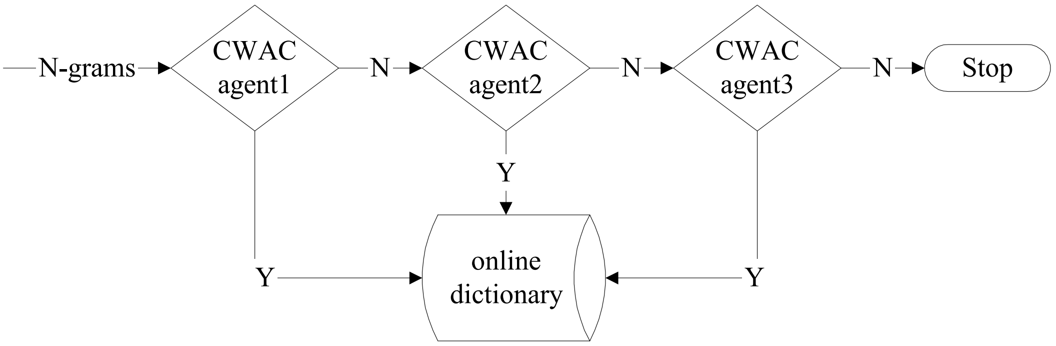


Figure 1. An illustration of a multi-CWAC agent system

This article is structured as follows. In Section 2, we will present a method for simulating a CWAC agent. Experimental results and analyses of the CWAC agent will be presented in section 3. Conclusion and future directions will be discussed in Section 4.

2. Development of the CWAC agent

The most frequent 50,000 words were selected from the CKIP lexicon (CKIP [1995]) to create the system dictionary. From this lexicon, we only use word and POS for our algorithm.

2.1 Major Processes of the CWAC Agent

A CWAC agent automatically identifies whether an n-gram input (or, say, n-char string) is a Chinese word. In this paper, an n-gram extractor is developed to extract all n-grams (n ≥ 2 and n-gram frequency ≥ 3) from test sentences as the n-gram input for our CWAC agent (see Figure 2). (Note that n-gram frequencies vary widely according to test sentences.)

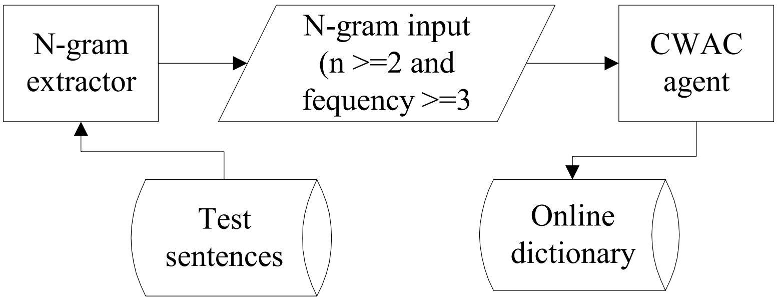


Figure 2. An illustration of n-gram extractor and CWAC agent

Figure 3 is the flow chart of the CWAC agent in which the major processes are labeled (1) to (6). The confirmation types, brief descriptions and examples of the CWAC agent, are given in Table 1. We apply linguistic approach, statistical approach and LFSL (linguistic first, statistical last) approach to develop the CWAC agent. Note in Figure 3, the processes (5) and (6) are statistical methods, and the remaining four processes are developed from linguistic knowledge. The LFSL approach means a combining process of a linguistic process (such as process 4) and a statistical process (such as process 5).

The details of these major processes are described below.

Process 1. System dictionary checking*:* If the n-gram input can be found in the system dictionary, it will be labeled **K0** (which means that the n-gram exists in the system dictionary). In Table 1, the n-gram 計程車 is a system word.

Process 2. Segmentation by system dictionary*:* In this stage, the n-gram input will be segmented by two strategies: left-to-right longest word first (LR-LWF), and right-to-left longest word first (RL-LWF). If LR-LWF and RL-LWF segmentations of the n-gram input are different, the CWAC agent will be triggered to compute the products of all word length for these

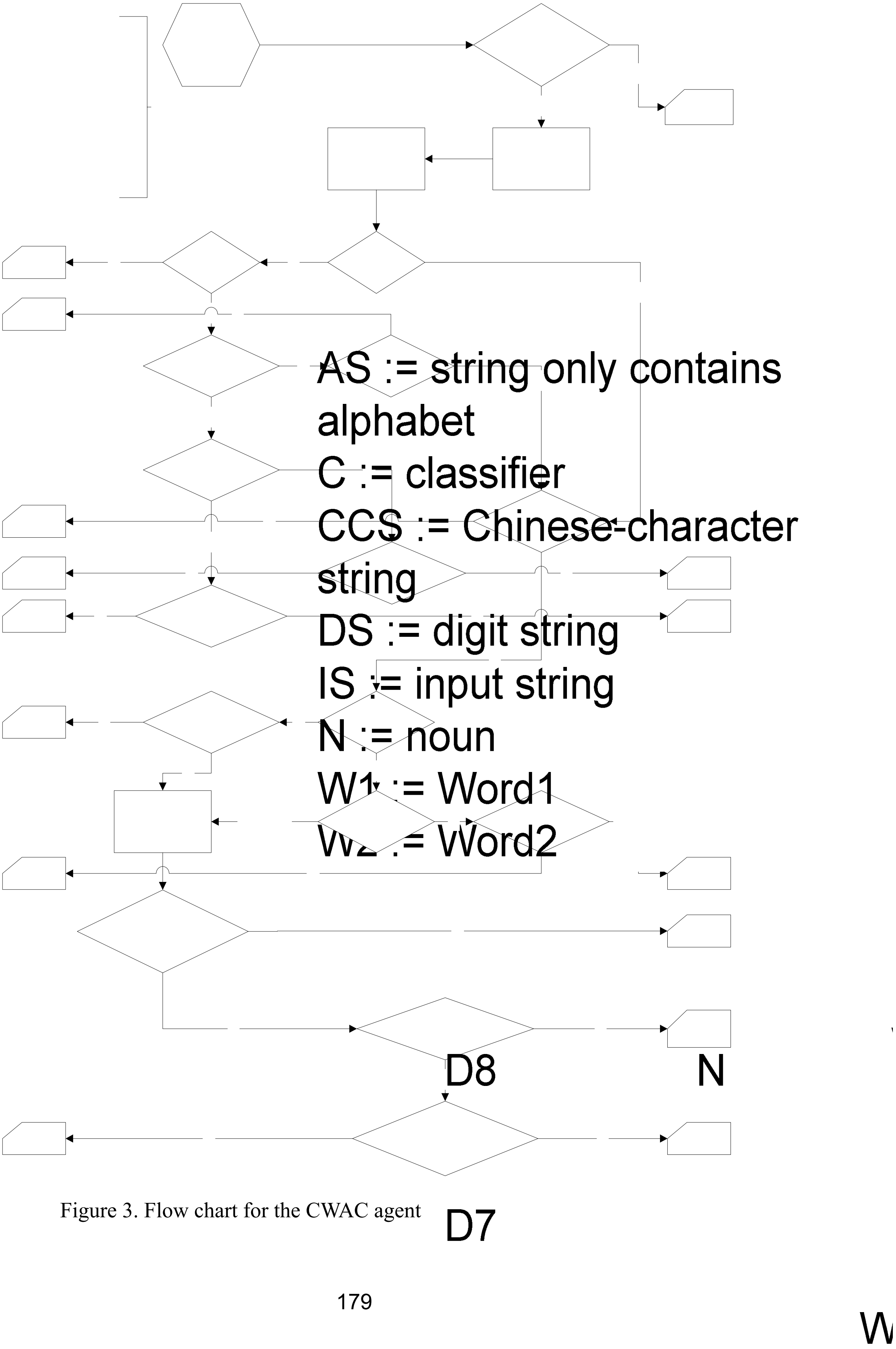


Figure 3. Flow chart for the CWAC agent

Table 1. Confirming results, types, descriptions and examples of the CWAC agent (The symbol / stands for word boundary according to system dictionary using RL-LWF)

Auto-Confirming Results	Types	Brief Descriptions	Examples	
			Input	Output
Word	K0	N-gram exists in system dictionary	計程車	計程車 [1]
	K1	Both polysyllabic words exist in online dictionary	接駁公車	接駁/公車 [1]
	K2	Two polysyllabic word compounds	食品公司	食品/公司 [1]
	K3	Both first and last word of segmented N-gram are polysyllabic words and N ≧ 3	東港黑鮪魚	東港/黑/鮪魚 [1]
	K4	Segmentation ambiguity is ≦ 50%	腸病毒	腸/病毒 [1]
	K5	N-gram frequency exceeds 10	阿爾巴尼亞裔	阿/爾/巴/尼/亞裔 [1]
Not Word	D1	Two polysyllabic word compounds with at least function word	問題一直	問題/一直 [2]
	D2	N-gram contains function word	市場指出	市場/指出 [2]
	D5	Segmentation ambiguity is > 50%	台北市立	台北市/立 [2]
	D6	Suffix Chinese digit string	隊伍	隊/伍 [1]
	D7	Digits suffix polysyllabic word	5 火鍋	5/火鍋 [2]
	D8	N-gram is a classifier-noun phrase	名學生	名/學生 [2]
	D9	N-gram includes unknown symbol	@公司	@/公司 [2]
	D0	Unknown reason	[3]	[3]

[1] These n-grams were manually confirmed as *is-word* in this study
[2] These n-grams were manually confirmed as *non-word* in this study
[3] There were no auto-confirming types "D0" and "K0" in this study

segmentations. If both products are equal, the RL-LWF segmentation will be selected. Otherwise, the segmentation with the greatest product will be selected. According to our experiment, the segmentation precision of RL-LWF is, on the average, 1% greater than that of LR-LWF. Take n-gram input 將軍用的毛毯 as an example. Its LR-LWF and RL-LWF segmentations are 將軍/用/的/毛毯 and 將/軍用/的/毛毯, respectively. Since both products are equal (2x1x1x2=1x2x1x2), the selected segmentation output for this process is 將/軍用/的/毛毯 as it is the RL-LFW.

Process 3. Stop word checking**:* The segmentation output from ***Process 2 is referred to as *segmentation2*. In this stage, all words in *segmentation2* will be compared with the stop word list. There are three types of stop words: **begining**, **middle**, and **end**. The stop word list used in this study is given in Appendix A. These stop words were selected by native Chinese speakers according to those computed beginning, middle, and end single-character words with < 1% of being the beginning, middle, or end words of Hownet [Dong 1999], respectively. If the first and last words of *segmentation2* can be found on the list of begining and end stop words, they will be eliminated from the *segmentation2*. For those cases in which the word number of *segmentation2* is greater than 2, middle stop word checking will be triggered. If a middle word in *segmentation2* can be found in the middle stop word list, the n-gram input will be split into new strings at any matched stop word. These new strings will be sent to ***Process 1*** as new n-gram input. For example, *segmentation2* of the n-gram input 可怕的腸病毒” is 可怕/的/腸/病毒. Since there is a middle stop word “的” in this *segmentation2*, the new strings 可怕 and 腸病毒 will be sent to ***Process 1*** as new n-gram input.

Process 4. Part-of-Speech (POS) pattern checking*:* Once *segmentation2* has been processed by ***Process 3***, the result is called *segmentation3*. If the word number of *segmentation3* is 2, POS pattern checking will be triggered. The CWAC agent will first generate all possible POS combinations of the two words using the system dictionary. If the number of generated POS combinations is one and that combination matches one of the POS patterns (**N/V**, **V/N**, **N/N**, **V/V**, **Adj/N**, **Adv/N**, **Adj/V**, **Adv/V**, **Adj/Adv**, **Adv/Adj**, **Adv/Adv** and **Adj/Adj**) the 2-word string will be tagged as a word and sent to ***Process 5***. This rule-based approach combines syntax knowledge and heuristic observation in order to identify compound words. For example, since the generated POS combination for *segmentation3* 食品/公司 is **N/N**, 食品公司 will be sent to ***Process 5***.

Process 5. Segmentation ambiguity checking*:* This stage consists of 4 steps:

1) Thirty randomly selected sentences that include the n-gram input will be extracted from either a large scale or a fixed size corpus. For example, the Chinese sentence “人人做環保” is a selected sentence that includes the n-gram input “人人”. The details of large scale and fixed size corpus used in this study will be addressed on Subsection 3.2. (Note that the number of selected sentences may be less than thirty and may even be zero due to corpus sparseness.)
2) These selected sentences will be segmented using the system dictionary, and will be segmented by the RL-LWF and LR-LWF technique.
3) For each selected sentence, if the RL-LWF and LR-LWF segmentations are different, the sentence will be regarded as an ambiguous sentence. For example, the Chinese sentence “人人做環保” is not an ambiguous sentence.
4) Compute the ambiguous ratio of ambiguous sentences to selected sentences. If the ambiguous ratio is less than 50%, the n-gram input will be confirmed as word type **K1**, **K2** or **K4** by ***Process 5*** (see Fig. 3) ; other-

wise, it will be labeled **D1** or **D2**. According to our observation, the ambiguous ratios of non word n-grams usually are greater than 50%.

Process 6. Threshold value checking**:* In this stage, if the frequency of an n-gram input is greater or equal to 10, it will be labeled as word type **K5** by ***Process 6. According to our experiment, if we directly regard an n-gram input whose frequency is greater than or equal to a certain threshold value as a word, the trade-off frequency of 99% precision rate occurs at the threshold value 7.

3. Experiment Results

The objective of the following experiments is to investigate the performance of the CWAC agent. By this objective, in ***process1*** of the CWAC agent, if an n-gram input is found to be a system word, a temporary system dictionary will be generated. The temporary system is the original system dictionary without this n-gram input. In this case, the n-gram input will be sent to ***process2*** and the temporary system dictionary will be used as system dictionary in both ***process2*** and ***process5***.

Three experiments are conducted in this study. Their results and analysis are given in Sections 3.3, 3.4 and 3.5.

3.1 Notion of Word and Evaluation Equations

The definition of word is not unique in Chinese [Sciullo *et al*. 1987, Sproat *et al*., 1996, Huang *et al*. 1997, Xia 2000]. As of our knowledge, the Segmentation Standard in China [Liu. *et al*. 1993] and the Segmentation Standard in Taiwan [CKIP 1996] are two of the most famous word-segmentation standards to Chinese. Since the Segmentation Guidelines for the Penn Chinese Treebank (3.0) [Xia 2000] has tried to accommodate the above famous segmentation standards in it, this segmentation was selected as our major guidelines for determining Chinese word. The notion of word in this study includes fixed-phrase words (such as 春夏秋冬, 你一句我一句, 奧林匹克運動會, etc.) , compounds (such as 腳踏車龍頭, 太陽眼鏡, etc.) and simple words (such as 房子, 老頭兒, 盤尼西林, etc.).

We use recall, precision and F-measure to evaluate the overall performance of the CWAC agent [Manning *et al*. 1999]. Precision, recall and F-measure are defined below. Note that the words in following equations (1) and (2) include new words and dictionary words.

$$recall = \#\ of\ correctly\ identified\ words\ /\ \#\ of\ words \quad (1)$$
$$precision = \#\ of\ correctly\ identified\ words\ /\ \#\ of\ identified\ words \quad (2)$$
$$F\text{-}measure = (2 \times recall \times precision)\ /\ (recall + precision) \quad (3)$$

3.2 Large Scale Corpus and Fixed Size Corpus

In Section 2, we mentioned that the corpus used in ***process5*** of the CWAC agent can be large scale or fixed size. The description of a large scale and a fixed size corpus is given below.

(1) ***Large scale corpus***: In our experiment, texts are collected daily. Texts collected in most Chinese web sites can be used as a large scale corpus. Here, we select **OPENFIND** [OPENFIND], one of the most popular Chinese search engines, to act as a large scale corpus. If ***process 5*** of the CWAC agent is in large scale corpus mode, it will extract the first thirty matching sentences, including the n-gram input, from the **OPENFIND** search results.

(2) ***Fixed size corpus***: A fixed size corpus is one whose text collection is limited. Here, we use a collection of 14,164,511 Chinese sentences extracted from whole 2002 articles obtained from *United Daily News (UDN)* web site [UDN] as our fixed size corpus, called 2002 *UDN* corpus.

3.3 The First Experiment

The objective of the first experiment is to investigate whether our CWAC agent satisfies criterion 1: the precision rate should be greater than 98% and the recall greater than 75%.

First, we create a testing corpus, called 2001 *UDN* corpus, consisting of 4,539,624 Chinese sentences extracted from all 2001 articles on the *UDN* Web site. The testing corpus includes 10 categories: 地方(local), 股市(stock), 科技(science), 旅遊(travel), 消費(consuming), 財經(financial), 國際(world), 運動(sport), 醫藥(health) and 藝文(arts). For each category, we randomly select 10,000 sentences to form a test sentence set. We then extract all n-grams from each test sentence set. We then obtain 10 test n-gram sets. All of the extracted n-grams have been manually confirmed as three types: *is-word*, *unsure-word* or *non-word*. In this study, the average percentages of n-grams manually confirmed as *is-word*, *unsure-word*, and *non-word* are 78%, 2% and 20%, respectively. When we compute precision, recall and F-measure, all *unsure-word* n-grams are excluded. Table 2 shows the results of the

CWAC agent in large scale corpus mode. Table 3 shows the results of the CWAC agent in fixed size corpus mode.

Table 2. The first experimental results of the CWAC agent in large scale corpus mode

Large scale Corpus						
n-grams frequency ≥ 3				n-grams frequency ≥ 4		
Class	P	R	F	P	R	F
地方	97.72%	76.37%	85.74%	98.54%	76.27%	85.99%
股市	94.32%	74.40%	83.19%	95.32%	75.51%	84.26%
科技	96.51%	76.33%	85.24%	97.64%	76.54%	85.81%
旅遊	97.51%	77.80%	86.55%	98.13%	78.09%	86.97%
消費	97.85%	79.41%	87.67%	98.56%	78.72%	87.53%
財經	95.68%	74.63%	83.86%	97.32%	75.74%	85.18%
國際	96.41%	78.64%	86.62%	97.26%	78.36%	86.79%
運動	94.17%	78.99%	85.92%	95.08%	78.66%	86.10%
醫藥	96.80%	78.09%	86.44%	98.60%	76.85%	86.38%
藝文	96.94%	76.87%	85.75%	98.20%	76.44%	85.96%
Avg.	96.31%	77.18%	85.69%	97.82%	77.11%	86.24%

Table 3. The first experimental results of the CWAC agent in fixed size corpus mode

Fixed size Corpus						
n-grams frequency ≥ 3				n-grams frequency ≥ 4		
Class	P	R	F	P	R	F
地方	97.93%	73.46%	83.95%	98.37%	73.91%	84.41%
股市	95.76%	69.60%	80.61%	96.63%	70.30%	81.39%
科技	97.70%	69.01%	80.89%	98.15%	68.99%	81.03%
旅遊	97.95%	70.09%	81.71%	98.61%	70.49%	82.21%
消費	98.20%	74.76%	84.89%	98.79%	74.73%	85.09%
財經	97.02%	67.41%	79.55%	97.76%	68.56%	80.60%
國際	97.06%	73.56%	83.69%	97.81%	73.00%	83.60%
運動	95.77%	74.03%	83.51%	97.02%	74.96%	84.57%
醫藥	97.68%	71.72%	82.71%	98.26%	71.64%	82.87%
藝文	98.22%	70.20%	81.88%	99.02%	69.40%	81.61%
Avg.	97.32%	71.44%	82.39%	98.11%	71.61%	82.79%

As shown in Table 2, the CWAC agent in large scale corpus mode can achieve 96.31% and 97.82% precisions, 77.18% and 77.11% recalls and 85.69% and 86.24%

F-measures for n-gram frequencies of ≥ 3 and ≥ 4, respectively. Table 3 shows that the CWAC agent in fixed size corpus mode can achieve 97.32% and 98.11% precisions, 71.44 and 71.61% recalls and 82.39% and 82.79% F-measures.

The hypothesis tests of whether the CWAC agent satisfies criterion 1, **H1a** and **H1b**, for this experiment are given below. (One-tailed t-test, reject H_0 if its p-value > 0.05)

H1a. H_0: avg. precision ≤ 98%, H_1: avg. precision > 98%

H1b. H_0: avg. recall ≤ 77%, H_1: avg. recall > 77%

From Tables 2 and 3, we compute the p-values of **H1a** and **H1b** for four CWAC modes in Table 4. Table 4 shows that the CWAC agent passes both hypotheses **H1a** and **H1b** in large scale corpus mode with an n-gram frequency of ≥ 4.

In Chen *et al.* (2002), a word that occurs no less than three times in a document is a high frequency word; otherwise, it is a low frequency word. Since a low frequency word in a document could be a high frequency word in our test sentence sets, the results in Tables 2 and 3 can be regarded as an overall evaluation of UWI for low and high frequency words.

Table 4. The p-values of the hypothesis tests, **H1a** and **H1b**, for four CWAC modes

CWAC mode	P-value (**H1a**)	P-value (**H1b**)
Large scale & Frequency ≥ 3	0.0018 (accept H_0)	0.3927 (reject H_0)
Large scale & Frequency ≥ 4	0.1114 (reject H_0)	0.3842 (reject H_0)
Fixed size & Frequency ≥ 3	0.0023 (accept H_0)	0.0 (accept H_0)
Fixed size & Frequency ≥ 4	0.4306 (reject H_0)	0.0 (accept H_0)

In Chen *et al.* (2002), researchers try to use as much information as possible to identify unknown words in hybrid fashion. Their results have 88%, 84% and 89% precision rates; 67%, 82% and 68% recall rates; 76%, 83%, 78% F-measure rates on low, high, and low/high frequency unknown words, respectively.

3.3.1 A Comparative Study

Table 5 compares some of the famous works on UWI (here, the performance of our CWAC agent was computed solely against "new words" exclude words that are already in system dictionary). In Table 5, the system of [Chen *et al.* 2002] is one of the most famous hybrid approaches on unknown word extraction. Although Lai's system [Lai *et al.* 2000] achieves the best F-measure 88.45%, but their identifying target (including words and phrases) is different from conventional UWI system. Thus, Lai's result is not included in Table 5.

Table 5. Comparison of works on UWI

System	Method	Target	Test size	P	R	F
[Our CWAC]	Hybrid	n-gram word	100,000 sentences	94.32	74.50	83.25
[Chen *et al*. 2002]	Hybrid	n-gram word	100 documents	89	68	77.10
[Sun et al. 2002]	Statistical	name entity	MET2 (Chen *et al*. 1997)	77.89	86.09	81.79
[Chang *et al*. 1997]	Statistical	bi-gram word	1,000 sentences	72.39	82.83	76.38

3.4 Second Experiment

The objective of this experiment is to investigate whether the CWAC agent satisfies criterion 2: the F-measure should be domain-independent.

The hypothesis test **H2** for this experiment is given below. (Two-tailed t-test, reject H_0 if its p-value < 0.05)

H2. H_0: avg. F-measure = μ_0; H_1: avg. F-measure $\neq \mu_0$

Table 6 lists the p-values of **H2** for four CWAC modes. Table 6 shows that the CWAC agent passes H2 and satisfies criterion 2 in all four CWAC modes.

Table 6. The p-values of the hypothesis test **H2** for four CWAC modes

CWAC mode	μ_0 (F-measure)	P-value
Large scale & Frequency ≥ 3	86%	0.4898 (accept H_0)
Large scale & Frequency ≥ 4	86%	0.7466 (accept H_0)
Fixed size & Frequency ≥ 3	83%	0.2496 (accept H_0)
Fixed size & Frequency ≥ 4	83%	0.6190 (accept H_0)

Summing up the results of first and second experiments, we conclude that our method can be used as a CWAC agent in large scale corpus mode when an n-gram frequency is ≥ 4.

3.5 Third Experiment

The objective of this experiment is to investigate whether the precision of our

CWAC agent is corpus-independent (**Q1**) and whether its recall is corpus-dependent (**Q2**). We use large scale and fixed size corpus modes to test **Q1** and **Q2**.

The hypothesis tests, **H3a** and **H3b**, for this experiment are given below. (Two-tailed t-test, reject H_0 if its p-value < 0.05)

H3a.H_0: avg. precision of large scale ($\mu1$) = avg. precision of fixed size ($\mu2$)
H_1: avg. precision of large scale ($\mu1$) $\neq$ avg. precision of fixed size ($\mu2$)

H3b.H_0: avg. recall of large scale ($\mu3$) = avg. recall of fixed size ($\mu4$)
H_1: avg. recall of large scale ($\mu3$) $\neq$ avg. recall of fixed size ($\mu4$)

Table 7 lists the p-values of **H3a** and **H3b** for n-gram frequencies of ≥ 3 and ≥ 4. Table 7 shows that **H3a** is accepted at the 5% *significance* level. This shows that the precision of the CWAC agent is corpus-independent, since the average precisions of both corpus modes equal at the 5% level. On the other hand, **H3b** is rejected at the 5% *significance* level. This shows the recall is corpus-dependent, since the average recalls of both corpus modes are not equal at the 5% level.

Table 7. The p-values of the hypothesis tests, **H3a** and **H3b**, for two frequency modes

Frequency mode	P-value (H3a)	P-value (H3b)
Frequency ≥ 3	0.079392 (accept H_0)	0.0000107 (reject H_0)
Frequency ≥ 4	0.238017 (accept H_0)	0.0000045 (reject H_0)

Tables 8 and 9 were created to sum up the experimental results in Tables 2 and 3. Table 8 gives the comparison of the linguistic, statistic and LFSL approaches in this study. From Table 8, it shows that the CWAC agent using the technique of LFSL achieves the best optimization of precision-and-recall with the greatest F-measure. Table 9 is the overall experimental results of the CWAC agent for n-gram frequencies of ≥ 3 to ≥ 10. From Table 9, it indicates the precisions, recalls and F-measures of the CWAC agent are close for different n-gram frequency conditions.

Table 8. Comparison of the linguistic, statistical and LFSL approaches results

N-grams frequency	Approach[1]	Precision (large, fixed)[2]	Recall (large, fixed)	F-measure (large, fixed)
≥ 3	Linguistic (L)	92.44%, 93.71%	67.41%, 48.96%	77.96%, 64.31%
≥ 3	Statistical (S)	89.15%, 100.00%	4.67%, 3.39%	8.88%, 6.56%
≥ 3	LFSL	96.72%, 97.43%	98.27%, 97.24%	97.49%, 97.34%

[1] The linguistic approaches include auto-confirmation types K3, D6, D7, D8 and D9; the statistical approaches include auto-confirmation types K1, K5, D1 and D5; the LFSL (linguistic approach first, statistical approach last) approaches include auto-confirmation types K2, K4 as shown in Fig. 3

[2] "large" means large scale corpus mode and "fixed" means fixed size corpus mode

Table 9. Overall experiment results

N-grams frequency	# of n-grams	Precision (large, fixed)[1]	Recall (large, fixed)	F-measure (large, fixed)
≥ 3	70502	96.31%, 97.32%	77.18%, 71.44%	85.69%, 82.39%
≥ 4	49500	97.82%, 98.11%	77.11%, 71.61%	86.24%, 82.79%
≥ 5	38179	97.49%, 98.52%	77.11%, 71.78%	86.11%, 83.05%
≥ 6	31382	97.64%, 98.76%	76.78%, 71.78%	85.96%, 83.14%
≥ 7	26185	97.77%, 99.00%	76.50%, 71.52%	85.84%, 83.05%
≥ 8	22573	97.86%, 99.11%	76.23%, 71.48%	85.70%, 83.06%
≥ 9	19473	97.84%, 99.16%	75.60%, 70.99%	85.29%, 82.74%
≥ 10	17048	97.72%, 99.17%	75.26%, 70.96%	85.03%, 82.73%

[1] "large" means large scale corpus mode and "fixed" means fixed size corpus mode

4. Conclusion and Directions for Future Research

UWI is the most important problem in handling unknown lexicons in NLP systems. A lexicon consists of words, POSs, word senses and word pronunciations. As shown in [Lin *et al*. 1993, Chang *et al*. 1997, Lai *et al*. 2000, Chen *et al*. 2002 and Sun *et al*. 2002], UWI is still a very difficult task for Chinese NLP systems. One important trend toward resolving unknown word problems is to follow a hybrid approach by combining the advantages of statistical and linguistic approaches. One of the most critical issues in identifying unknown words is to overcome the problem of precision-and-recall trade-off.

In this paper, we create a CWAC agent adopting a hybrid method to auto-confirm n-gram input. Our experiment shows that the LFSL (linguistic approach first, statistical approach last) approach achieves the best precision-and-recall optimization. Our results demonstrate that, for n-gram frequency ≥ 4 in large corpus mode, our CWAC agent can achieve 97.82% precision, 77.11% recall, and 86.24% F-measure. Thus, it satisfies that two criteria. Moreover, we discover that the use of large scale corpus in this method increases recall but not precision. On the other hand, we find that the precision of using either a large scale corpus or a fixed size corpus is not statistical significantly different at the 5% level.

This method is our first attempt to create a CWAC agent. We have also considered a building-block approach to construct a multi-CWAC agent. We believe a multi-CWAC agent could be able to maintain the 98% precision rate and increase recall rate by integrating more CWAC agents.

In the future, we will continue addressing agent-oriented and service-oriented approaches for handling unknown lexicons, such as unknown word POS auto-tagging agent and unknown word-sense auto-determining agent. Furthermore, the method to achieve corpus-independent recall will also be considered.

5. Acknowledgements

We are grateful to the our colleagues in the Intelligent Agent Systems Lab., Gladys Hsieh, Yi-Fan Chang and Yin-Ru Chen, who helped us to manually-confirm n-grams and create tools for this study.

References

Chen, K.J. and W.Y. Ma, "Unknown Word Extraction for Chinese Documents," *Proceedings of 19th COLING 2002*, Taipei, pp.169-175

Chieu, H.L. and H.T. Ng, "Named Entity Recognition: A Maximum Entropy Approach Using Global Information," *Proceedings of 19th COLING 2002*, Taipei, pp.190-196

Chang, J.S. and K.Y. Su, "An Unsupervised Iterative Method for Chinese New Lexicon Extraction," *International Journal of Computational Linguistics & Chinese language Processing*, 1997

Chen, K.J. and S.H. Liu, "Word Identification for mandarin Chinese Sentences," *Proceedings of 14th COLING*, pp. 101-107

Church, K.W., "Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p*p," *Proceedings of 18th COLING 2000*, pp.180-186

CKIP (Chinese Knowledge Information processing Group), Technical *Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Taiwan, Taipei, Academia Sinica, 1995. http://godel.iis.sinica.edu.tw/CKIP/r_content.html

CKIP (Chinese Knowledge Information processing Group), *A study of Chinese Word Boundaries and Segmentation Standard for Information processing (in Chinese)*. Technical Report, Taiwan, Taipei, Academia Sinica, 1996.

Dreyfus, H.L., *What computers still can't do: a critique of artificial reason*, Cambridge, Mass. : MIT Press, 1992

Dong, Z. and Q. Dong, Hownet, 1999, http://www.keenage.com/

Dunnin, T., "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, n 1., 1993

Florian, R. and D. Yarowsky, "Dynamic nonlocal language modeling via hierarchical topic-based adaptation," Proceedings of ACL99, 1999, pp. 167-174

Graham, S., S. Simeonov, T. Boubez, D. Davis, G. Daniels, Y. Nakamura and R. Neyama, *Building Web Services With Java*, Pearson Education, 2002

Huang, C.R., Chen, K.C., Chen, F.Y. and Chang, L.L., “Segmentation Standard for Chinese natural language Processing,” Computational Linguistics and Chinese Language Processing, 2(2), Aug., 1997, pp.47-62

Lai, Y.S. and Wu, C.H., “Unknown Word and Phrase Extraction Using a Phrase-Like-Unit-based Likelihood Ratio,” *International Journal of Computer Processing Oriental Language*, 13(1), pp.83-95

Lin, M.Y., T.H. Chiang and K.Y. Su, “A preliminary Study on Unkown Word Problem in Chinese Word Segmentation,” *Proceedings of ROCLING VI*, 1993, pp. 119-137

Manning, C.D. and Schuetze, H., *Fundations of Statistical Natural Language Processing*, MIT Press, 1999, pp.534-538

OPENFIND, OPENDFIN Chinese Search Web Site, http://www.openfind.com.tw/

Sciullo, A.M.D. and Williams, E., *On the Definition of Word*, MIT press, 1987

Smadja, F., “Retrieving Collocations from Text: Xtract,” *Computational Linguistics*, 22(1)

Sproat, R., C. Shih, W. Gale and N. Chang, “A Stochastic Finite-State Word-Segmentation Algorithm for Chinese,” *Computational Linguistics*, 22(3), 1996, pp.377-404

Sun, J., J. Gao, L. Zhang, M. Zhou and C. Huang, “Chinese Named Entity Identification Using Class-based Language Model,” *Proceedings of 19th COLING 2002*, Taipei, pp.967-973

UDN, On-Line United Daily News , http://udnnews.com/NEWS/

Xia, F., *The Segmentation Guidelines for the Penn Chinese Treebank (3.0)*, October 17, 2000

Yu, S., S. Bai, and P. Wu, “Description of the Kent Ridge Digital Labs System Used for MUC-7,” *Proceedings of the 7th Message Understanding Conference*, 1998

Appendix A. Stop Words List

I. Begining stop word list

/兒/呀/嗎/吧/呢/呼/了/是/你/我/他/又/等/既/或/有/到/去/在/爲/
/及/和/與/之/的/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/
/未/能/將/此/可/與/到/向/以/用/乃/入/又/下/久/乎/者/小/已/才/
/互/仍/勿/太/欠/且/乎/去/只/必/再/吁/多/好/如/早/而/至/行/但/
/別/即/吧/呀/更/沒/矣/並/和/呢/或/所/則/卻/哉/很/後/怎/既/甚/
/皆/相/若/唷/哼/哩/唉/哦/啊/得/都/最/喂/喔/喳/喲/等/著/嗎/嗨/
/嗚/嗡/愈/跟/較/過/嘛/嘎/嘟/嘻/嘿/噓/噗/罷/噹/噯/還/雖/嚕/

II. Middle stop word list

/可/已/各/被/到/等/既/但/且/而/並/同/又/爲/是/有/或/及/和/與/
/之/的/在/的/在/以/已/將/與/和/是/及/也/或/之/於/由/都/並/卻/
/且/只/則/但/又/才/仍/該/各/其/有/時/前/後/上/中/下/再/更/不/
/很/最/多/非/稍/否/至/了/吧/嗎/但/因/爲/而/且/就/對/雖/裡/裏/
/等/要/把/到/去/給/打/做/作/個/你/妳/我/他/她/它/們/這/那/此/
/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/未/能/將/此/可/
/與/到/向/以/用/乃/入/又/下/久/乎/者/已/互/仍/勿/欠/且/乎/去/
/只/必/再/吁/多/好/如/早/而/至/但/別/即/吧/呀/更/沒/矣/並/呢/
/或/所/則/卻/哉/很/後/怎/既/甚/皆/相/若/唷/哼/哩/唉/哦/啊/得/
/都/最/喂/喔/喳/喲/等/著/嗎/嗨/嗚/嗡/愈/跟/較/過/嘛/嘎/嘟/嘻/
/嘿/噓/噗/罷/噹/噯/還/雖/嚕/

III. End stop word list

/等/及/與/的/是/個/不/的/有/要/對/於/就/了/爲/也/在/及/之/未/
/能/將/此/可/會/與/到/向/以/用/乃/入/又/下/久/乎/者/小/已/才/
/互/仍/勿/太/欠/且/乎/去/只/必/再/吁/多/好/如/早/而/至/行/但/
/別/即/吧/呀/更/沒/矣/並/和/呢/或/所/則/卻/哉/很/後/怎/既/甚/
/皆/相/若/唷/哼/哩/唉/哦/啊/得/都/最/喂/喔/喳/喲/等/著/嗎/嗨/
/嗚/嗡/愈/跟/較/過/嘛/嘎/嘟/嘻/嘿/噓/噗/罷/噹/噯/還/雖/嚕/

Mencius: A Chinese Named Entity Recognizer Using Hybrid Model

Tzong-Han Tsai*†, Shih-Hung Wu†, and Wen-Lian Hsu†

*Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
d90013@csie.ntu.edu.tw

†Institute of Information Science, Academia Sinica.
Taipei, Taiwan, R.O.C.
{thtsai, shwu, hsu}@iis.sinica.edu.tw

Abstract

This paper presents a maximum entropy based Chinese named entity recognizer (NER): Mencius. It aims to address Chinese NER problems by combining the advantages of rule-based and machine learning (ML) based NER systems. Rule-based NER systems can explicitly encode human comprehension and can be tuned conveniently, while ML-based systems are robust, portable and inexpensive to develop. Our hybrid system incorporates a rule-based knowledge representation and template-matching tool, InfoMap [1], into a maximum entropy (ME) framework. Named entities are represented in InfoMap as templates, which serve as ME features in Mencius. These features are edited manually and their weights are estimated by the ME framework according to the training data. To avoid the errors caused by word segmentation, we model the NER problem as a character-based tagging problem. In our experiments, Mencius outperforms both pure rule-based and pure ME-based NER systems. The F-Measures of person names (PER), location names (LOC) and organization names (ORG) in the experiment are respectively 92.4%, 73.7% and 75.3%.

1 Introduction

Information Extraction (IE) is the task of extracting information of interest from unconstrained text. IE involves two main tasks: the recognition of named entities, and the recognition of the relationships among these named entities. Named Entity

Recognition (NER) involves the identification of proper names in text and their classification into different types of named entities (e.g., persons, organizations, locations). NER is not only important in IE [3] but also in lexical acquisition for the development of robust NLP systems [4]. Moreover, NER has proven fruitful for tasks such as documents indexing, and maintenance of databases containing identified named entities.

During the last decade, NER has drawn much attention at Message Understanding Conferences (MUC) [5] [6]. Both rule-based and machine learning NER systems have had some success. Previous rule-based approaches have used manually constructed finite state patterns, which match text against a sequence of words. Such system (like University of Edinburgh's LTG [7]) do not need too much training data and can encode expert human knowledge. However, rule-based approaches lack robustness and portability. Each new source of text requires a significant tweaking of the rules to maintain optimal performance; the maintenance costs can be quite steep.

Another popular approach in NER is machine-learning (ML). ML is more attractive in that it is more portable and less expensive to maintain. The representative ML approaches used in NER are HMM (BBN's IdentiFinder in [8, 9] and Maximum Entropy (ME) (New York Univ.'s MEME in [10] [11]). Although ML systems are relatively inexpensive to develop, the outputs of these systems are difficult to interpret. As well, it is difficult to improve the system performance through error analysis. The performance of a ML system can be very poor when training data is insufficient. Furthermore, the performance of ML systems is worse than that of rule-based ones by about 2% as witnessed in MUC-6 [12] and MUC-7 [13]. This might be due to the fact that current ML approaches can capture non-parametric factors less effectively than human experts who handcraft the rules. Nonetheless, ML approaches do provide important statistical information that is unattainable by human experts. Currently, the F-measure in English rule-based and ML NER systems are 85% ~ 94% on MUC-7 data [14]. This is higher than the average performance of Chinese NER systems, which ranges from 79% to 86% [14].

In this paper, we address the problem of Chinese NER. In Chinese sentences, there are no spaces between words, no capital letters to denote proper names or sentence breaks, and, worst of all, no standard definition of “words”. As a result, word boundaries cannot, at times, be discerned without context. As well, the length of a named entity is longer on average than an English one, thus, the complexity of a Chinese NER system is greater.

Previous works [15] [16] [2] on Chinese NER rely on the word segmentation module. However, an error in the word segmentation step could lead to errors in NER results. Therefore, we bypass word segmentation and use a character-based tagger, treat each character as a token, and combine the tagged outcomes of continuing characters to form an NER output.

Borthwick [11] uses an ME framework to integrate many NLP resources, including previous systems such as Proteus, a POS tagger. In this paper, Mencius incorporates a rule-based knowledge representation and template-matching tool, InfoMap [1], into a maximum entropy (ME) framework. Named entities are represented in InfoMap as templates, which serve as ME features in Mencius. These features are edited manually and their weights are estimated by the ME framework according to the training data.

This paper is organized as follows. Section 2 provides the ME-based framework for NER. Section 3 describes features and how to represent them in our knowledge representation system, InfoMap. The data set and experimental results are discussed in Section 4. Section 5 gives our conclusions and possible extensions of the current work.

2. Maximum Entropy-Based NER Framework

For our purpose, we regard each character as a token. Consider a test corpus and a set of *n* named entity categories. Since a named entity can have more than one token, we associate two tags to each category *x*: *x_begin* and *x_continue*. In addition, we use the tag *unknown* to indicate that a token is not part of a named entity. The NER problem can then be rephrased as the problem of assigning one of 2*n* + 1 tags to each token. In Mencius, there are 3 named entity categories and 7 tags: *person_begin*, *person_continue*, *location_begin*, *location_continue*, *organization_begin*, *organization_continue* and *unknown*. For example, the phrase [李 遠 哲 在 高 雄 市] (Lee, Yuan Tseh in Kaohsiung City) could be tagged as [*person_begin*, *person_continue*, *person_continue*, *unknown*, *location_begin*, *location_continue*, *location_continue*].

2.1 Maximum Entropy

ME is a flexible statistical model which assigns an *outcome* for each token based on its *history* and *features*. Outcome space is comprised of the seven Mencius tags for an ME formulation of NER. ME computes the probability $p(o|h)$ for any o from the space of all possible outcomes O, and for every h from the space of all possible histories H. A *history* is all the conditioning data that enables one to assign probabilities to the space of outcomes. In NER, *history* could be viewed as all information derivable from the test corpus relative to the current token.

The computation of $p(o|h)$ in ME depends on a set of binary-valued *features*, which are helpful in making a prediction about the outcome. For instance, one of our features is: when the current character is a known surname, it is likely to be the leading character of a person name. More formally, we can represent this feature as

$$f(h,o) = \begin{cases} 1: \text{if Current - Char - Surname(h) = true and } o = \mathit{person\_begin} \\ 0: \text{else} \end{cases} \quad (1)$$

Here, *Current-Char-Surname(h)* is a binary function that returns the value *true* if the *current character* of the history h is in the surname list.

Given a set of features and a training corpus, the ME estimation process produces a model in which every feature f_i has a weight α_i. This allows us to compute the conditional probability as follows [17].

$$p(o \mid h) = \frac{1}{Z(h)} \prod_i \alpha_i^{f_i(h,o)} \quad (2)$$

Intuitively, the probability is the multiplication of weights of active features (i.e. those f_i $(h,o) = 1$). The weight α_i is estimated by a procedure called Generalized Iterative Scaling (GIS) [18]. This is an iterative method that improves the estimation of the weights at each iteration. The ME estimation technique guarantees that for every feature f_i, the expected value of α_i equals the empirical expectation of α_i in the training corpus.

As Borthwick [11] remarked, ME allows the modeler to concentrate on finding the features that characterize the problem while letting the ME estimation routine deal with assigning relative weights to the features.

2.2 Decoding

After having trained an ME model and assigned the proper weight α_i to each feature f_i, decoding (i.e. *marking up*) a new piece of text becomes a simple task. First, Mencius tokenizes the text and preprocesses the testing sentence. Then for each token we check which features are active and combine the α_i of the active features according to equation 2. Finally, a Viterbi search is run to find the highest probability path through the lattice of conditional probabilities that does not produce any invalid tag sequences (for instance the sequence [*person_begin*, *location_continue*] is invalid). Further details on the Viterbi search can be found in [19].

3 Features

We divide features that can be used to recognize named entities into four categories according to whether they are external and whether they are category dependent. McDonald defined internal and external features in [20]. The internal evidence is found within the entity, while the external evidence is gathered from its context. We use category-independent features to distinguish named entities from non-named entities (e.g., first-character-of-a-sentence, capital-letter, out-of-vocabulary), and category-dependent features to distinguish between different named entity categories (for example, surname and given name lists are used for recognizing person names). However, to simplify our design, we only use internal features that are category-dependent in this paper.

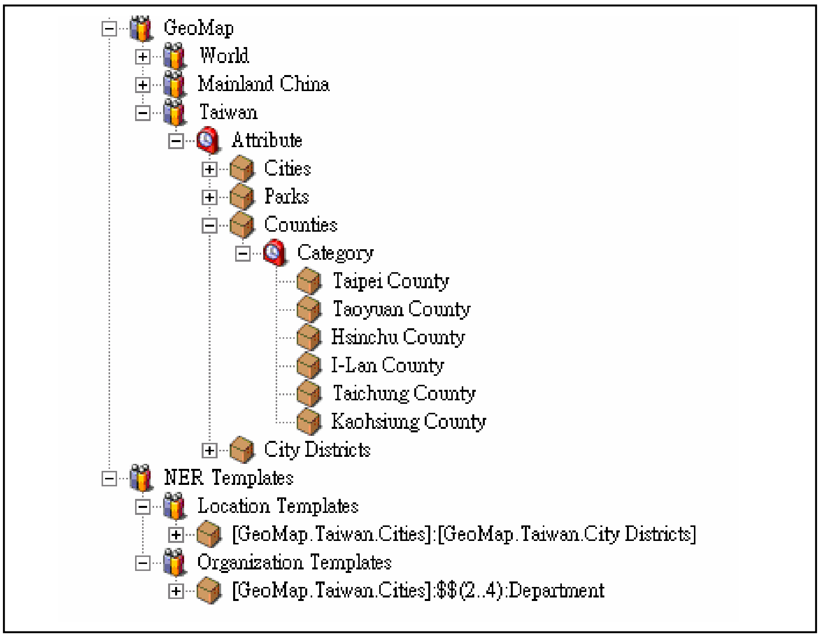
3.1 InfoMap – Our Knowledge Representation System

To calculate values of location features and organization features, Mencius uses InfoMap. InfoMap is our knowledge representation and template matching tool, which represents location or organization names as templates. An input string (sentence) is first matched to one or more location or organization templates by InfoMap and then passed to Mencius, there it is assigned feature values which further distinguish which named entity category it falls into.

3.1.1 Knowledge Representation Scheme in InfoMap

InfoMap is a hierarchical knowledge representation scheme, consisting of several domains, each with a tree-like taxonomy. The basic units of information in InfoMap are called generic nodes which represent concepts, and function nodes which represent the relationships among generic nodes of one specific domain. In addition, generic nodes can also contain cross references to other nodes to avoid needless repetition.

In Mencius, we apply the geographical taxonomy of InfoMap called GeoMap. Our location and organization templates refer to generic nodes in Geomap. In Figure 1, GeoMap has three sub-domains: World, Mainland China, and Taiwan. Under the sub-domain Taiwan, there are four attributes: Cities, Parks, Counties and City Districts. Moreover, these attributes can be further divided, for example, Counties separates into individual counties: Taipei County, Taoyuan County, etc. In InfoMap, we refer to generic nodes (or concept node) by paths. A path of generic nodes consists of all node names from the root of the domain to the specific generic node, in which function nodes are omitted. The node names are separated by periods. For example, the path for the "Taipei County" node is "GeoMap.Counties.Taipei County."



- GeoMap
  - World
  - Mainland China
  - Taiwan
    - Attribute
      - Cities
      - Parks
      - Counties
        - Category
          - Taipei County
          - Taoyuan County
          - Hsinchu County
          - I-Lan County
          - Taichung County
          - Kaohsiung County
      - City Districts
- NER Templates
  - Location Templates
    - [GeoMap.Taiwan.Cities]:[GeoMap.Taiwan.City Districts]
  - Organization Templates
    - [GeoMap.Taiwan.Cities]:$$(2..4):Department

Figure 1. A partial view of GeoMap

3.1.2 InfoMap Templates

In InfoMap, text templates are stored in generic nodes. Templates can consist of character strings, wildcards (see $$ in Table 1), and references to other generic nodes in InfoMap. For example, the template, [通用地理.台灣.縣]:$$(2..4):局 ([GeoMap.Taiwan.Counties]:$$(2..4):Department), can be used to recognize county level governmental departments in Taiwan. The syntax used in InfoMap templates are shown in Table 1. The first part of our sample template above (enclosed by "[]") is a path that refers to the generic node "Counties". The second element is a wildcard ($$) which must be 2 to 4 characters in length. The third element is a specified character "局" (Department).

Table 1. InfoMap template syntax

Symbol	Semantics	Example Template	Sample Matching String
:	Concatenate two strings	A:B	AB
$$(m..n)	Wildcards (number of characters can be from m to n; both m and n have to be non-negative integers)	A:$$(1..2):B	ACB, ADDB, ACDB
[p]	A path to a generic node.	[GeoMap.Taiwan.Counties]	Taipei County,

			Taoyuan County, Hsinchu County, etc.

3.2 Category-Dependent Internal Features

Recall that category-dependent features are used to distinguish among different named entity categories.

3.2.1 Features for Recognizing Person Names

Mencius only deals with surname plus first name (usually with two characters), for example, 陳水扁 (Chen Shui-bian). There are various way to express a person in a sentence, such as 陳先生 (Mr. Chen) and老陳 (Old Chen), which have not been incorporated into the current system. Furthermore, we do not target transliterated names, such as 布希 (Bush), since they do not follow Chinese name composition rules. We use a table of frequently occurring names to process our candidate test data. If a character and its context (history) correspond to a feature condition, the value of the current character for that feature will be set to 1. Feature conditions, examples and explanations for each feature are shown in Table 2. In the feature conditions column, c_{-1}, c_0, and c_1 represent the preceding character, the current character, and the following character respectively.

Table 2. Person Features

Feature	Feature Conditions	Example	Explanation
Current-Char-Person-Surname	$c_0c_1c_2$ or c_0c_1 are in the name list	“陳”水扁, “連”戰	Probably the first character of a person name
Current-Char-Person-Given-Name	$c_{-2}c_{-1}c_0$ or $c_{-1}c_0$ or $c_{-1}c_0c_1$ are in the name list	陳“水”扁, 陳水“扁”, 連“戰”	Probably the second or third character of a person name
Current-Char-Surname	c_0 are in the surname list	“陳”, “林”, “李”	Probably a surname
Current-Char-Given-Name	c_0c_1 or $c_{-1}c_0$ are in the given name list	黃“其”聖, 黃其“聖”	Probably part of a popular given name
Current-Char-Freq-Given-Name-Character	Both c_0, c_1 or c_{-1}, c_1 are in the frequent given name character list	羅“方”全, 羅方“全”	Probably a given name character
Current-Char-Speaking-Verb	c_0 or c_0c_1 or $c_{-1}c_0$ are in the list of verbs indicating speech	“說”, “表”示, 表“示”	Probably part of a verb indicating speech (ex: John said he was tired)
Current-Char-Title	c_0 or c_0c_1 or $c_{-1}c_0$ are in the title list	“先”生, 先 “生”	Probably part of a title

Current-Char-Person-Surname: This feature is set to 1 if $c_0c_1c_2$ or c_0c_1 are in the person name database. For example, in the case $c_0c_1c_2$ = 陳水扁, the feature

Current-Char-Person-Surname for 陳 is active since c_0 and its following characters c_1c_2 satisfy the feature condition.

Current-Char-Person-Given-Name: This feature is set to 1 if $c_{-2}c_{-1}c_0$, $c_{-1}c_0$, or $c_{-1}c_0c_1$ are in the person name database.

Current-Char-Surname: This feature is set to 1 if c_0 is in the top 300 popular surname list.

Current-Char-Given-Name: This feature is set to 1 if c_0c_1 or $c_{-1}c_0$ are in the given name database.

Current-Char-Freq-Given-Name-Character: (c_0 and c_1) or (c_{-1} and c_0) are in the frequently given name character list

Current-Char-Speaking-Verb: c_0 or c_0c_1 or $c_{-1}c_0$ are in the speaking verb list. This feature distinguishes a trigram containing a speaking verb such as 陳沖說 (Chen Chong said) from a real person name.

Current-Char-Title: c_0 or c_0c_1 or $c_{-1}c_0$ are in the title list. This feature distinguishes a trigram containing a title such as 陳先生 (Mr. Chen) from a real person name.

3.2.2 Features for Recognizing Location Names

In general, locations are divided into four types: administrative division, public area (park, airport, or port), landmark (road, road section, cross section or address), and landform (mountain, river, sea, or ocean). An administrative division name usually contains one or more than one location names in hierarchical order, such as 安大略省多倫多市 (Toronto, Ontario). A public area name is composed of a Region-Name and a Place-Name. However, the Region-Name is usually omitted in news content if it was previously mentioned. For example, 倫敦海德公園 (Hyde Park, London) contains a Region-Name 倫敦 (London) and a Place-Name 海德公園 (Hyde Park). But "Hyde Park, London" is usually abbreviated as "Hyde Park" within the report. The same rule can be applied to landmark names. A landmark name includes a Region-Name and a Position-Name. In a news article, the Region-Name can be omitted if the Place-Name has been mentioned previously. For example, 溫哥華市羅伯遜街五號 (No. 5, Robson St., Vancouver City), will be stated as 羅伯遜街五號 (No. 5, Robson St.) in the report later.

In Mencius, we build templates to recognize three types of location names. Our administrative division templates contain more than one set of location names in hierarchical order. For example, the template, [通用地理.台灣.市]:[通用地理.台灣.各市行政區] ([GeoMap.Taiwan.Cities]:[GeoMap.Taiwan.City Districts]), is for recognizing all Taiwanese city districts. In addition, public area templates contain one set of location names and a set of Place-Name. For example, [通用地理.台灣.市]:[通

用地理.台灣.公園] ([GeoMap.Taiwan.Cities]:[GeoMap.Taiwan.Parks]) is for recognizing all Taiwanese city parks. Landmark templates are built in the same way. E.g., [通用地理.台灣.市]:$$(2..4):路 ([GeoMap.Taiwan.Cities]:$$(2..4):Road), is for recognizing roads in Taiwan.

For each InfoMap template category x (e.g., location and organization), there are two features associated with it. The first is Current-Char-InfoMap-x-Begin, which is set to 1 for the first character of matched string and set to 0 for the remaining characters. The other is Current-Char-InfoMap-x-Continue, which is set to 1 for all the characters of matched string except for the first character and set to 0 for the first character. The intuition is: using InfoMap to help ME detect which character in the sentence is the first character of a location name and which characters are the remaining characters of a location name. That is, Current-Char-InfoMap-x-Begin is helpful for determining which character is tagged as *x_begin* while Current-Char-InfoMap-x-Continue is helpful for determining which character is tagged as *x_continue* if we build InfoMap template for that category x. The two features associated with x category are showed below.

$$f(h,o)=\begin{cases}1:\text{if Current - Char - InfoMap - x - Begin} = \text{true and } o = x\_begin \\ 0:\text{else}\end{cases} \quad (3)$$

$$f(h,o)=\begin{cases}1:\text{if Current- Char- InfoMap- x - Continue= true and } o = x\_continue \\ 0:\text{else}\end{cases} \quad (4)$$

In recognizing a location name in a sentence, we test if any location templates match the sentence. If several matched templates overlap, we select the longest matched one. As we mentioned above, the feature Current-Character-InfoMap-Location-Begin of the first character of the matched string is set to 1 while the feature Current-Character-InfoMap-Location-Continue of the remaining characters of the matched string is set to 1. Table 3 shows the necessary conditions for each organization feature and gives examples of matched data.

Table 3. Location Features

Feature	Feature Conditions	Example	Explanations
Current-Char-InfoMap-Location-Begin	c_0~c_{n-1} matches an InfoMap location template, where the character length of the template is n	“台”北縣板橋市	Probably the leading character of a location
Current-Char-InfoMap-Location-Continue	c_a…c_0…c_b matches an InfoMap location	台”北”縣板橋市	Probably the continuing

	template where a is a negative integer and b is a non-negative integer		character of a location

3.2.3 Features for Recognizing Organization Names

Organizations include named corporate, governmental, or other organizational entity. The difficulty of recognizing an organization name is that an organization name is usually led by location names, such as 台北市地檢署 (Taipei District Public Prosecutors Office). Therefore, traditional machine learning NER systems only identify the location part rather than the full organization name. For example, the system only extracts 台北市 (Taipei City) from 台北市 SOGO 百貨週年慶 (Taipei SOGO Department Store Anniversary) rather than 台北市 SOGO 百貨 (Taipei SOGO Department Store). According to our analysis of the structure of Chinese organization names, we found that organization names are mostly ended with a specific keyword or led by a location name. Therefore, we use those keywords and location names as the boundary markers of organization names. Based on our observation, we categorize organization names into four types by boundary markers:

Type I: With left and right boundary markers:

The organization name in this category is led by one or more than one geographical names and ended by an organization keyword. For example, 台北市 (Taipei City) is the left boundary marker of 台北市捷運公司 (Taipei City Rapid Transit Corporation) while an organization keyword, 公司 (Corporation), is the right boundary marker.

Type II: With left boundary markers:

The organization name in this category is led by one or more than one geographical names but the organization keyword (e.g., 公司 (Corporation)) is omitted. For example, 台灣捷安特 (Giant Taiwan) only contains the left boundary 台灣 (Taiwan).

Type III: With right boundary marker:

The organization name in this category is ended by an organization keyword. For example, 捷安特公司 (Giant Corporation) only contains the right boundary 公司 (Corporation).

Type IV: No boundary marker:

In this category, both left and right boundaries as above mentioned are omitted, such as 捷安特 (Giant). The organization names in this category are usually in the abbreviated form.

In Mencius, we build templates for recognizing Type I organization names. Each organization template begins with a location name in GeoMap and ends with an organization keyword. For example, we build [通用地理.台灣.市]:$$(2..4):局 ([GeoMap.Taiwan.Cities]:$$(2..4):Department) for recognizing county level government departments in Taiwan. However, in Type II, III, IV, organization names cannot be recognized by templates. Therefore, the maximum entropy model uses features of characters (from c_{-2} to $c_{2)}$, tags (from t_{-2} to t_2), and organization keywords, e.g., 公司 (Corporation), to find the most likely tag sequences and recognize them.

Once a string matches an organization template, the feature Current-Character-InfoMap-Organization-Start of the first character is set to 1. In addition, the feature Current-Character-InfoMap-Organization-Continue of the remaining characters is set to 1. The necessary conditions for each organization feature and examples of matched data are shown in Table 4. These features are helpful in recognizing organization names.

Table 4. Organization Features

Feature	Feature Conditions	Example	Explanations
Current-Char-InfoMap-Organization-Begin	c_0~c_{n-1} is matches an InfoMap organization template, where the character length of the template is n	“台”北市捷運公司	Probably the leading character of an organization
Current-Char-InfoMap-Organization-Continue	$c_a…c_0…c_b$ matches an InfoMap organization template, where a is a negative integer and b is a non-negative integer	台”北”市捷運公司	Probably the leading character of an organization
Current-Char-Organization-Keyword	c_0 or c_0c_1 or $c_{-1}c_0$ are in the organization keyword list	“公”司, 公“司”	Probably part of an organization keyword

4 Experiments

4.1 Data Sets

For Chinese NER, the most famous corpus is MET-2 [6]. There are two main differences between our corpus and MET-2: the number of domains and the amount of data. First, MET-2 contains only one domain (Accident) while our corpus, which is collected from the online United Daily News in December 2002 (http://www.udn.com.tw), contains six domains: Local News, Social Affairs, Investment, Politics, Headline news and Business, which provides more varieties of organization names than single domain corpus does. The full location names and organization names are comparatively longer in length and our corpus contains more location names under county level and addresses. Therefore, the patterns of location names and organization names are more complex in our corpus.

Secondly, our corpus is much larger than MET2. MET2 contains 174 Chinese PER, 750 LOC, and 377 ORG while our corpus contains 1,242 Chinese PER, 954 LOC, and 1,147 ORG in 10,000 sentences (about 126,872 Chinese characters). The statistics of our data is shown in Table 5.

Table 5. Statistics of Data Set

Domain	Number of Named Entities			Size (in characters)
	PER	LOC	ORG	
Local News	84	139	97	11835
Social Affairs	310	287	354	37719
Investment	20	63	33	14397
Politics	419	209	233	17168
Headline News	267	70	243	19938
Business	142	186	187	25815
Total	1242	954	1147	126872

4.2 Experimental Results

To demonstrate that Mencius performs better than pure rule-based and ML systems, we conduct the following three experiments. We use a 4-fold cross validation to test our system.

4.2.1 Name Lists and Templates (Rule-based)

In this experiment, we use a person name list and InfoMap templates to recognize all named entities. The number of lexicons in person name lists and gazetteers is 32000. As shown in Table 6, the results indicate the F-Measures of PER, LOC and ORG are 83.6%, 71.2% and 76.8%, respectively.

Table 6. Performance of Name Lists and Templates

NE	P(%)	R(%)	F(%)
PER	72.98	**97.93**	83.63
LOC	67.96	**74.67**	71.16
ORG	**95.77**	**64.07**	**76.78**
Total	75.62	**82.13**	78.74

4.2.2 Pure Maximum Entropy Model (ML-based)

In this experiment, we apply the pure ME model, which only uses context information of characters from c_{-2} to c_2 and tags from t_{-2} to t_2. As shown in Table 7, the results indicate that the F-Measures of PER, LOC and ORG are 32.1%, 29.3% and 2.2%, respectively.

Table 7. Performance of Pure Maximum Entropy

NE	P(%)	R(%)	F(%)
PER	62.38	21.64	32.13
LOC	72.83	18.31	29.26
ORG	38.24	1.15	2.23
Total	65.03	13.89	22.89

4.2.3 Integrating Name Lists and Templates into A Maximum Entropy-Based Framework (Hybrid)

In this experiment, we integrate name lists, location templates, and organization templates into a maximum-Entropy-Based framework. As shown in Table 8, the results indicate that the performance of PER, LOC, ORG is better than those in 4.2.1 and 4.2.2.

Table 8. Hybrid Performance

NE	P(%)	R(%)	F(%)
PER	**97.94**	87.39	**92.36**
LOC	**78.60**	69.35	**73.69**
ORG	94.39	62.57	75.25
Total	**90.56**	73.70	**81.26**

4.3 Discussions

In this section, we discuss problems encountered by Mencius.

4.3.1 Data Sparseness

As shown in Tables 6, 7 and 8, Mencius outperforms the rule-based method (Lists and Templates) and ML-based method (pure ME) in the total F-Measure. However, rule-based approach outperforms Mencius in the ORG category. It is due to the data sparseness problem. For example, 中壢天晟醫院 is tagged as [*organization_begin*, *organization_continue*, *unknown*, *unknown*, *organization_continue*, *organization_continue*]. Because 中壢天晟醫院 rarely occurs, it might not appear as an organization name in training set during the 4-fold cross validation experiment. The Viterbi search cannot deal with sequences containing *unknown* tags. With an appropriate post-processing procedure, this kind of error can be resolved. We can treat the *unknown* tag as *x_continue* in a certain window size.

4.3.2 Other Errors

In this section, we show error cases associated with each named entity category.

A. Person Names

The summary report in Table 8 shows that the precision and recall rates for person names are 97.9% and 87.4%, respectively. The major errors are listed below.

(1) The surname character of a person name is not in surname list or the given-name character is not in the given-name character list. Therefore, some of the person features are not set to 1. For example, 李哖 (Lee Nian) are not recognized because 哖 (Nian) is not in the given-name character list.

(2) A person name follows a single-character word which can be a surname. For example, 戴 is both a surname (Dai) and a verb (wear) in Chinese lexicon. However, in 頭戴李應元的帽子 (wear Lee Ying Yuan's Hat), 戴 means *wear* while Mencius mistakenly considers 戴 as a surname. Therefore, Mencius mistakenly recognizes 戴李應 (Dai Lee-Ying) as a person name rather than the correct person name 李應元 (Lee, Ying-Yuan).

(3) Several person names appear consecutively while all of their given names are omitted. Since the context of two person names and one person name are similar, Mencius may mistakenly extract an incorrect name. For example, in the sentence 吳、黃二人在他就職前兩天 , Mencius extracts 黃二人 from it. However, 二人 in English means "both", not the given name.

(4) Transliterated names are not defined in the person name category in Mencius. However, some transliterated person names look like Chinese person names. Therefore, Mencius mistakenly extracts 柯林頓 (Clinton), 夏馨 (Shaheen) from sentences.

(5) Some Japanese and Korean person names look like Chinese person names. For example, Mencius mistakenly extracts 盧武鉉 (Roh, Moo Hyun) from sentences.

B. Location Names

The summary report in Table 8 shows the precision and recall rates for location names are 78.6% and 69.4%, respectively. The major errors are listed below.

(1) Location names within an organization name are extracted but the organization name is not recognized. For example, 韓國東洋製果 (Korea Orion Food) is not recognized as an organization name, but 韓國 (Korea) is recognized as a location name.

(2) The location name is abbreviated. For example, 台 (Tai), the abbreviated form of 台灣 (Taiwan), is not recognized in some cases.

(3) The Chinese usually call a market *street*. For example, 電子街 (Electronics St.) represents an electronics market. However, this is an informal name.

C. Organization Names

Table 8 shows the precision and recall rate for organization name recognition are 94.4% and 62.6%, respectively. We illustrate standard error analysis with examples.

(1) The organization name is a bilingual term. For example, eBay 台灣 (eBay Taiwan) is not recognized.
(2) The organization name is in Type II, III, or IV category (defined in Section 3.2.3). For example, 韓國東洋製果 (Korea Orion Food), 東洋製果公司 (Korea Orion Food Corporation), and 東洋製果 (Orion Food).
(3) Several organization names appear consecutively while part of each name is omitted. For example, in 台北市龍安, 信義, 吳興等國小 (Taipei Long-Ann, Hsin-Yi, and Wu-Xin elementary schools) , 龍安 (Long-Ann), 信義 (Xin-Yi) and 吳興 (Wu-Xin) are not recognized as organizations because the organization ending boundary markers are abbreviated.
(4) The organization name is a foreign organization name, which is not considered by our organization template. For example, 日本農林中央金庫 (The Norinchukin Bank) is not recognized as an organization name.
(5) The organization name is an exception. In 台北縣第二所國中 (the second junior high school in Taipei county), 第二所 means "the second", and appear in the wildcard part of template [通用地理.台灣.縣]:$$(2..13):國中 ([GeoMap.Taiwan.Counties]: $$(2..13):Junior-High-School). We need more out of vocabulary (OOV) knowledge to represent all the number plus quantifier patterns.

5 Conclusions

In this paper, we developed a Chinese NER system, Mencius, which does not rely on the word segmentation module. Instead, we model the NER problem as a character-based tagging problem. Mencius uses ME modeling combining advantages of rule-based and ML-based NER systems. Our hybrid system uses a rule-based knowledge representation system, InfoMap, and incorporates it into the ME framework. The F-Measures of person names (PER), location names (LOC) and organization names (ORG) in the experiment are respectively 92.4%, 73.7% and 75.3%. These are comparatively better than the results obtained by pure rule-based and pure ME-based method.

We are persuaded Mencius can be improved in the following directions. We only use internal features that are category-dependent in this version. In the future, we will collect more features, especially external ones. In addition, we will design a post-processing module to deal with the data sparseness problem. Moreover, we will use document level context information to recognize abbreviated names which cannot be recognized at present.

References

[1] S. H. Wu, M. Y. Day, T. H. Tsai, and W. L. Hsu, "FAQ-centered Organizational Memory," in *Knowledge Management and Organizational Memories*, R. Dieng-Kuntz, Ed. Boston: Kluwer Academic Publishers, 2002.

[2] J. Sun, J. F. Gao, L. Zhang, M. Zhou, and C. N. Huang, "Chinese Named Entity Identification Using Class-based Language Model," presented at the 19th International Conference on Computational Linguistics,, 2002.

[3] R. Grishman, "Information Extraction: Techniques and Challenges," in *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, J. G. Carbonell, Ed. Frascati, Italy: Springer, 1997, pp. 10-26.

[4] S. Coates-Stephens, "The Analysis and Acquisition of Proper Names for Robust Text Understanding," in *Dept. of Computer Science*. London: City University, 1992.

[5] N. Chinchor, "MUC-6 Named Entity Task Definition (Version 2.1)," presented at the 6th Message Understanding Conference, Columbia, Maryland, 1995.

[6] N. Chinchor, "MUC-7 Named Entity Task Definition (Version 3.5)," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[7] A. Mikheev, C. Grover, and M. Moensk, "Description of the LTG System Used for MUC-7," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[8] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, and R. Weischedel, "BBN: Description of the SIFT System as Used for MUC-7," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[9] D. Bikel, R. Schwartz, and R. Weischedel, "An Algorithm that Learns What's in a Name," *Machine Learning*, 1999.

[10] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "NYU: Description of the MENE Named Entity System as Used in MUC-7," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[11] A. Borthwick, "A Maximum Entropy Approach to Named Entity Recognition," New York University, 1999.

[12] N. Chinchor, "Statistical Significance of MUC-6 Results," presented at the 6th Message Understanding Conference, Columbia, Maryland, 1995.

[13] N. Chinchor, "Statistical Significance of MUC-7 Results," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[14] N. Chinchor, "MUC-7 Test Score Reports for all Participants and all Tasks," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[15] H. H. Chen, Y. W. Ding, S. C. Tsai, and G. W. Bian, "Description of the NTU System Used for MET2," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[16] S. H. Yu, S. H. Bai, and P. Wu, "Description of the Kent Ridge Digital Labs System Used for MUC-7," presented at the 7th Message Understanding Conference, Fairfax, Virginia, 1998.

[17] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, pp. 39-71, 1996.

[18] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Annals of Mathematicl Statistics*, vol. 43, pp. 1470-1480, 1972.

[19] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. IT, pp. 260-269, 1967.

[20] D. McDonald, "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names," in *Corpus Processing for Lexical Acquisition*, J. Pustejovsky, Ed. Cambridge, MA: MIT Press, 1996, pp. 21-39.

以網際網路內容為基礎之問答系統“Why”問句研究

沈天佐　林川傑　陳信希

國立台灣大學資訊工程學系

{tzshen,cjlin}@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

摘要

以“Why”開頭的問句，問題的答案是“原因”。“原因”有不同的型態，可能是一個片語、一個子句、一個句子，甚至跨越句子的範圍。目前的問答系統特別針對“Why 問句”研究的並不多，本文探討如何從文件中擷取出“Why 問句”的答案，文件的來源設定在網際網路。我們運用搜尋引擎取得相關文件，以描述因果關係的句型來擷取答案。由於句型本身可能會有歧義性，某個句型的出現並不代表一定是問句的答案，本文也針對這項議題進一步分析。我們並將所發展的問答系統，與另外兩個以網際網路為基礎的問答系統—AnswerBus 和 LCC，作了效能的評估。在以 50 個問句的測試中，我們的系統、AnswerBus 和 LCC 的 MRR 值分別為 0.623、0.429 和 0.229，顯示我們的系統的效能優於這兩個系統。

1. 緒論

問答系統接受使用者的自然語言問句，從一堆文件集中，找出問句的答案。透過問答系統，使用者可以直接得到答案，而不必自己瀏覽資訊檢索系統所傳回的一堆相關文件尋找答案。TREC (Text Retrieval Conference) 自 1999 年開始舉辦問答系統的效能評比 (Voorhees, 1999)，帶動近年來問答系統的研究風潮。TREC 評比的重點隨著研究成果的進展，每年都進行調整。以 2002 年為例，評比的重點在於參賽者的系統是否能夠準確地定出答案的範圍，而不是以一個固定長度的文字片段當作答案。

完整的問答系統分為兩步驟，第一個步驟是從所有文件中找出與問句相關的

文件，此即「資訊檢索」的部分。如何將自然語言問句轉換為適合資訊檢索系統的查詢字串，是個研究課題。第二個步驟是從相關文件中找出問句的答案，此稱為「答案擷取」，這個部分是問答系統主要研究重點。進行「答案擷取」，問答系統必須針對問句進行分析，以取得答案的類型。常見的「答案擷取」方法是利用 “Named Entity Tagging” 的技術，再加上 “問句與上下文相似度的計算”。從簡單的關鍵字比對，到較複雜的語意一致性判斷，都是可能的上下文與問句相似度計算方法 (Harabagiu *et al*., 2000a; Moldovan and Rus, 2001)。

以網際網路為基礎的問答系統研究，主要是利用網路上常見的搜尋引擎進行資訊檢索，以取得相關文件，再利用與 TREC 問答系統類似的技術來擷取答案。這種類型的問答系統，必須考量即時性，避免太複雜技術帶來的負擔。目前的研究有 Radev *et al*. (2001)、Radev *et al*. (2002)、Zheng (2002)、Lin (2002)。另外，網頁文件的一些特性，例如 HTML 標記、超鏈結、風格差異、內容正確性等，也是在研究上必須考量的議題。

目前大部分問答系統擷取答案方法，主要針對答案類型為 Named Entities。對於答案較複雜，沒有固定形式的問句類型，如 “Why … ?” 和 “How does *S V*?”，則較少有深入的探討與分析。Girju 與 Moldovan (2002) 曾經探討過回答 “cause-effect questions”，研究因果關係在文中的表達方法。不過這篇文章的重點擺在 <NP1 VERB NP2> 這種 pattern 上，其中的動詞必須是個 “causative verb”，例如：“cause”、“lead to”、“make”等。由於這些動詞未必一定代表因果關係，如 “make” 有時的意義為 “製造”，所以研究重點在於如何由 VERB、NP1 和 NP2 來判斷是否描述因果關係。

在閱讀測驗問答系統 (reading comprehension) 的研究上, Anand *et al*. (2000) 和 Riloff and Thelen (2000) 也有相關研究。系統針對一篇文章，找到問句的答案。TREC 問答系統與這類問答系統主要的不同點是答案來源為多篇相關文件，答案可能重複出現多次，有較多機會找到答案，但雜訊也會比較多。閱讀測驗問答系統則相反，答案可能只出現在文章中一次，所以需要較複雜的方法來找到不

是那麼明顯的答案，但另一方面雜訊會比較少。

第 2 節說明實作系統的架構，以及各個子系統。第 3 節引用 Penn Treebank 語料庫，分析擷取答案 patterns 的準確率。第 4 節為本系統的效能評估，並與另外兩個以網際網路為基礎的問答系統比較。第 5 節是結論與未來研究方向。

2. 系統概觀

2.1 資訊檢索系統

本文所提的問答系統架構如圖 1，只針對單一的問句類型 (也就是以“why”開頭的問句) 進行處理，所以並未包含問句分析子系統，同時我們選擇 Google 來找出與問句相關的網頁文件。首先將問句轉為查詢字串，去掉問句中的停用詞 (stop words，包括疑問詞、介系詞、連接詞、代名詞、助動詞、某些副詞……等) 與標點符號，剩下來的字以空白相連接，為交給 Google 的原始查詢字串。由於 Google 採 AND 的方式來解讀關鍵字，一定要含所有關鍵字的文件才會被取回，所以有可能取回的文件篇數很少。

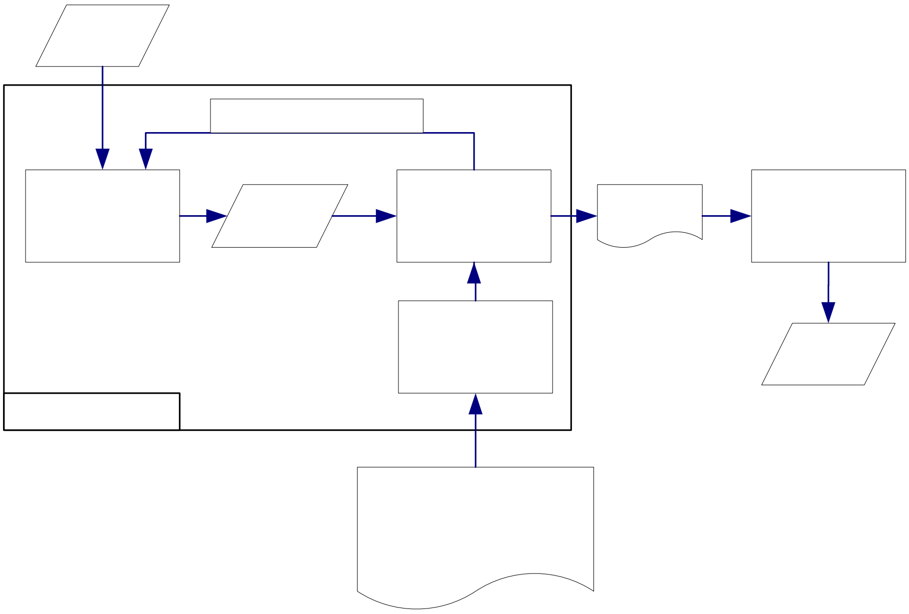


圖 1. 問答系統架構

當 Google 所找回的相關文件數量不足時，我們會修改查詢字串，再進行一次查詢以補足不足的部分。查詢字串修改的方法是刪除查詢字串中的某個關鍵字，產生新的查詢字串。我們選擇“權重”較小的先刪除，設定權重如下：

> 專有名詞 > 名詞詞組的 Head > 動詞詞組的 Head > 名詞詞組的其他字 > 動詞詞組的其他字 > 不在名詞詞組或動詞詞組的其他字

權重越大的關鍵字，與文件主題的關係越密切。

當文件中含有某些特殊字如“reason”時，可能表示此文件中提到某個事件的原因。因此，若在進行資訊檢索時，能夠將這些特殊字加到原有的查詢字串中，所檢索到的排名較前面的網頁文件，就會是那些既含有問句中的關鍵字 (表示和問句所問的主題相關)，而且內容又描述了某種因果關係的網頁文件。同時 Google 在檢索文件時，也考慮了各關鍵字在文中的接近程度。當各關鍵字在文件中越接近，文件的排名會越前面。可以協助尋找因果關係的特殊字，包括“reason”、“why”和“because”等。

2.2 答案擷取系統

要在文件中找尋表達因果關係的資訊，目前已知有四種情形：

一、 利用因果 patterns 來判斷文件中描述因果關係的部分。

二、 以整篇文章來解釋原因和理由。網際網路上較常看到這種情形，作者在問句處提供一個指向答案的超連結。

三、 原因和結果出現在前後文，兩者間並無明顯關連詞出現。

四、 某些動詞隱含因果關係，如 Girju and Moldovan (2002) 所做的研究，以及在 WordNet 中也有動詞間 causation 關係的資訊。

以下針對各情形詳細說明：

一、 利用因果 patterns 來判斷文件中描述因果關係的部分

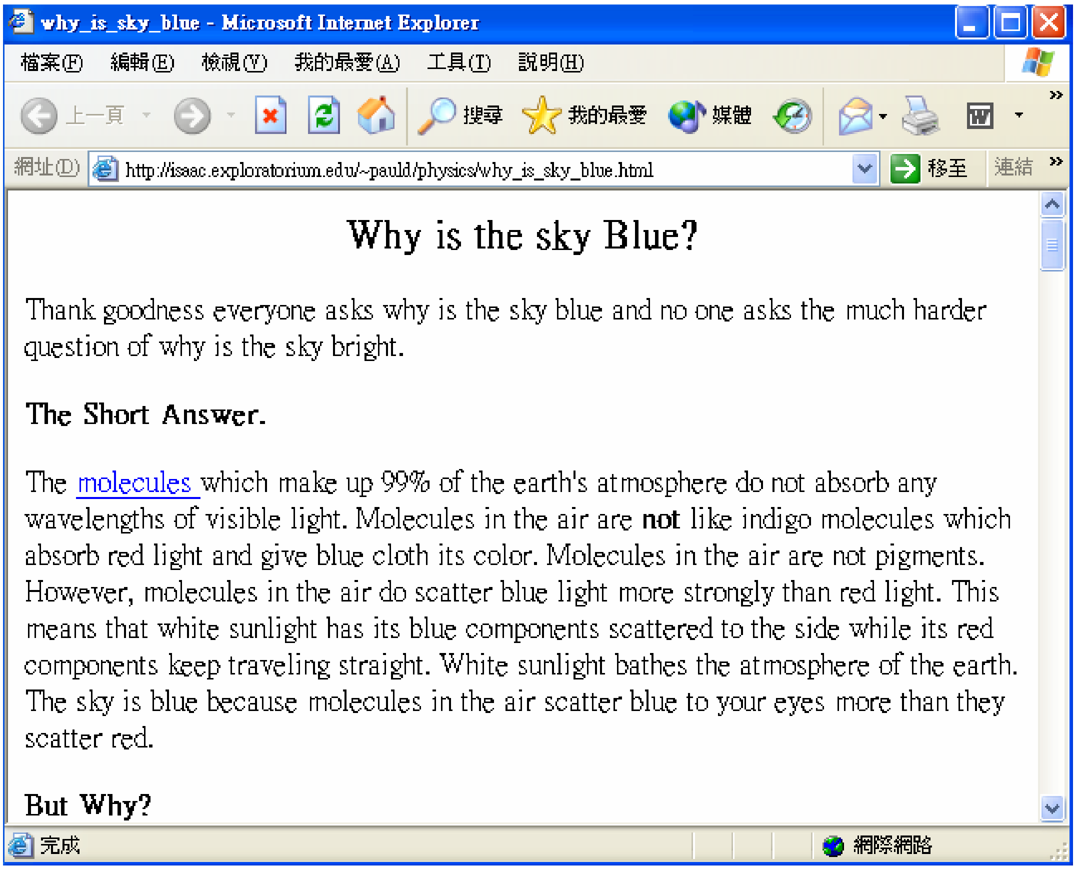
在文法及修辭學上，有不少句型可用來描述兩件事之間的因果關係。我們試著利用這樣的句型來找出“原因”的部份。這樣的句型包括：

[EVENT] because [REASON].
[REASON], therefore [EVENT].
[EVENT] in order to [REASON].

其中 [EVENT] 代表結果事件，[REASON] 表示其發生原因。這些句型所得到的 patterns 不但可以用來判斷因果關係的資訊，也可以用來決定“原因”部份的邊界。

二、 以整篇文章來解釋原因和理由

在某些以教育為主題，或是提供常見問答集 (FAQ) 的網站中，就可看到這類以整個段落或整篇文章來解釋或回答一個問題的網頁。例如圖 2 即為“Why is the sky blue?” 答案的網頁，其中答案的描述長達一整篇文章。



網址(D) http://isaac.exploratorium.edu/~pauld/physics/why_is_sky_blue.html

Why is the sky Blue?

Thank goodness everyone asks why is the sky blue and no one asks the much harder question of why is the sky bright.

The Short Answer.

The molecules which make up 99% of the earth's atmosphere do not absorb any wavelengths of visible light. Molecules in the air are **not** like indigo molecules which absorb red light and give blue cloth its color. Molecules in the air are not pigments. However, molecules in the air do scatter blue light more strongly than red light. This means that white sunlight has its blue components scattered to the side while its red components keep traveling straight. White sunlight bathes the atmosphere of the earth. The sky is blue because molecules in the air scatter blue to your eyes more than they scatter red.

But Why?

圖 2. 常見問題集答案網頁的範例

因為是人工建置的，這種情形所取得的答案無疑會是正確答案。我們所需要做的工作是，尋找此類的網頁並找到何者是它的原始問句。如果原始問句與使用者所問的問題是一樣的，則以此段文字 (或網頁連結) 提供給使用者做為答案。這方面的研究比較接近 FAQ Finding。

三、 原因和結果出現在前後文，兩者間並無明顯關連詞出現

在下面這個例子中：

問： Why can't ostriches fly?
答： The flightless birds … include ostriches, … . These birds have only small or rudimentary wings.

答句中的 “These birds have only small or rudimentary wings.” 就與前一話沒有直接的因果關連詞，但是人類仍可以知道這句和前句有著因果關係。

四、 動詞隱含因果關係

在 Girju 與 Moldovan 的研究中，某些動詞帶有因果次序的訊息，例如 “provoke”、“induce”。WordNet 中則提供了動詞之間 “cause to” 的關係，例如『“kill” cause to “die”』。然而這樣的動詞並不多，並不是描述因果關係的最主要方式。

由以上的說明，我們可以發現，方法一不但較為簡單，適用性又廣，非常適合於網際網路環境的問答系統建構。因此我們將重點放在因果 patterns 的建置，以及比對出文件敘述因果關係部分的方法。另外，為了處理第三種情形，我們設計了一個 pattern 稱為 “final pattern”，將在第 2.2.1 節中介紹。動詞隱含因果關係則可做為未來加強系統回答能力的有用資訊。

2.2.1 因果關係 patterns

在文法及修辭學上，有不少句型可用來描述兩件事之間的因果關係。我們試著利用這樣的句型來找出 “原因” 的部份。這樣的句型包括：

[EVENT] because [REASON].
[REASON], therefore [EVENT].
[EVENT] in order to [REASON].

其中 [EVENT] 代表結果事件，[REASON] 表示其發生原因。這些句型所得到的 patterns 不但可以用來判斷因果關係的資訊，也可以用來決定“原因”部份的邊界。

我們從一些相關書籍及研究成果中蒐集到許多表示因果關係的句型。當一個“why”問句被提出後，我們的問答系統會先利用上述 patterns 找出所有包含因果關係的句子，並且評估 patterns 中對應 [EVENT] 的部份與問句的相似度。當有個文句符合其中某一條因果 pattern，且對應 [EVENT] 部份高度相似時，系統就可以抽取出文句中對應 [REASON] 部份，做為回應給使用者的答案。

然而在實際實驗時，我們發現有些應用上的問題。首先，有的 patterns 是描述句子之間 (而不是句子之內) 的關係的。舉例來說：

> Molecules in the air scatter blue to your eyes more than they scatter red. Therefore, the sky is blue.

上面段落上，“the sky is blue” 的原因位於前一個句子。因此，這些 patterns 被修正為以兩句或三句話做為比對單位來決定 [EVENT] 及 [REASON] 的位置。像在上面的例子中，pattern 即為 “[REASON]. Therefore, [EVENT].”。

此外，有些 patterns 並不僅只代表因果關係，同時也有其他含意的用法，因此會有歧義性存在。舉例來說，“since” 這個字就有“由於”和“自從”兩種不同意義。在下面兩句話中：

(1) Since their enemies had been destroyed, they sent back their army.
(2) Since that day, the flowers she had planted had spread all over the hill.

第一句在 “since” 後面所接的文字表達了一件事情的原因，而在第二句中 “since” 則是點出某個事件的起始時間，而不是原因。

為了能夠正確地使用 patterns，我們必須更進一步地瞭解各 patterns 應用上的準確性，且做可能的修正。因果 patterns 準確率的預估方法將於第 3 節中敘述。一旦有了準確率的資訊後，尋找答案時就由準確率較高的 pattern 開始比對起，以最先符合的 pattern 來考慮是否可能為正確答案。

有時，在文字的表現上，[EVENT] 和 [REASON] 之間並沒有很強烈的字面訊息。人類是由上下文以及人類具有的知識得知它們的因果關係。唯一的字面線索是 [EVENT] 和 [REASON] 僅出現在前後文。為了也能捕捉到這種情形，我們加了一個 pattern 為 “[REASON]. [EVENT]. [REASON].”，稱之為 “final pattern”。設定其擁有最低的準確率，成為最後一個被比對的 pattern。

2.2.2 答案擷取步驟

以下為因果 patterns 比對的步驟：

一、 先使用一個詞性標記系統對問句進行詞性標記。我們使用的是 QTAG 3.1[1]。

二、 去除掉問句中的 “why”，餘下的字做為比較相似度時的關鍵字，每個字的權重依其詞性而定。名詞、動詞的權重為 5，形容詞、副詞、數詞、符號或公式的權重為 4，連接詞、冠詞、介系詞及不定詞中的 “to” 的權重為 1，其餘詞性的關鍵字權重為 2。

三、 利用 Porter Stemmer 對所有的關鍵字進行字根還原，由字根還原得到的字稱為“字根關鍵字”。這些字的權重為原始關鍵字權重的一半。

四、 當文件中的句子符合某因果 pattern 時，切出文句中對應 [EVENT] 的部分。

五、 計算與問句之間的相似度。針對問句中的每一個關鍵字，如果出現在 [EVENT] 部份，則加此關鍵字的權重於相似度的分數中。若是僅為字根關鍵字，則加上字根關鍵字的權重。若未出現則不加分。[EVENT] 和問句的相似度即為所有關鍵字所貢獻之分數總和。

六、 最後，此句子可能為正確答案的分數為：(所符合 pattern 的權重)×([EVENT] 與問句的相似度)。其中 pattern 的權重定義為：0.5 + (0.5×pattern 準確率)。在此定義 “final pattern” 的準確率為 0。

問答系統依照上面的流程，對每個相關文件中的句子做比對並計算可能成為答案的分數。最後依照步驟六所得分數排序，將分數高的答案回覆給使用者。

[1] http://web.bham.ac.uk/O.Mason/software/tagger/

3. 句型歧義性分析

為了觀察各 patterns 的正確性，需要一個較大規模的測試集來測試。測試集中需包含 patterns 的出現，並標示出“原因”所在的位置。然而目前並沒有這樣的測試集存在。底下我們利用兩種方法來求得各因果 patterns 的準確率。

3.1 Penn Treebank 之 PRP 標記

Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993) 裡有一個功能標記是“PRP”，用來標示該詞組帶有“目的”或是“理由”的角色。例如，底下這句話的剖析樹中：

> Chevron had to shut down a crude-oil pipeline in the Bay area to check for leaks.
>
> ((S (NP-SBJ-1 Chevron) (VP had (S (NP-SBJ *-1) (VP to (VP shut (PRT down) (NP (NP a crude-oil pipeline) (PP-LOC in (NP the Bay area) (S-***PRP*** (NP-SBJ *-1) (VP to (VP check (PP-CLR for (NP leaks))))))) .))

在此句中，“to check for leaks”就是“shut down a crude-oil pipeline”的原因。

Penn Treebank 中有 9,613 個含有 PRP 標記的句子。有些句子含有一以上的 PRP 標記，故 PRP 出現總數為 10,720 次。

然而有些同樣表達因果關係的情形，卻沒有被標上 PRP 標記。Penn Treebank 只標示帶有原因及目的角色的附屬子句或是介系詞片語，若是整個句子代表理由時 (例如“Because he was young.” “Therefore, he will not attend.”)，就不會帶有 PRP 標記。如此一來，不是每個與因果 patterns 相符的句子都會標上 PRP 標記。

因此，我們改利用 PRP 標記這項資訊來估算 patterns 關鍵字的準確率。Patterns 關鍵字就是 patterns 中 [EVENT] 和 [REASON] 之外的部份。之後以 patterns 中關鍵字的準確率來做為 patterns 的準確率。

抽取出 Penn Treebank 中所有被標上 PRP 標記的詞組，統計出現在詞組開頭的因果 patterns 關鍵字個數。再計算每個 patterns 關鍵字出現在整個 Penn Treebank 中的次數，就可得出當一個 pattern 關鍵字出現時，它會被標為 PRP 的比例是多

少。我們以此比例視為 patterns 關鍵字的準確率。統計結果如表 1 所示。

由表 1 的統計資料中，我們可以看到某些 patterns 關鍵字準確率很高，像是 “because”、“in order to” 等等。然而除了包含 “because” 出現次數很高以外，其他 patterns 關鍵字的出現次數並不高。相反的，“to”、“for” 以及 “since” 這幾個在 PRP 詞組中常出現的 patterns 關鍵字，卻因為準確率不夠高而無法直接用來判斷因果關係。這是因為這些 patterns 關鍵字有歧義性的原因，以下我們針對這三個 patterns 關鍵字再做進一步的分析。

表 1. 各因果 patterns 準確率

Patterns 關鍵字	PRP 個數	總次數	準確率
‘cause	9	9	1
because	3750	3861	0.971
because of	641	661	0.97
in order to/for/that	108	116	0.931
so as to	6	7	0.857
as a result of	61	85	0.718
on account of	5	7	0.714
as a result	39	86	0.453
so that	180	416	0.432
so as	4	10	0.4
due to	40	110	0.364
cause	82	249	0.329
since	310	1169	0.265
why	133	824	0.161
to	3318	55272	0.06
for	731	18075	0.044
so	173	8768	0.02
as	46	10481	0.004
that	16	36897	0.0004
for \w+ing	66	823	0.08

(1) “to + 原形動詞”

出現 “to” 而表示因果關係的情形中，通常是先以完整句子描述事件，再接以 “to” 開頭的不定詞子句說明原因，因此 “to” 之後必定接原形動

詞。我們觀察 Penn Treebank 中所有“w_1 to w_2”或是“To w_2”的情形，若是限制 w_2 為原形動詞而 w_1 不為動詞時才判定為因果關係，準確率可由 6%提升至 15.1%。

(2) “for”

以“for”來表達因果關係時，用法如同“because”，是用以連接描述原因和結果的兩個句子的。因此“for”多半出現在句首，或是在句中但以逗號與前句隔開。我們於是將 pattern 改為“For …” (限定在句首) 以及“…, for …” (在句中前接逗號)。然而 Penn Treebank 中並不會將所有表示因果關係的情形都標上 PRP (此點將在第 3.2 節中討論)，因此我們改以人工判斷的方式，隨機取出的 25 句符合本 patterns 的句子觀察，準確率是 7/25=28%。

(3) “since”

為了排除以“since”描述時間起始點的情形，只要“since”之後接有年份、月份，或是“year”、“day”等等之類代表時間的關鍵字，以及“ever since”、“since then”的話，都不視為因果關係。如此一來，準確率提升至 38.4%。如果更進一步限制“since”只能出現在句首或是逗號之後，準確率可達 64%，但是這樣只能判斷出一半以“since”為起首的 PRP 詞組。為了彌補召回率的不足，除了“since”在句首或逗號後並做時間詞判斷的 patterns 外，我們也保留了完全不做判斷的“since” pattern (準確率 26.5%)。

3.2 人工觀察

因為 Penn Treebank 只標示帶有原因及目的角色的附屬子句或是介系詞片語，有些 patterns 關鍵字就不會被標上 PRP，像是“therefore”。要得到這些 patterns 關鍵字的準確率，我們改以人工的方式來進行。我們隨機至 Penn Treebank 中抽出至多 25 句出現 patterns 關鍵字的句子，再以人工判定是否為因果關係。所得到的觀察結果列在表 2。

在表 2 中，“due to” 會加上 “不接原形動詞” 這一項條件，是因為我們觀察發現，當 “due to” 意義為 “預定要” 的時候，其後會接動詞的原形，而這就不是描述因果關係的句子了。

表 2. 由人工觀察之 patterns 準確率 (%)

Patterns	**人工判定表因果關係個數**	**準確率**
therefore	25/25	100
Thus	20/25	80
hence	21/21	100
So … 或 …, so …	15/25	60
due to (不接原形動詞)	25/25	100
as a result	25/25	100

此外，以 “so” 起始的子句會被標上 PRP 者，都是等同於 “so that” 的情形。為了計估 “so” 當連接詞、表示 “所以” 的比例，我們也以人工的方式觀察了 “so” 出現在句首或逗號之後的情況。這樣的 patterns 準確率可由 2% 提升到 60%。準確率仍然不夠高的原因是因為 “so” 的用法實在太廣，無法單以字面就能決定 “so” 的真正角色。

3.3 因果 patterns 的比對

由 Penn Treebank 及人工判斷得到各準確率之後，我們就可以依照 patterns 的準確率排序。要尋找文件中含有因果關係的句子時，會由準確率較高的 pattern 開始比對起。如此一來，如果同一段文字中出現兩個以上符合因果 patterns 的部份，將會優先判斷準確率較高的部份是否為可能答案。

由於有些 patterns 的準確率來自小量測試集的人工評估，我們於是將由 Penn Treebank 評估所得高準確率 patterns 的優先順序往前挪。先比對在 Penn Treebank 中準確率大於 80%的 patterns，再依照其餘 patterns 的準確率由高到低分別比對。

4. 實驗與討論

我們根據第 2 節及第 3 節所得到的 patterns 及其準確率，實作了一個針對 “why

問句”回答的問答系統。本節描述我們如何進行效能評估的實驗，並且也與另外兩個以網際網路為基礎的問答系統 AnswerBus (http://www.answerbus.com/) 及 LCC (Language Computer Company, http://www.languagecomputer.com/) 做比較。

4.1 實驗資料

在 TREC QA-Tracks 歷屆的題目中，只有 8 題是屬於“why 問句”。為了擴大實驗規模，我們至 AskJeeves (http://www.ask.com/) 網站蒐集之前使用者曾經提過的問題。在十萬多個問句中，僅有 87 句是“why 問句”。另外還有 50 題是由 AnswerBus (http://www.answerbus.com/) 網站中“Sample questions from Excite” 網頁內容所整理出來的，總共得到 145 個問句可進行實驗。

受限於人力的不足，我們先以其中的 50 題來進行系統的效能評估。先去掉這 145 題中意義重複的問句，以及某些沒有標準答案或是詢問建議、需結合使用者背景資料才能回答的題目，例如 “Why is my monitor only showing 16 colors ?”、“Why should I go to college?” 等。之後隨機選取 42 題，連同來自 TREC 的 8 題共 50 題“why 問句”來進行評估。

檢索相關文件以備尋找答案時，我們會在由問句所建構成的原始查詢中，分別加入“reason”、“why” 和 “because” 這三個特殊字，成為新的查詢，分別利用 Google 找出最相關的前 10 篇共 30 篇文件備查。之後再利用原始查詢檢索出不與這 30 篇重複的 200 篇相關文件。因果 patterns 比對及答案擷取就在這 230 篇文件中進行。

4.2 答案評估

AnswerBus 和 LCC 在給使用者答案時，都是以句子為單位回覆。為了要和這兩個系統比較，我們的系統也以完整的句子做為給答單位。但如果比對成功的 pattern 會跨過句子邊界，則系統會將所有此 pattern 所涵蓋的句子都抓出來做為一個答案。

測試時，將第 4.1 節選出的 50 題問句分別向這三個系統提出。由每個系統

的回答中，各題都挑出前五名的答案以進行評估。AnswerBus 和 LCC 常常只回覆了 5 個以下的答案，AnswerBus 平均回答 4 個，LCC 平均回答 4.88 個。我們的系統則是一定提供前五名比對到的答案。

標定各答案是否正確是由人工來進行。我們將答案打散，讓評估者無從得知各答案是由哪個系統所回答的。每一題都給三個評估者評估，以多數人的意見決定是否為正確答案。得到的評估結果列在表 3 之中。

表 3. 問答系統回答“why 問句”的效能評估

系統	正解在第一名	正解在前五名	MRR
AnswerBus	15	31	0.429
LCC	8	20	0.229
我們的系統	26	39	0.623

表 3 中第四欄的 MRR (Mean Reciprocal of Rank) 是在 TREC QA 比賽中所用的評比標準 (Voorhees, 1999)。其計算方法為，針對一問句，若系統所給出第一名的答案即為正確答案的話，得一分。若第二名的答案才正確的話，得 1/2 分。若第三名才正確的話，得 1/3。也就是以正解所在的最高名次的倒數為得分，最後的 MRR 值為每一題所得分數的平均。

由表 3 我們可以看到，我們的系統利用因果 patterns 的幫助，系統效能優於其他兩個線上系統。

4.3 分析

4.3.1 增加特殊字查詢相關文件的幫助

如第 2.1 節所提，在檢索相關文件時，系統會在查詢中加入“reason”、“why”和“because”等特殊字，以期找回的相關文件中能含有因果關係的文句。但是這個動作的幫助有多少？

首先我們觀察加不加入特殊字，對於檢索所得相關文件的影響。分別以 145 個問句的原始查詢字串與加入特殊字查詢字串做檢索，原始查詢 (包括刪去查詢字以求足量相關文件的動作) 取前 200 名。特殊字查詢字串所得到的 145×3×10 =

4,350 篇中，有 1,216 篇完全未出現在以原始字串查詢的結果中，顯示特殊字確實可以幫助抓到更多可能含有因果關係描述的文件。

而在系統評估時，我們針對各正確答案的來源文件做了統計，結果在表 4。其中 Rn、Wn、Bn 分別表示加入 “reason”、“why”、“because” 查詢所得的第 n 篇相關文件，Nn 則表示利用原始查詢所得、但排去已由特殊字查詢檢索出文件的第 n 篇相關文件。

表 4. 答案與文件來源的關係

範圍	總數	正解	範圍	總數	正解
R1-R10	28	13	N91-N100	3	1
W1-W10	47	20	N101-N110	8	3
B1-B10	45	25	N111-N120	6	1
N1-N10	14	5	N121-N130	7	2
N11-N20	3	1	N131-N140	8	3
N21-N30	3	2	N141-N150	12	4
N31-N40	8	4	N151-N160	7	2
N41-N50	11	5	N161-N170	8	3
N51-N60	9	5	N171-N180	2	0
N61-N70	6	4	N181-N190	3	2
N71-N80	6	2	N191-N200	1	1
N81-N90	5	2	總計	250	110

針對實驗的 50 個問句，我們的系統提出了 250 個可能答案，有 110 個被評估為正確。由表 4 可知，有一半以上 (13+20+25=58) 的正確答案來自加入特殊字查詢所得的相關文件中。顯示這利用特殊字所查得的 30 篇相關文件，提供因果關係的資訊遠多於原始查詢的相關文件。

此外，有趣的是，是否找到正確答案，與排去特殊字的原始查詢相關文件的名次不很相關，這和問答系統研究中的一個性質相吻合：正確答案不一定出現在所謂「最相關」的文件中。

4.3.2 各因果 patterns 的答題正確率

表 5 中列出了各 patterns 提供答案的個數，以及被評估為正確答案的個數 (以關

鍵字做為分類統計)。由表 5 中可以看到，有許多 patterns 在這次評估中並未被用到，“because” 和 “because of” 則佔了一半以上。這表示 “because” 確實是一個很常用的句型，它們在我們整理出的 patterns 排名中又很前面，所以容易先被比對到。

然而在表 5 中 “because” 各 patterns 答題正確率卻只有 50%上下。經過觀察，發現這並不是 patterns 的錯誤。許多 patterns 比對的錯誤都來自 [EVENT] 部份與問句比對這階段。由於我們的系統在計算相似度時是以關鍵字比對為主，會發生比對錯誤的情形。比方說，問句為 “Why is the sky blue?”，而有一個句子是 “Blue ocean is beautiful because…”，這時問句和 [EVENT] 部份有不小的相似度，造成答案抽取的錯誤。

同樣的情形也會造成正確答案未被找到的狀況。當含有正確答案且符合因果 patterns 的句子中，[EVENT] 部份使用了與問句語意相同但字面差異很大的說法時，這個句子就無法比對成功，答案也就無法被找到。由此可知，短文句間的相似度比對及語意比對是影響問答系統效能的重要因素。

其次的錯誤就來自於 patterns 關鍵字本身的歧義性，如同我們在第 3 節中所討論的一樣。在擷取答案的過程中，仍會找到並非因果關係描述的文句。

“Why [EVENT]? [REASON].” 這個 pattern 有另一個錯誤情形。在文章中，會以 “Why…?” 提詞的敘述方式，其後可能會以三四句甚至一整段文字來解釋這個原因。而我們的步驟至多只抽取往後一個句子，因此會因答案不完整而被判斷錯誤。

表 5. 答案與 patterns 之關係

Pattern 關鍵字	總數	正解	正確率
because of	24	12	50.0%
because	102	56	54.9%
‘cause	0	0	-
In order to	0	0	-
so as to	0	0	-
as a result of	2	0	0.0%
as a result	0	0	-
therefore	8	3	37.5%
hence	3	3	100.0%
due to	8	5	62.5%
thus	2	1	50.0%
on account of	0	0	-
that is why	2	2	100.0%
for this reason	0	0	-
Why ?	31	13	41.9%
reason that	7	1	14.2%
so as	0	0	-
so that	1	1	100.0%
since（經判別）	6	3	50.0%
So/,so	14	3	21.4%
For/,for	2	1	50.0%
to-V（經判別）	3	1	33.3%
since（未判別）	0	0	-
to-V（未判別）	8	1	12.5%
for	11	2	18.2%
so	0	0	-
as	2	1	50.0%
“final pattern”	13	0	0.0%
總和	250	110	44.0%

5. 結論與未來工作

本論文建構了一個以網際網路為基礎的問答系統，自動回答“why”類型的問句。我們使用了搜尋引擎檢索出相關的網頁文件以用來尋找可能答案。接著利用描述因果關係的 patterns，評估 patterns 中 [EVENT] 部份與問句本身的相似度。最後以問句相似度和符合之 pattern 權重的乘積做為這個可能答案的分數，將分數較高的答案優先回覆給使用者。

設定 patterns 權重時，我們以 Penn Treebank 及人工評估的方式，得到各 patterns 關鍵字的準確率，準確率越高的 pattern 有越高的權重。

進行效能評估時，我們以另兩個以網際網路為基礎的問答系統 (AnswerBus 和 LCC) 來與我們的系統做比較，發現我們系統的效能優於另外兩個線上系統。以 TREC 中 QA 評比的 MRR 值來評估，AnswerBus、LCC 和我們系統的 MRR 值分別為 0.429、0.229 和 0.623。

在未來的工作中，[EVENT] 與問句相似度的比較會是一個重要的研究議題。除了關鍵字與字根比對外，還可試著加入語法或語意上的比較，或者使用 WordNet 來進行關鍵字的擴充，甚至是處理代名詞指涉問題等來加強相似度比較的正確性。不過如果是基於網際網路的問答系統，必須考慮反應時間的長短，所以也不適宜使用太複雜的相似度比較方法。

而如何修改 patterns，或增加比對上的限制，藉以提升 patterns 的準確率，是未來研究的另一個重點。此外，當 patterns 涵蓋兩句以上的段落時，如何確定答案的邊界就是一個值得研究的課題。

參考文獻

Anand, Pranav; Breck, Eric; Brown, Brianne; Light, Marc; Mann, Gideon; Riloff, Ellen; Rooth, Mats and Thelen, Michael (2000) “Fun with Reading Comprehension,” Final report, Reading Comprehension group, Johns Hopkins Center for Language and Speech Processing Summer Workshop 2000. Johns Hopkins University, Baltimore MD. [Online] Available URL:

http://www.clsp.jhu.edu/ws2000/groups/reading/ WS00_readcomp_final_rpt.pdf

Girju, Roxana and Moldovan, Dan (2002) "Mining Answers for Causation Questions," *Proceedings of the American Association for Artificial Intelligence (AAAI) - Spring Symposium*, Stanford University, California, USA, March 2002.

Harabagiu, Sanda; Moldovan, Dan; Pasca, Marius; Mihalcea, Rada; Surdeanu, Mihai; Bunescu, Razvan; Girju, Roxana; Rus, Vasile and Morarescu, Paul (2000a) "FALCON: Boosting Knowledge for Answer Engines," *Proceedings of the Text REtrieval Conference (TREC-9)*, Gaithersburg, Maryland, USA, November 2000, pp. 479-488.

Lin, Jimmy (2002) "The Web as a Resource for Question Answering: Perspective and Challenges," *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, May 2002.

Marcus, Mitchell P.; Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993) "Building a large annotated corpus of English: the Penn Treebank," *Computational Linguistics*, Vol. 19, No. 2, June 1993. pp. 313-330.

Moldovan, D. and Rus, V. (2001) "Logic Form Transformation of WordNet and its Applicability to Question Answering," *Proceedings of the ACL 2001 Conference*, Toulouse France, July 2001, pp. 394-401.

Radev, Dragomir R.; Qi, Hong; Zheng, Zhiping; Blair-Goldensohn, Sasha; Zhang, Zhu; Fan, Weiguo and Prager, John (2001) "Mining the Web for Answer to Natural Language Questions," *Proceedings of the ACM CIKM-2001*: *Tenth International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA, November 2001.

Radev, Dragomir; Fan, Weiguo; Qi, Hong; Wu, Harris and Grewal, Amardeep (2002) "Probabilistic Question Answering on the Web," *Proceedings of the eleventh International World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, USA, May 2002.

Riloff, Ellen and Thelen, Michael (2000) "A Rule-based Question Answering System for Reading Comprehension Tests," *Proceedings of the ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Seattle, Washington, USA, May 2000.

Voorhees, E. (1999) "The TREC-8 Question Answering Track Evaluation," *Proceedings of the Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland, USA, November 1999, pp. 23-37.

Zheng, Zhiping (2002) "AnswerBus Question Answer System," *Proceedings of Human Language Technology Conference (HLT 2002)*, San Diego, California, USA, March 2002.

基於自然語言處理技術的研究主題抽取與分析

Extraction and Analysis of Research Topics Based on NLP Technologies

世新大學資訊傳播學系
Department of Information and Communications, Shih-Hsin University
林頌堅
Sung-Chen Lin
Email: scl@cc.shu.edu.tw

摘要

本論文針對研究主題分析的問題，提出一系列以自然語言處理為基礎的技術，從學術領域中發表的論文資料中抽取重要的關鍵詞語，並將這些詞語依據彼此間共現關係進行叢集，以叢集所得到的詞語集合表示領域中重要的研究主題。研究主題分析在學術領域的應用上，可以提供研究人員一個清楚的梗概；在資訊檢索的過程中，則可以幫助使用者釐清資訊需求。我們並將所提出的方法應用到 ROCLING 研討會的論文資料上，抽取計算語言學領域的重要研究主題。結果顯示這個方法可以應用於國內學術領域的特殊環境，同時抽取出中文和英文的關鍵詞語，所得到的詞語叢集結果也可以表示領域中重要的研究主題。這樣的結果初步的驗證了本論文所提出方法的可行性。從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，抽取出來的詞語叢集中有許多與機器翻譯、語音處理和資訊檢索相關，在語言的計算模式上，語法模式與剖析、斷詞和統計式語言

模型的建立則是國內計算語言學家所關心的主題。

一、緒論

資訊檢索研究著重的問題是人與資訊之間的介面，近來的研究趨勢注重於使用者所具有的背景知識、在檢索過程中對問題的認知[Wilson, 1999]及資料的嫻熟程度(material mastery)[Bishop, 1999][Covi, 1999]。為了對一個學術領域的資訊傳播現象進行全面的了解，所謂的「領域分析」(domain analysis)藉由對學術領域內重要的學術活動，諸如研究、論文發表、會議參與等等進行分析，探討研究人員所使用或產生的知識組織、結構、合作模式、語言和通訊形式、資訊系統以及相關標準等[Hjørland and Albrechtsen, 1995] 。而研究主題分析可以說是領域分析的一項要務，了解重要的研究主題可以掌握領域中的知識組織，幫助使用者釐清資訊需求(information need)，迅速取得所需的資訊。此外，藉由有系統的方法抽取研究主題並加以分析，可以展示學術領域研究一個完整的面貌，提供新進學者在初期進入領域時的參考，也可以作為學術研究領域發展的指引(road map)，提供已經深入的研究人員擴展學術研究的範疇。

本論文提出一個自動化的研究主題抽取方法，從學術領域中發表的論文集合中選出關鍵詞語，再依據詞語彼此間出現在相同論文中具有特定意義的共現(co-occurrences)現象，辨認每一篇論文中可能具有的研究主題，作為分析這個領域重要研究主題的依據。我們認為論文的豐富詞彙訊息蘊含了研究主題。在論文發

表的過程中，作者藉由論文題名、摘要以及本文中的詞語將研究的問題、方法與結果等主題傳達給讀者，甚至論文所引用的參考文獻題名也包含許多與主題相關的詞語訊息；而讀者在閱讀論文時，便可以依據這些詞語判斷與本身研究興趣上的相關性，同時將這些資訊建構與融入個人的知識結構中[Harter, 1992]。以本論文做一例子，在本論文的題名、摘要和本文中包含了許多『學術領域』、『研究主題』、『論文』等等詞語，目的是希望讀者在閱讀時，可以從這些詞語的共同出現與使用，了解我們所研究的主題是從學術論文中抽取重要的研究主題，而有興趣的讀者在閱讀後，便可在研究與發表上加以利用。進一步地，在一個學術領域中，可以發現某些受到重視的研究主題相關的詞語在許多論文中出現。以計算語言學領域來看，便可以發現諸如『語料庫』、『剖析』、『資訊檢索』等等的詞語在許多論文中出現，這些都是這個領域中的重要研究主題。而且與研究主題相關的一組詞語會重複出現在許多論文中。因此，如果對學術領域出版的論文進行分析，選取具有代表主題意義的詞語，統計這些詞語間的共現現象，利用這些資訊將經常一起出現的一組詞語叢聚成一個集合，所形成的詞語集合可以視為是某一特定的研究主題。在分析某一論文的主題時，便可以估算代表各研究主題的詞語叢聚與該論文的相關性，作為判斷該論文是否具有此一主題的資訊。因此，本論文嘗試利用自然語言處理技術來分析學術領域中發表的論文，確認論文中出現的詞語，抽取蘊含在其中詞語的共現訊息，再進行詞語叢聚(term clustering)，作為辨認主題分析的資訊。

我們並將所發展出來的技術應用於國內計算語言學領域的主題分析。選擇以計算語言學作為研究對象的主要原因是這個領域具有科際整合研究(interdisciplinary research)的特色，並且成功地將發展出的理論和技術應用到學術研究與實際的系統和產品研發[Lenders, 2001]。參與這個領域研究的研究人員主要來自於語言學和計算機科學兩個學科，對於計算機科學家來說，主要的研究工作在於建構一個實用的電腦系統來處理有關自然語言的問題，比方說機器翻譯、字型辨認、語音辨認、資訊檢索等等。語言學家的工作則在於計算性理論的規範與應用，用來解釋自然語言的認知現象模式及模擬驗證的能力[王士元, 1988]。計算機科學家的工作需要依賴語言學家所形成的語言理論來建立合理而有效率的電腦系統；而語言學家則是利用計算機科學家所發展的計算理論與系統探究自然語言的規律[Huang, 2000]。在這個領域中的重要研究除了將計算機方法應用於自然語言理論的探討之外，最受矚目的研究還包括利用語料庫(corpus)所發展出來的語言理論及利用這些理論設計與發展各種實務系統[Church and Mercer, 1993]。所以，從這個領域所進行的學術活動，可以觀察語言學和計算機科學兩種不同學科的學者在互相激盪下產生的成果，也可以觀察到從理論的研究到技術發展，再到實務的應用，對研究主題分析是一項具有挑戰且有意義的研究。除此之外，另一方面則是我們對於這個領域的熟悉，將有助於研究方法的發展，對於所得到的初步結果做出合理的詮釋，並作為下一階段改進的參考。

在使用 ROCLING 一到十四屆的學術研討會論文資料，共 235 篇，我們共抽

取出 343 個關鍵詞語。研究主題叢集後得到 34 個代表重要研究主題的詞語集合。結果顯示所發展的詞語抽取法可以同時抽取出中文和英文的關鍵詞語，所得到的詞語叢集結果也可以表示領域中重要的研究主題。初步驗證了本論文所提出方法的可行性。從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，抽取出來的詞語叢集中有許多與機器翻譯、語音處理和資訊檢索相關，在語言的計算模式上，語法模式與剖析、斷詞和統計式語言模型的建立則是國內計算語言學家所關心的主題。

本論文其餘的章節架構如下：在第二節中首先說明一些相關研究及本論文所提出研究主題分析方法的概念與合理性，並且依據這些概念設計出利用一系列自然語言處理技術進行研究領域分析的方法。接著在第三節和第四節中分述這個方法中的核心技術：詞語抽取和研究主題叢聚。第三節中，我們提出了在多語環境下的關鍵詞語抽取方法，可以從中英文論文資料中取得代表研究主題的關鍵詞語。第四節則提出一個詞語叢聚方法，利用詞語的共現關係，將詞語進行多重叢聚來代表可能的研究主題；本節中並且說明研究主題與論文之間相關程度的計算方式。第五節中報告將此分析方法應用到國內計算語言學研究的結果。最後，第六節則是結論。

二、本論文提出的研究主題分析方法

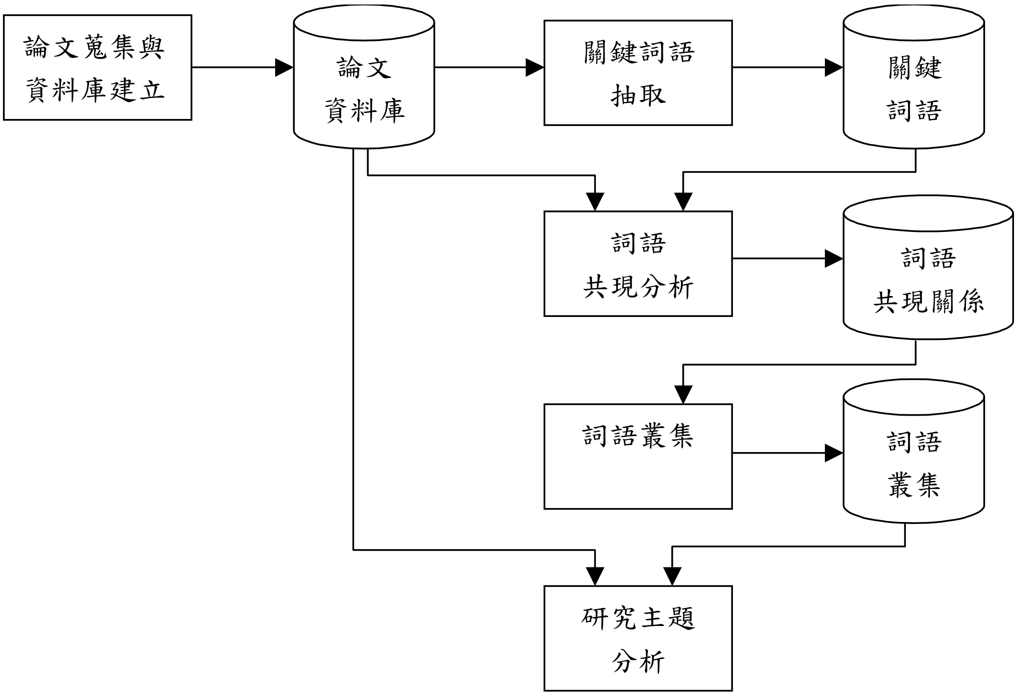
本論文希望發展一套研究主題抽取方法，可以從特定學術領域中出版的論文

中，抽取用來表達研究人員共識的重要研究主題，作為進一步分析或資訊檢索應用的資訊。在資訊檢索研究範疇中與這個問題相近的研究有主題偵測(topic detection)。主題偵測是希望從一序列來源各不相同的新聞中，偵測出與某些『事件』(events)相關的連續報導[Wayne, 2000]。目前研究人員認為『叢聚假說』(cluster hypothesis)可以適用於解決這個問題[Yang, Pierce and Carbonell, 1998][Hatzivassiloglou, Gravano and Maganti, 2000]，利用具有相關主題的文件具有相似的詞語分布情形，以文件叢聚(document clustering)技術，偵測新進文件是否與現有文件集合具有相似的詞語分布情形，將文件歸入相關事件的集合中；若文件與現有集合皆不相近，則視為是一個新事件，產生一個新的集合。因此，我們嘗試應用叢聚假說，探索領域中可能的研究主題。再者，主題偵測研究已應用專有名詞(proper nouns)等相關詞彙作為區隔不同新聞事件的重要訊息[Hatzivassiloglou, Gravano and Maganti, 2000]，本論文也將嘗試利用論文中與領域相關的詞語抽取出來作為分析的主要訊息。此外，主題偵測應用所謂的『新聞熱潮』(news bursts)現象，將時間訊息加入叢聚演算法，提昇偵測的結果[Yang, Pierce and Carbonell, 1998]。但是學術論文雖然有所謂『資訊流行』(information epidemic)[Tabah, 1996]的說法，也就是在某一項新的理論、研究方法或技術提出後，如果得到很大的成功，將可以吸引許多研究人員投入後續的研究中，造成一股相關研究發表的風潮，然而在實證研究中卻發現此一現象雖然存在但並不常見[Tabah, 1996]，所以在本論文並不考慮加入時間訊息。

在本論文中，我們利用相同研究主題的論文中具有相似詞彙訊息的概念，利用在論文中詞語的共現關係，找出詞語的叢聚情形來代表研究主題。以論文中出現的詞語取代整篇論文作為分析對象的主要原因是希望能獲得較可信賴的統計訊息。並非所有的研究領域都有足夠多的學術論文發表可供進行研究，較小的學術領域所出版的論文數量較為不足，以整篇論文進行分析，統計上不容易得到研究主題的分析結果。以關鍵詞語作為分析對象，可以獲得充足的統計訊息，克服文件數量較少的問題。某些論文具有多個研究主題也可藉由詞語的多重叢聚加以表示，進而探索研究主題之間的關係。此外，文件叢聚不易詮釋結果所代表的主題，詞語叢聚則容易直接由成員的語義進行解釋。

本論文方法的架構如圖一所示。首先對需要進行分析的學術領域蒐集相關論文資料，建立論文資料庫。資料庫中收錄的資料包括論文的題名、摘要和參考文獻的題名等作為詞語抽取與叢聚分析的資訊，論文作者和出版年等項目則用來作為後續研究主題的分析工作上。特別值得一提的是，國內的學術論文基本上是中、英語雙語並行，許多領域皆接受論文以中文或英文發表，然而並非所有的論文同時具有中、英文雙語的題名和摘要，無法單就某一種語言的文本進行分析。若只考慮以某一種語言發表的論文進行分析，而忽略另一種語言，有可能造成某些特殊的研究主題被遺漏的情形。若是分別處理各種語言的論文，缺乏分屬兩種語言的詞語在論文中的共現訊息，無法分析出這些詞語的相關性，在整合上有相當大的困難。因此需要考慮這個特殊的論文發表現象，提出可以同時分析兩種語言論

文的方法。本論文所提出的解決之道是加入論文中參考文獻的題名進行分析，通常參考文獻的題名與研究的理論、方法及技術等也有密切的關係，而且參考文獻的題名可能包含兩種語言，若能提出適當的多語詞語抽取方法，便可以統計分屬兩種語言的相關詞語的共現現象，整合兩種語言的詞語訊息，而得到較佳的研究主題分析結果。



圖一　本論文提出的研究主題抽取與分析方法

在建立好論文資料庫後，接著便利用多語的關鍵詞語抽取方法從論文資料中自動抽取領域中具有意義的詞語，統計詞語在論文中的共現關係，利用這些資訊將相關的詞語叢集成集合，用來代表某一個特定研究主題。在進行研究領域分析時，當詞語叢集與某一論文的相關性(relevance)足夠強時，可以假定該論文具有該詞語叢集所代表研究主題。下面的兩節中，將針對多語環境下的重要詞語抽取以

及詞語叢聚技術詳細說明。

三、多語環境下的關鍵詞語抽取

為了抽取可以代表學術領域研究主題的關鍵詞語，我們首先確認論文資料中重要的中英文詞組以及中文的多字詞，增強詞語的語彙訊息，再選擇具有代表研究主題意義的詞語，作為這一階段的結果。在學術論文中，常以詞組的形式表達重要的研究主題，比方在計算語言學領域的論文中，可以發現諸如英文的“language model”、“machine translation”或是中文的“語言模型”、“機器翻譯”等等。此外，中文的文本裡，詞與詞之間沒有明顯的界限，進行自然語言處理前，需要先進行分詞，確認文本內可能的詞。所以要進行研究主題分析，首要工作是從論文中確認重要的中英文詞組以及中文的多字詞。然而學術論文中經常有許多新的詞語出現，來代表新的概念、方法和技術，我們無法事先收錄各個領域裡所有可能的詞語來製作十分完整的詞典，進行斷詞。而且利用構詞律的規則式斷詞方法，需要處理同時中文和英文兩種語言的文本，難以整合應用。所以本論文採用統計式的處理方法[Chien, et. al., 1999]，以便同時解決中文的多字詞及中英文的詞組問題。

本論文所使用的方法如下：首先利用題名、摘要和參考文獻的題名等論文資料裡所有的文句建立一個 PAT-tree 資料結構，用來儲存所有出現在論文資料中的字串及它們所在的論文資料[Chien, 1997]。接著在 PAT-tree 中擷取可能的字串作為候選詞語，以統計訊息及經驗法則(heuristic rules)作為判斷字串是否為詞語的標準。

在本論文中，所使用的統計訊息包括字串在所有資料中的出現總頻次、字串在出現論文中的平均頻次和標準差(standard deviation)以及字串前後接字的複雜度。字串的出現總頻次代表該字串在領域中的重要性，出現頻次高表示這個字串在領域裡的論文經常出現而具有重要意義。字串在出現論文中的平均頻次和標準差用來表示字串對出現論文的重要程度，如式(1)

$$R_S \overset{def}{=} m_S + \sigma_S \tag{1}$$

在式(1)，m_S和 σ_S分別代表字串 S 在出現論文中的平均頻次和標準差。當字串 S 的平均頻次超過某一閾值時，表示此字串極有可能在許多論文中出現多次，是這些論文的關鍵詞語，應該被選取出來。或是雖然字串 S 在論文的平均頻次較低，但在某些論文中出現多次，是這些論文的關鍵詞語，也需要被選取出來，此時字串 S 會有一個較大的標準差 σ_S。因此，我們可以利用字串在出現論文中的平均頻次和標準差的總和 R_S代表字串對出現論文的重要程度，R_S值愈高的字串對出現論文愈重要。

字串前後接字的複雜度則可以判斷是否是一個完整的詞語或是其他詞語的部分，字串 S 的前後接字複雜度 C_{1S}和 C_{2S}分別如式(2a)和(2b)所示

$$C_{1S} \overset{def}{=} -\sum_a \frac{F_{aS}}{F_S} \log(\frac{F_{aS}}{F_S}) \tag{2a}$$

$$C_{2S} \overset{def}{=} -\sum_b \frac{F_{Sb}}{F_S} \log(\frac{F_{Sb}}{F_S}) \tag{2b}$$

式(2a)和(2b)中，a 和 b 代表字串 S 在論文資料中任一個可能的前接字和後接字，F_S、F_{aS}和 F_{Sb} 分別是字串 S、aS 和 Sb 的出現總頻次。以式(2a)前接字的情形來看，若是字串 S 有愈多種類的前接字，而且每一種前接字出現的次數越接近時，C_{IS} 的值愈大，反之，當字串前只有一種前接字時，C_{IS} 的值等於 0，或是有一個前接字出現的機會較其他大非常多時，則 C_{IS} 的值接近於 0，表示該字串再加上這個前接字可能才是一個詞語。愈大的前接字複雜度代表該字串愈有可能是獨立的詞語而不是其他詞語的一部分；後接字的情形也是相同的道理。

通過上面條件的字串，再利用停用詞(stop words)不能出現在字串首尾的經驗法則，進一步過濾去不完整的詞語。在過去的經驗中，介詞和定詞等停用詞常出現在抽取出字串的首尾，如 “名詞+的”、“名詞+of”或“to+動詞”等詞組結構。但停用詞出現在字串的中間代表特定的詞組，例如“part of speech”，因此，將這種情形加以保留。

在確認論文資料中重要的中英文詞組以及中文的多字詞後，以這些詞語建立斷詞處理所需的詞典。我們使用長詞優先法則與詞語的出現總頻次將所有論文資料加以斷詞，確認所有在論文資料中出現的詞語。此時，我們分出來的詞語包括了一些中英文的詞組、詞和一些中文單字。在本論文中並非需要確認論文資料中所有可能的詞語，而是希望抽取所有可能代表研究主題的關鍵詞語，因此我們過濾具有以下情形的詞語。首先是中文資料中的單字(characters)，多半是一些介詞、

停用詞或是在上一步驟中無法組成詞的詞語，加以濾掉。其次，出現總頻次與式(1)之 R_S 值太小的詞語，也加以過濾，其理由如前面所述。剩下的詞語與它們在論文資料中的出現情形則是下一階段分析的對象。

四、研究主題叢聚

為了探索學術領域中重要的研究主題，本論文依據詞語在論文資料的共現關係，建立詞語之間的相關程度，將詞語進行叢聚，以一組叢聚的相關詞語作為一個研究主題。由於有些詞語可能包含在不同的研究主題中，本節中提出一個可以對詞語進行多重叢集的演算法。

首先，我們將上一階段抽取出來的詞語，利用可以進行多重叢聚的 cliques 叢集演算法[Kowalski and Maybury, 2000]進行詞語叢集。在選定最小相關程度的情形下，我們可以得到若干個詞語叢集，在這些詞語叢集中的詞語，彼此間的相關程度都在所選定的最小相關程度之上；而且詞語因它們與其他詞語相關程度的不同，可以叢集在多個集合中。本論文所使用的詞語相關程度的計算方式如下：我們先計算每一詞語在每一篇論文的題名、摘要和參考文獻的題名等資料中出現的頻次，作為詞語的特徵值運算的資訊。但因為只取用上述論文資料的資料量較小，詞語在其中出現的頻次不會太高，為了使低頻次的詞語差異不會太大，以詞語在每一篇論文資料的頻次的平方根作為一個特徵值。如此一來，對每一詞語有一組特徵向量(feature vector)，計算詞語間的相關程度便可以利用所對應特徵向量間夾

角的餘弦值(cosine value)來估算。如式(3)和式(4)分別表示詞語 A 的特徵向量和詞語 A 與 B 間的相關程度估算方式。

$$\vec{v}_A' \overset{def}{=} [\sqrt{f_{1,A}}, \sqrt{f_{2,A}}, \ldots, \sqrt{f_{N,A}}] \tag{3}$$

$$R(A,B) \overset{def}{=} \frac{\vec{v}_A \cdot \vec{v}_B}{\|\vec{v}_A\| \|\vec{v}_B\|} \tag{4}$$

式(3)中，$f_{i,A}$ 代表詞語 A 在第 i 篇論文資料中出現的頻次。式(4)中，分子部分是詞語 A 和 B 的特徵向量 $\vec{v}_A$ 和 $\vec{v}_B$ 內積(inner product)的值，分母部分則是兩個特徵向量長度 $\|\vec{v}_A\|$ 和 $\|\vec{v}_B\|$ 的乘積。經過 cliques 演算法所得到的結果是相當嚴格的，只有以詞語的共現關係所估算的相關程度在某一閾值以上的一對詞語才有可能叢集在一個集合內。然而，在研究主題中相同或相近的概念可能以不同詞語來表示，這些詞語不一定出現在相同的論文資料中，利用上述以詞語共現現象的相關程度估算方法將會得到很小的相關程度估算值，無法利用 cliques 演算法將這些詞語叢集起來。本論文採用以下兩種技術來解決上述的問題。

首先我們利用 LSI(Latent Semantics Indexing)技術對上述的特徵向量所形成的『詞語-特徵』矩陣 M 進行奇異值分解 (SVD, singular value decomposition)運算 [Deerwester, et. al., 1990]，將矩陣 M 分解成三個矩陣，T_o、S_o 和 D_o，使得 $M = T_o S_o D_o'$。此處 T_o 和 D_o 為 M 的左、右奇異向量(singular vectors)所形成的矩陣，其大小分別為 t×r 和 d× r，t 和 d 分別為詞語和特徵的數目，r 則為矩陣 M 的秩(rank)，而 S_o 為一個大小為 r×r 的對角線矩陣(diagonal matrix)，其對角線上的值為

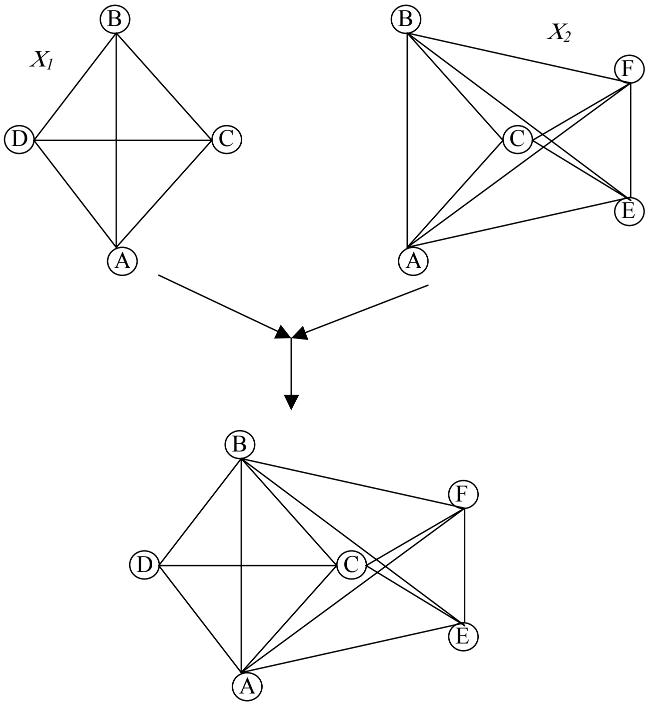
M 的奇異值(singular values)，且依據遞減的方式排列。若我們希望取得一個秩為 k 的矩陣 $\hat{M}$ ，$k<=r$，並使得 $\hat{M}$ 與 M 的最小平方差(least square error)最接近，可以取 S_o 對角線上的前 k 個奇異值，產生一個大小為 $k\times k$ 的新矩陣 S，同時 T_o 和 D_o 也分別取前 k 個行向量(column vectors)，形成矩陣 T 和 D，大小分別為 $t\times k$ 和 $d\times\ k$。矩陣 $\hat{M}$ 便可由 $\hat{M} = TSD'$ 計算得到。在使用 LSI 技術的檢索過程，當進行詞語的相關程度估算時，以 $\hat{M}\hat{M}'$ 來估算原先以 *MM'*計算兩兩詞語特徵向量間的內積值，如式(5)所表示，

$$MM' \approx \hat{M}\hat{M}' = TSD'(TSD')' = TSD'DS'T' = TSS'T' = TS^2T' \qquad (5)$$

在式(5)中，由於矩陣 D 中的行向量彼此互為單位正交(othonormal)，*DD'=I*，而且 S 為對角線矩陣，*S'*= S，所以 $\hat{M}\hat{M}' = TS^2T'$。利用 SVD 取得隱含語義結構(latent semantic structure)的特性，使得原先因為共現關係較弱或是不存在，而相關程度較估算得很小的兩個相關詞語，可以獲得較大的估算值[Deerwester, et. al., 1990]。

其次，進行 cliques 叢集演算法後，我們對於所得到的結果依據它們成員間重疊的情形再次進行叢集。假設兩個叢集之間有三個以上的成員是相同的，而且其餘的成員間雖然沒有很強的詞語共現關係，但是也曾在某些論文資料中一起出現，我們即將這兩個叢集的詞語集合進行聯集，產生新叢集。如圖二所示， 在 A、B、C、D、E 和 F 六個相關詞語中，依據它們的共現關係進行 cliques 叢集，叢集成X_1和X_2兩個詞語集合。在這兩個叢集間，有三個詞語 A、B 和 C 是相同，而且

經過比對，叢集X_1剩下的成員 D 和叢集X_2剩下的成員 E 和 F 出現的論文資料有一些是相同的，我們便將X_1和X_2兩個叢集進行合併，使得所得到的叢集更具有研究主題的代表性。



圖二 將X_1和X_2兩個具有相同成員的叢集進行合併的示意圖

最後經過上述的叢集處理後，可以得到一些代表領域中重要研究主題的詞語叢集。在分析研究主題時，我們計算每一叢集與論文間的相關程度，計算方式依據為 LSI 的相關估計方式[Deerwester, et. al., 1990]，如式(6)計算叢集X對所有論文的相關程度。

$$R_X = \chi TSD' \tag{6}$$

式(6)中，χ為一個行向量，$\chi' = [e_1, e_2, \ldots, e_t]$，每一個元素代表一個特定詞語是

否出現在叢集X之中，換言之，如果第 i 個詞語包含於這個叢集中，則 e_i 的值為 1；否則若是這個叢集不包含這個詞語，e_i 的值為 0。式(6)所得到的結果 R_X也是一個行向量，大小為 $1\times d$，每一個元素所代表的值為詞語叢集X與所對應的論文之間的相關程度估算值。最後依據這個結果，將相關程度大的論文資料取出，作為研究主題相關的論文資料來進行分析。

五、國內計算語言學的研究主題分析的實驗結果

計算語言學研討會 ROCLING 是國內的計算語言學領域相當重要的學術活動。因此，ROCLING 的研討會論文集中論文資料，可以說是歷年來國內計算語言學領域學者的心血結晶，所蘊含的研究主題也是他們所關心的研究主題。因此，本論文將以 ROCLING 研討會的論文資料做為分析國內計算語言學研究主題的素材。

分析資料為從第一屆(1988)到第十四屆(2001)的 ROCLING 研討會論文，共 235 篇。進行詞語抽取時，首先抽取重要的多字詞及詞組，所設定的字串出現總頻次的閾值，較短的字串(2 或 3 字)設定為 15 次，較長的字串(4~5 字)則設定為 10 次，字串對出現論文資料的重要程度 R_S(平均頻次和標準差的總和)設為 2.5，前後接字的複雜度設定為 0.5。這些抽取出來的多字詞或詞組加入詞典後，對論文資料進行分詞，依據第三節的方法對所有詞語進行統計，過濾去不重要的詞語。最後的結果共得到 343 個關鍵詞語。由於篇幅所限，無法將所有的詞語一一列出，我們將

出現總頻次最高的前 50 個詞語及出現的總頻次列表於表一。

表一 關鍵詞語抽取所得到的前 50 個出現總頻次最高的詞語及總頻次

次序	詞名	出現總頻次	次序	詞名	出現總頻次
1	parsing	209	26	parser	80
2	speech	184	27	probabilistic	78
3	系統	175	28	動詞	78
4	sentences	141	29	語音	78
5	lexical	138	30	knowledge	74
6	mandarin	134	31	語法	74
7	speech recognition	132	32	chinese text	73
8	方法	131	33	語言	73
9	semantic	130	34	semantics	72
10	corpus	129	35	corpora	71
11	syntactic	107	36	used	71
12	recognition	106	37	國語	71
13	data	105	38	discourse	70
14	分析	104	39	處理	70
15	learning	102	40	dictionary	68
16	mandarin chinese	97	41	problem	65
17	sentence	97	42	分類	65
18	machine translation	92	43	corpus based	64
19	words	92	44	design	62
20	theory	87	45	information retrieval	62
21	rules	84	46	syntax	61
22	models	83	47	generation	60
23	phrase	83	48	語料庫	60
24	漢語	82	49	應用	60
25	classification	80	50	character	59

接著將取出來的詞語進行研究主題叢聚。進行詞語的 cliques 叢集時，我們分別以第四節中原先的詞語特徵向量與經過 SVD 處理的特徵向量，*k* 值為 30、60 及 120，進行相關程度估算。將相關程度的閾值設為 0.4，經過 cliques 叢集與叢集合

併後，所得到三個詞語以上的叢集的數目，如表二所示。

表二 不同相關程度估算方法進行研究主題叢集所得到的叢集數目

	Original feature vectors	SVD k=120	SVD k=60	SVD k=30
cliques 叢集	65	78	85	74
叢集合併	27	34	34	32

從表二中，可以觀察到經過 SVD 處理的 cliques 叢集數目較原先的特徵向量來得多，顯然 LSI 技術有助於捕捉詞語不共現卻相關的隱含語義結構，產生較多 cliques 叢集。因此，我們以經 SVD 處理 *k* 值為 60 的特徵向量進行詞語相關程度估算，將所得到的 34 個詞語叢集作為進一步的分析的對象，這 34 個詞語叢集列表於附錄一。

從詞語叢集的結果我們可以看到幾個現象。第一、若干叢集同時具有中文詞語與英文詞語，甚至包含縮寫與相同概念但不同詞名的詞語，比方說，叢集 12 包含了‘machine translation’、‘mt’、‘機器翻譯’等詞語；或是又如叢集 18 包含了 ‘word identification’、‘word segmentation’、‘斷詞’等詞語。可見將參考文獻的題名加入論文資料，可以獲得中文和英文兩種語言的詞彙訊息，而且利用詞語的共現關係可以將相關的詞語叢聚起來。第二、大部分的詞語叢聚都可以明顯地用來代表一個特定的研究主題。除了叢集 3、叢集 11 與叢集 29 由意義較廣泛的詞語形成之外，其餘叢集的詞語間都具有相關性，而且可以用來代表計算語言學領域中的特定研究主題。比方說，叢集 7 為語音辨認的相關詞語、叢集 9 則為文件分類的相關詞

語。因此，本論文所提出來的研究主題抽取方法的可行性便可以得到初步驗證。

表三 與語言的計算模式相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
23 語法模式與剖析	分析，表達，剖析，格位，訊息，動詞，結構，詞類，漢語，語法，語法模式，語意，模式，關係	1989 "訊息為本的格位語法--一個適用於表達中文的語法模式" 1991 "連接詞的語法表達模式-以中文訊息格位語法(ICG)為本的表達形式" 1992 "漢語的動詞名物化初探--漢語中帶論元的名物化派生名詞"
18 斷詞	chinese text, chinese word segmentation, segmentation, unknown word, word identification, word segmentation, words, 斷詞	1994 "Chinese-Word Segmentation Based on Maximal-Matching and Bigram Techniques" 1995 "A Unifying Approach to Segmentation of Chinese and Its Application to Text Retrieval" 1997 "Unknown Word Detection for Chinese by a Corpus-based Learning Method" 1997 "Chinese Word Segmentation and Part-of-Speech Tagging in One Step" 1997 "A Simple Heuristic Approach for Word Segmentation"
22 統計式語言模型的建立	bigram, class based, clustering, entropy, language model, language modeling, language models, n gram	1994 "An Estimation of the Entropy of Chinese - A New Approach to Constructing Class-based n-gram Models" 1997 "Truncation on Combined Word-Based and Class-Based Language Model Using Kullback-Leibler Distance Criterion" 2001 "使用關聯法則為主之語言模型於擷取長距離中文文字關聯性"

由於篇幅的限制，本論文無法對所有抽取出來詞語叢集一一進行詳盡的報告，以下針對幾個主題較明確的詞語叢集進行說明。表三是與語言的計算模式相關的詞語叢集及相關論文的列表，論文前的數值是論文在 ROCLING 研討會中發表的年份。表三可以驗證早期的計算語言學多以規則式的語法模式與剖析為主，近來則較多發展統計式語言模型，而斷詞則是一直以來國內計算語言學領域相當重視的獨特問題。

此外從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，表四到表六分別列出與機器翻譯、語音處理和資訊檢索相關的集合。從表四的結果，說明機器翻譯是計算語言學最早的應用問題之一[Lenders, 2001]，而其發展從規則式的自動翻譯到統計式，近期的應用則是在跨語言檢索部分。

表四 與機器翻譯相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
12 機器翻譯	'bilingual', 'machine translation', 'mt', 'transfer', '機器翻譯'	1991 "Lexicon-Driven Transfer In English-Chinese Machine Translation" 1992 "A Modular and Statistical Approach to Machine Translation" (只有與叢集 12 相關) 1995 "THE NEW GENERATION BEHAVIORTRAN: DESIGN PHILOSOPHY AND SYSTEM ARCHITECTURE" 1996 "介詞翻譯法則的自動擷取" 2001 "統計式片語翻譯模型"
32 機器翻譯	'bilingual', 'machine translation', 'translation', '機器翻譯'	

在過去計算語言學所處理的對象多為書寫語言(orthographic languages)，近年來語音處理已經成為計算語言學相當重視的研究主題。從 ROCLING 的論文資料中所得到的結果可以分析成語言模型、聲學辨認以及語音合成三個研究主題(表五)。國內計算語言學較早進行研究的主題是語言模型和語音合成，近年在聲學辨認研究上，也有許多研究人員進入這個領域發表相關論文。在表五，另外還可將語音合成研究分成系統製作(叢集 30)與聲學訊息研究(叢集 31)兩個部分。

表五 與語音處理相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
13 語言模型	dictation, large vocabulary, 語言模型, 語音辨認	1993 "國語語音辨認中詞群雙連語言模型的解碼方法" 1994 "國語語音辨認中詞群語言模型之分群方法與應用" 1995 "應用於'音中仙'國語聽寫機之短語規則分析與建立" 1996 "國語語音辨認中多領域語言模型之訓練、偵測與調適" 1999 "國語電話語音辨認之強健性特徵參數及其調整方法" (只有與叢集 17 相關)
17 語言模型	國語, 語言模型, 語音辨認, 辨認	
7 聲學辨認	hidden markov, maximum, robust speech recognition, speech recognition	1998 "Speaker-Independent Continuous Mandarin Speech Recognition Under Telephone Environments" 1999 "國語電話語音辨認之強健性特徵參數及其調整方法" 2000 "具有累進學習能力之貝氏預測法則在汽車語音辨識之應用" 2000 "綜合麥克風陣列及模型調整技術之遠距離語音辨識系統"
30 語音合成	speech, synthesis, 文句翻語音, 合成, 系統, 音節, 國語, 連音, 語音, 輸入	1995 "以 CELP 為基礎之文句翻語音中韻律訊息之產生與調整" 1996 "時間比例基週波形內差--一個國語音節信號合成之新方法" 1996 "中英文文句翻語音系統中連音處理之研究" 1999 "台語多聲調音節合成單元資料庫暨文字轉語音雛形系統之發展" (只有與叢集 30 相關) 1999 "國語文句翻台語語音系統之研究" (只有與叢集 30 相關) 2001 "Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method", (只有與叢集 31 相關)
31 語音合成	mandarin text to speech, pitch, prosodic, speech, synthesis, 文句翻語音, 合成	

在計算語言學領域中，資訊檢索比起其他研究可說是一個較新的主題，然而由於網際網路與電子文件的發展使得這項應用成為相當具有潛力的研究主題。我們可以從表六中發現國內計算語言學在這方面的重要研究包括資訊檢索和文件分類。

表六 與資訊檢索相關的詞語叢集及相關論文

叢集編號	詞語	相關論文資料
25 資訊檢索	csmart, databases, document, indexing, information retrieval, retrieval, text retrieval, 檢索	1995 "適合大量中文文件全文檢索的索引及資料壓縮技術" 1996 "尋易(Csmart-II):智慧型網路中文資訊檢索系統" 1997 "An Assessment on Character-based Chinese News Filtering Using Latent Semantic Indexing" 1999 "A New Syllable-Based Approach for Retrieving Mandarin Spoken Documents Using Short Speech Queries"
9 文件分類	document, hierarchical, text categorization, 分類, 文件, 文件分類, 特徵	1993 "中文文件自動分類之研究" 1999 "階層式文件自動分類之特徵選取研究" 2001 "基於階層式神經網路之自動文件分類方法" 2001 "適應性文件分類系統"
28 文件分類	document, text categorization, 分類, 文件分類, 文件自動, 關鍵詞	

六、結論

本論文針對研究主題分析的問題，提出一系列以自然語言處理為基礎的技術，從學術領域中發表的論文資料中抽取重要的關鍵詞語，並將這些詞語依據彼此間共現關係進行叢集，以叢集所得到的詞語集合表示領域中重要的研究主題。在本論文中，我們將所提出的方法應用到 ROCLING 研討會的論文資料上，抽取計算語言學領域的重要研究主題，結果顯示這個方法可以同時抽取出中文和英文的關鍵詞語，所得到的詞語叢集結果也可以表示領域中重要的研究主題。這樣的結果初步驗證了本論文所提出方法的可行性。從研究結果中，我們也發現計算語言學研究與實務應用有密切的關係，抽取出來的詞語叢集中有許多與機器翻譯、語音處理和資訊檢索相關，在語言的計算模式上，語法模式與剖析、斷詞和語言

模型則是國內計算語言學家所關心的主題。

在後續的研究上，除了進一步改善目前所提出來的方法，並且深入探討各研究主題的起源、發展與演變之外，我們將探索各個研究主題之間的相關性，並嘗試將結果以圖形化的方式加以呈現。另外，對於不同學術領域間的相關研究主題的發掘和分析，比方說資訊檢索同樣是圖書資訊學所關心的研究主題，如何利用自然語言處理技術來分析兩個領域間的共通與相異，是一項值得探討的研究。

致謝

本研究受國科會「國內計算語言學學術資訊交流之研究(I)」(編號NSC91-2413-H-128-004-)計畫案補助。另外，也感謝三位審查委員所提供的意見。

參考文獻

[Bishop, 1999] A. P. Bishop, “Document Structure and Digital Libraries: How Researchers Mobilize Information in Journal Articles”, Information Processing and Management, 35, p255-279.

[Chien, 1997] Lee-Feng Chien, “PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval”, SIGIR’97, p50-58.

[Chien, et. al., 1999] Lee-Feng Chien, Chun-Liang Chen, Wen-Hsiang Lu, and Yuan-Lu Chang, “Recent Results on Domain-Specific Term Extraction From Online Chinese Text Resources”, ROCLING XII, p203-218.

[Church and Mercer, 1993] K. W. Church and R. L. Mercer, “Introduction to the Special Issue on Computational Linguistics Using Large Corpora”, Computational Linguistics, 19(1), p1-24.

[Covi, 1999] L. M. Covi, “Material Mastery: Situating Digital Library Use in University Research Practices”, Information Processing and Management, 35, p293-316.

[Deerwester, et. al., 1990] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis”, Journal of the American Society for Information Science, 41(6), p391-407.

[Harter, 1992] S. P. Harter, “Psychological Relevance and Information Science”, Journal of the American Society for Information Science, 43(9), p602-615.

[Hatzivassiloglou, Gravano and Maganti, 2000] V. Hatzivassiloglou, L. Gravano and A. Maganti, “An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering”, SIGIR’2000, p224-231.

[Hjørland and Albrechtsen, 1995] B. Hjørland and H. Albrechtsen, “Towards a New Horizon in Information Science: Domain-Analysis”, Journal of the American Society for Information Science, 46(6), p400-425.

[Huang, 2000] Chu-Ren Huang, “From Quantitative to Qualitative Studies: Developments in Chinese Computational and Corpus Linguistics”, 漢學研究, 第十八卷特刊, p473-509.

[Kowalski and Maybury, 2000] G. J. Kowalski and M. T. Maybury, “Document and Term Clustering”, Information Storage and Retrieval Systems: Theory and Implementation, 2nd ed., Chapter 6, p139-163.

[Lenders, 2001] W. Lenders, “Past and Future Goals of Computational Linguistics”, ROCLING XIV, p213-236.

[Tabah, 1996] A. N. Tabah, Information Epidemics and the Growth of Physics, Ph. D. Dissertation of McGill University, Canada.

[Wayne, 2000] C. L. Wayne, “Topic Detection and Tracking in English and Chinese”, IRAL 5, p165-172.

[Wilson, 1999] T. D. Wilson, “Models in Information Behaviour Research”, Journal of Documentation, 55(3), p249-270.

[Yang, Pierce and Carbonell, 1998] Y. Yang, T. Pierce, and J. Carbonell, “A Study on Retrospective and On-Line Event Detection”, SIGIR’98, p28-36.

[王士元, 1988] “電腦在語言學裡的運用”, ROCLING I, p257-287.

附錄一 ROCLING 研討會論文資料所得到的詞語叢集

叢集編號	詞語
1	generation, generator, systemic, text generation
2	acquisition, explanation, generalization, learning
3	方法, 系統, 問題, 處理
4	initial, min, taiwanese, 台語, 台灣, 資料庫

叢集編號	詞語
5	atn, attachment, pp, preference
6	complexity, computational, gpsg, morphology
7	hidden markov, maximum, robust speech recognition, speech recognition
8	aspect, logic, temporal, tense
9	document, hierarchical, text categorization, 分類, 文件, 文件分類, 特徵
10	classifiers, decision, non, symbols
11	分析, 系統, 處理, 語言
12	bilingual, machine translation, mt, transfer, 機器翻譯
13	dictation, large vocabulary, 語言模型, 語音辨認
14	adaptation, maximum, robust speech recognition, 語音辨識
15	attachment, pp, preference, score
16	系統, 設計, 輸入, 鍵盤
17	國語, 語言模型, 語音辨認, 辨認
18	chinese text, chinese word segmentation, segmentation, unknown word, word identification, word segmentation, words, 斷詞
19	attention, conversation, discourse, elicitation, interaction
20	continuous, hidden markov, maximum, speech recognition
21	統計, 詞彙, 語言, 語料
22	bigram, class based, clustering, entropy, language model, language modeling, language models, n gram
23	分析, 表達, 剖析, 格位, 訊息, 動詞, 結構, 詞類, 漢語, 語法, 語法模式, 語意, 模式, 關係
24	adaptive, compression, scheme, 英文, 資料, 調整, 壓縮
25	csmart, databases, document, indexing, information retrieval, retrieval, text retrieval, 檢索
26	grammars, parser, parsing, sentence
27	continuous, large vocabulary, mandarin, speaker, speech, speech recognition, telephone
28	document, text categorization, 分類, 文件分類, 文件自動, 關鍵詞
29	方法, 系統, 設計, 應用
30	speech, synthesis, 文句翻語音, 合成, 系統, 音節, 國語, 連音, 語音, 輸入
31	mandarin text to speech, pitch, prosodic, speech, synthesis, 文句翻語音, 合成
32	bilingual, machine translation, translation, 機器翻譯
33	explanation, generalization, learning, parse
34	aspect, functional, lexical, lexical semantic, mandarin chinese, meaning, parsing, phrase, roles, semantic, semantics, syntactic, syntax, thematic, theory, verb, verbal, verbs

以知識概念模型為基礎之多主題對話管理系統
Ontology-Based Dialog Management for Multiple Service Integration

陳銘軍 葉瑞峰 吳宗憲

國立成功大學資訊工程學系

rusk@csie.ncku.edu.tw, p7890101@ccmail.ncku.edu.tw, chwu@csie.ncku.edu.tw

摘要

目前的對話系統，多侷限於單一功能或單一領域的應用，而在實際情況，使用者經常需要跨多個服務或多個主題。本文提出一快速且有效整合現有對話系統之方法。為了實現此一研究主題，必須先半自動地建立一個醫療概念模型作為對話系統的知識表示法，建立及整合三個不同但具有相關性之服務模組：分別為掛號諮詢模組、科別諮詢模組以及常見問答集(FAQ)模組。藉由對話管理模組依其意圖加以整合。在多主題或多服務的整合上，提出以部分樣本樹做為意圖偵測之評估方式，配合語意框架記錄並控制對話流程，並利用樣本產生系統之相對回應文句。

為了評估本文所提出的方法之可行性，由 50 個大專以上教育程度，但未參與本研究之使用者，以文字輸入的模式實際測試本系統。在效能評估上，其整體意圖偵測正確率為 86.2%、系統成功率為 77%，每筆對話的平均長度為 9.2 回合，而回答自然度則為 78.5%，足見本文所提之方法是有效可行的。

1. 緒論

語音及語言處理技術的日趨成熟，使得對話系統的實現成為可能[1][2][3][4]。對話系統的研究，在國外方面有 MIT 的 JUPITER [5] 、AT&T 的線上服務系統[6]、Philips 的 Automatic Train Timetable Information System[7]、英國的 Nuance Automatic Banking System 以及法國的 LIMSI Arise system 旅遊資訊導覽系統[8]等。在國內方面，台大則提出在語意與知識之擷取[9]以及在分散式網路環境中對話系統代理人之架構[10]，工研院有智慧型總機、氣象查詢系統[11]等多項成果，成大在對話系統的研究上也投入相當多之研究人力[12][13][14][15]。

就對話系統言，其中一個重要課題便是如何理解人類語言，讓電腦能夠接受人類的語言和指令，直至今日仍是一個無法完全解決的問題。知識概念模型定義

除了定義了概念本身之語義定義外，同時也定義了概念與概念間的關係，並存在有推論規則可供邏輯推理，是一種具有推論能力的知識表示法[16][17]。而對話系統目前的研究大多侷限在單一功能或單一領域的應用。然而使用者之意圖，可能會牽涉到多個領域或多個服務，因此如何整合多套對話系統，進而提供更完整的資訊，是目前許多研究的主題[18] 。本文提出利用意圖偵測機制，整合多種服務，此一作法不僅在整合既有系統提供解決方案，更在未來新增功能預留空間。

為驗證所提方法之正確性，我們建立一套智慧型的醫療對話查詢系統，提供三個功能模組，分別為掛號資訊諮詢模組、科別資訊諮詢模組及常見問答集資訊諮詢模組。其中

(1) 掛號諮詢模組：乃利用自然語言處理的技術，使用者除了從網路超連結中以點選的方式進行網路掛號外，更可以符合使用者習慣之口語方式輸入。

(2) 科別諮詢模組：此一模組利用知識概念模型中的規則來進行推論、藉以解決使用者在掛號時，可能做症狀描述但不知科別之窘境。

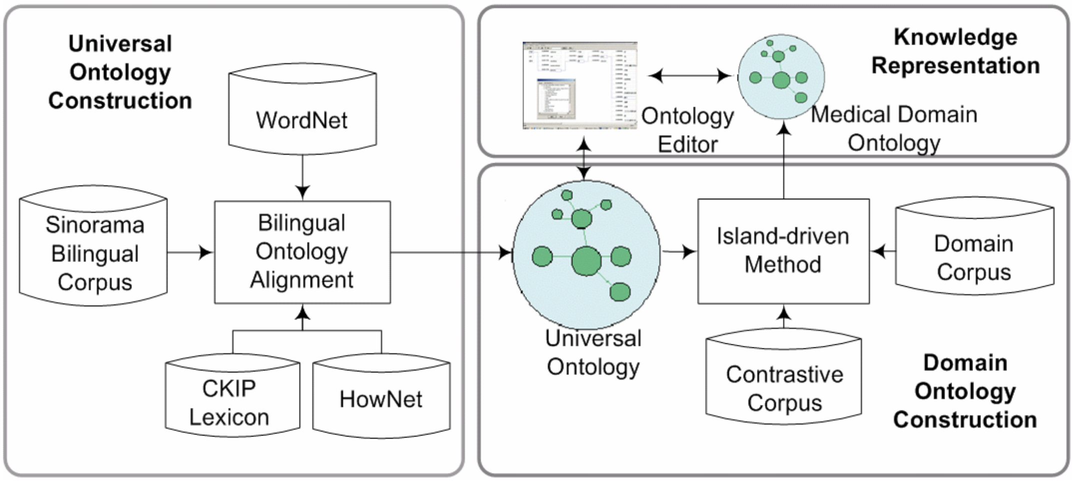
(3) FAQ 諮詢模組：相關文獻的查詢，從對話過程中累積的資訊分析得到，與以單一文句查詢之 QA 系統不同。

知識表示法乃採用 WordNet 的結構為基礎，將 HowNet 的資訊整合進去，首先使用雙語語料為統計對象，基於雙語字典整合 HowNet 和 WordNet 來建立出 Universal Ontology，再利用醫療類語料和島嶼演算法[19]半自動萃取出屬於 Medical Ontology 之概念，並加入 1213 則推論規則(axioms)，形成醫療概念模型，此醫療概念模型(Medical Ontology)即為系統底層的知識庫。

在多項服務的整合上，本文提出利用部分樣本樹 PPT(Partial Pattern Tree)[20]來擷取使用者的意圖(Intention)，再由使用者的意圖搭配相關之語意框架(Semantic Frame)來控制並完成整個對話流程。

2. 知識概念模型之建構

本文提出一個兩階段建構知識概念模型之方法，如下圖一所示：



圖一、知識概念模型建構流程圖

以 WordNet 的階層架構為主，將 HowNet 整合進來，其中所用之雙語語料為光華智慧藏部分語料。建構出通用知識概念模型（Universal Ontology），再利用島嶼演算法將屬於領域特性的知識概念模型擷取出來成為領域知識概念模型（Domain Ontology）。

在知識概念模型的整合上採取由下而上(Bottom-up algorithm)的策略。詞與詞或概念與概念間的上下位關係為一樹狀結構。節點之相似程度乃由終端節點開始向上傳遞累積，若子節點相似且節點本身相似程度高，則此兩個節點所代表之意義應屬於同一概念。而目前系統中的關聯整合乃建構於字網的關聯架構上，將知網的關聯分析後將之對應至字網對應的依據乃是前面所得的中、英文概念詞之對應關係加權計分，和他們彼此之間的結構上之相似度 c 如下式(1)所示：

$$\begin{aligned}&\Pr(synset^k \mid CW_i)\\&=\sum_{j=1}^{m}\Pr(synset^k, EW_j^k \mid CW_i)\qquad (1)\\&=\sum_{j=1}^{m}(\Pr(synset^k \mid EW_j^k, CW_i)\times\Pr(EW_j^k \mid CW_i))\end{aligned}$$

其中

$$\Pr\left(synset^k \mid EW_j^k, CW_i\right) = \frac{N\left(synset_j^k, EW_j^k, CW_i\right)}{\sum_l N\left(synset_j^l, EW_j^k, CW_i\right)} \tag{2}$$

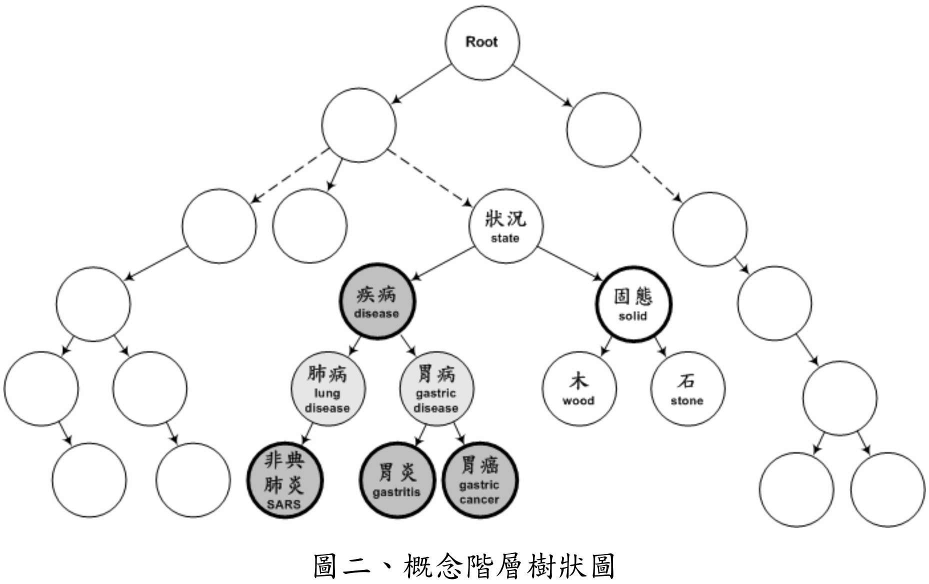
如上式所示，$N\left(synset_j^k, EW_j^k, CW_i\right)$為$CW_i$、$EW_j^k$和$synset_j^k$共同出現的次數。在定義於 HowNet 中之中文詞，至少有一主要特徵$PF_i^l\left(CW_i\right)$和定義於 WordNet 中英文詞之上位詞中存在有一個同義詞集合，$synset_j^k(EW_j)$，有一致之概念時，其機率$\Pr\left(EW_j \mid CW_i\right)$ 為 1，否則為 0。

$$\Pr\left(EW_j \mid CW_i\right) = \begin{cases} 1 & if \quad \left(\bigcup_l PF_i^l\left(CW_i\right)\right) \cap \left(\bigcup ancestor\left(\bigcup_k synset_j^k(EW_j)\right)\right) \neq \varnothing \\ 0 & otherwise \end{cases} \tag{3}$$

最後中文概念詞CW_i將依其機率值$\Pr(synset^k \mid CW_i)$被整合至英文同義詞集合$synset_j^k$中，成為其中的一個元素。如此一來便完成了通用知識概念模型的建構。

接下來我們便從此通用知識概念模型中擷取出，屬於醫療領域之領域概念模型，其步驟有四：

(1) 階層線性化：知識概念模型中的階層是一樹狀的結構，如圖二所示。線性化的目的，即是將樹狀結構分解為節點之有序串列(由終端節點到根節點的路徑)所成的集合。



圖二、概念階層樹狀圖

(2) 以語料為基礎之概念抽取：語料分為目標領域語料(Target Domain Corpus)與對比領域語料(Contrastive Corpus)兩類。其中領域語料是從網路上蒐集得到的一千二百二十二篇中文醫療類 FAQ，而對比領域語料是由光華雜誌智慧藏當中選取二千一百八十篇非醫療類文章構成。並依 Tf-Idf 決定節點是否屬於目標領域，若屬於目標領域則稱此一節點為有效節點。

$$operative\_node(W_i) = \begin{cases} 1, & if\ Tf-idf_{Domain}(W_i) > Tf-idf_{Contrastive}(W_i) \\ 0, & Otherwise \end{cases} \quad (4)$$

其中

$$Tf-idf_{Domain}(W_i) = freq_{i,Domain} \times \log \frac{N}{n_{i,Domain}}$$

$$Tf-idf_{Contrastive}(W_i) = freq_{i,Contrastive} \times \log \frac{N}{n_{i,Contrastive}}$$

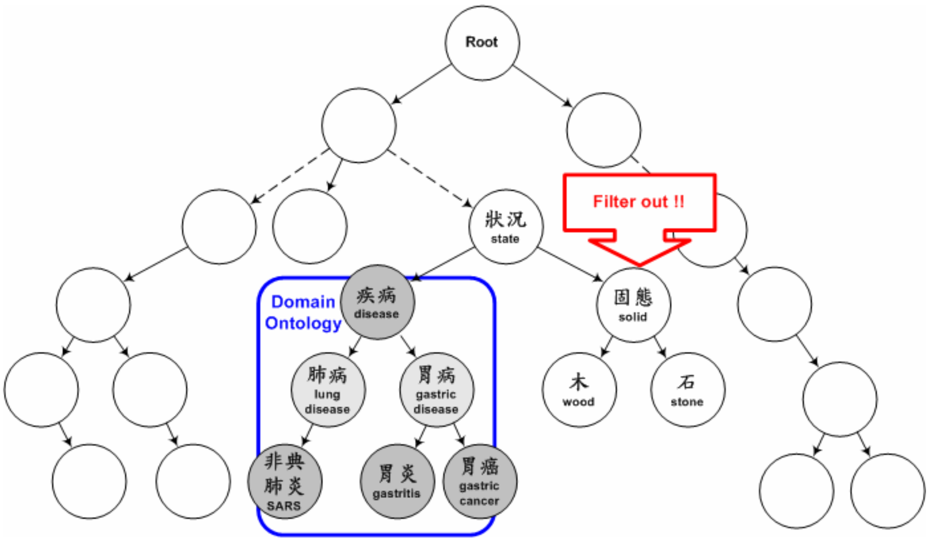
$$N = n_{i,Domain} + n_{i,Contrastive}$$

其中 W_i 為待測概念詞，$freq_{i,Domain}$ 和 $freq_{i,Contrastive}$ 分別代表 W_i 出現在目標領域語料和對比領域語料中的頻率。$n_{i,Domain}$ 和 $n_{i,Contrastive}$ 則為目標領域語料和對比領域語料中含有 W_i 之文章數。

(3) 島嶼演算法：由於統計語料之不足，所以利用知識概念模型中的上下位關係，將一些未出現在統計語料，但屬於該領域的概念也一併萃取出來。事實上，對於一個概念而言，若是其上位關係與下位關係皆是屬於某一領域，則該概念同屬此一概念的機率是很大的。基於這樣的假設，利用島嶼演算法[19]來將潛藏於兩個領域概念之間的概念一併收錄。

(4) 合併線性成分且過濾獨立概念：經過島嶼演算法之後，可以得到一個具擴充的節點串列，原則上只要保這些節點串列合併起來，便可以建構出領域知識概念模型的雛型。但對於一些具有多個有效節點的節點串列，將予以保留並加以合併。而對於僅含有一個有效節點的節點串列將予以刪除，其原因是這些單獨存在的有效節點，多為停用詞(Stop word)或其他自然語言處理程序中所產生的干擾，如斷詞錯誤等，完成後如圖三所示。

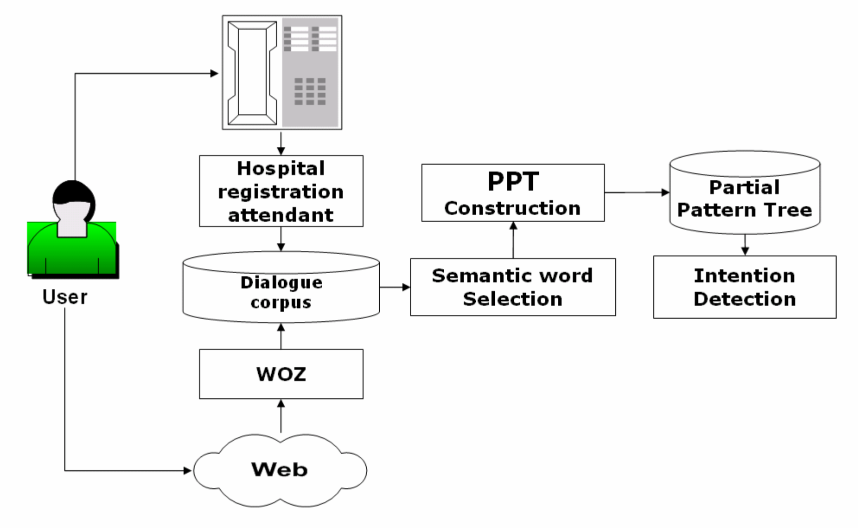
建立醫療概念模型後，依疾病、症狀及所對應的科別，加入了 1213 條規則。



圖三、領域知識概念之取得

3. 部分樣本樹之建立和意圖偵測

本文提出利用意圖來整合多項服務的方法，在對話系統的發展中經常面臨語料不足的問題，為了克服語料不足以及增強系統之包容性(Robustness)，我們使用部分樣本樹來做意圖偵測，其建構流程如圖四所示：



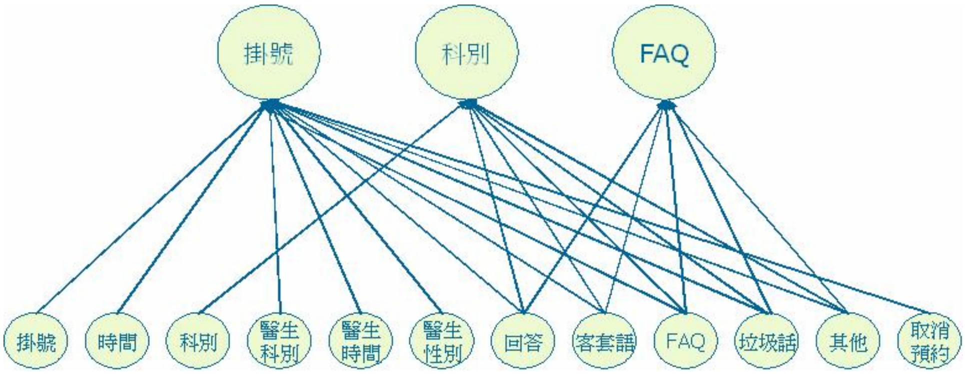
圖四、部分樣本樹和意圖偵測架構圖

以統計方法建構一套實際的對話系統時，面臨的第一個問題就是對話語料的收集，語料收集之良窳直接影響整個對話系統的建構。我們採用兩種方式收集語料，分別為 woz 以及成大醫院實際語料。

(1) Wizard-of-Oz 方式之語料蒐集：本論文使用 ASP 和 IIS5.0 版來實做 WOZ 之語料收集平台，透過此一系統，本論文初步收集了三十四位使用者的對話共計 234(turns)之語料，用以建立系統之雛型。

(2)**成大醫院實際掛號語料：**在實際語料收集部分，在 88 年一月間於成大醫院錄製之實際電話預約語音資料，再以打字的方式轉成文字檔，共有十萬多字，四千零八九筆對話資料。

在收集完語料之後，由 WOZ 所收集的語料和電話語料中選擇出 1395 句來做人工標記其意圖，做為建立部分樣本樹之訓練語料，依據系統所整合之服務，並觀察語料可歸納如下之意圖示意圖，意圖共分為 12 類，如圖五所示：



圖五、意圖示意圖

為了整合各種不同而相關的服務，必須從各種服務間找出其差異性，在此我們使用意圖(Intension)來區分服務之機制，就可從標記為各意圖的語料裡找出一組主要語意詞來代表這個意圖，簡單地說，主要語意詞即為具有鑑別性之語意概念，例如使用者想查詢有關醫生看診時間，就會提到醫生的姓名，醫生姓名便可以用來代表醫師姓名這項意圖。本文採用 LSA(Latent Semantic Analysis)的方法，其原理乃利用奇異值分解，除了對於各意圖選擇最具鑑別性的詞，作為各意圖的主要語意詞(Semantic words)。

語言的多樣性經常導致收集樣本之不充分，如使用者想要掛號可能出現的語

句可能是「我要掛號」,「我想要掛號」,「我要預約掛號」..等等，但從語料的觀察，可以發現意圖通常與主要語意詞(Semantic word)有極高之共現率，其他功能性詞彙則有可能被省略，因此本文使用部分樣本據來建立意圖辨識模組，在這裡本論文將句子是為一連串的功能性詞彙和主要語意詞的組合，可表示如下：

$$S_i = \left\{FP_1^i, FP_2^i, \cdots, FP_{NB_i}^i, SP^i, FP_{NB_{i+1}}^i, \cdots, FP_{NB_i+NA_i}\right\} \tag{5}$$

在式子(5)中 SP^i 代表主要語意詞，而 FP_j^i 代表第 j 各功能性詞彙 NBi 和 NAi 為在主要語意詞前和在主要語意詞之後的功能性詞彙數。根據上述定義，部分樣本句為包含主要語意詞 SP^i 的子序列，所以最常的部分樣本句即為句子本身，而最短的部分樣本句則只有主要語意詞一個詞彙，而每一個功能性詞彙都有可能被省略，所以對式子(5)定義的句子共有 $2^{NA_i+NB_i}$ 句部分樣本句。舉例說明，若有一句子為〝ABC〞且 A，C 為功能性詞彙而 B 為主要語意辭彙，則有四句部份樣本句〝ABC〞,〝AB〞,〝BC〞以及〝B〞。

部分樣本樹是利用語料庫中的句子，將其分解成部分樣本句後，所建立的模組，它有兩點特點，第一點為具自動學習之能力，由訓練語料所得之的部分樣本句文法，第二點為可處理贅詞及部分詞彙錯誤的情形。因此根據收集到的訓練語料首先將訓練的句子分解成所有的部分樣本句，然後以樹狀結構將所有的部分樣本句的句型資訊儲存起來即為部分樣本樹。

在實際建立部分樣本樹的過程中，每一個內部節點代表部分樣本樹上的一個獨立詞彙，因此對於每一個內部節點本論文可以表示成

$$IN_i = \left\{PH_i, FR_i, Ns_i, Son_i\right\}$$

包含的參數描述如下：

PH_i:此節點在部分樣本句上所代表的辭彙

FR_i:此節點在訓練語料出現的頻率

Ns_i:其下所接的內部節點個數

Son_i:記錄所有到子節點的連結

在部分樣本樹中，外部節點代表著依據部分樣本句的結尾，因此可以利用外部節點很容易的回溯找到其所代表的部分樣本句. 所以在此將外部節點表示為

$$EN_i = \left\{PP_i, Ptr_i, IT_i\right\}$$

其中 PP_i :代表此外部節點所表示的部分樣本句

Ptr_i:紀錄此部分樣本句是從哪些較完整的部分樣本句中因部分功能性詞彙被省略而來

IT_i:記錄這條路徑所代表的意圖

整個訓練過程，總共有三個步驟來建立部分樣本樹：

(1) 將訓練的句子斷詞成為一連串的詞彙序列，對於每一具訓練語料中的主要語意詞在這個步驟都將其標記唯一特殊詞彙「Semantic word」，也就是說在訓練過程中，將所有主要語意詞都看成同一個詞彙。

(2) 將斷好詞的句子拆解成部分樣本句。

(3) 利用接下來介紹的演算法來建立部分樣本句。

部分樣本樹建立演算法：

步驟一: Initialization

設定部分樣本樹的根節點，R

步驟二： Recursion

對所有的部分樣本句，$PP_i = \{Ph_1^i Ph_2^i ... Ph_{N_i}^i\}$，$N_i$ 為部分樣本句 PP_i 的詞彙個數，執行步驟 2.1 到步驟 2.5

步驟 2.1: 根據部分樣本句的詞彙順序，由根節點 R 搜尋已建利立的部分樣本樹，最後停在節點 IN_s ，也就是說由根節點 R 到停止節點 IN_s ，這條路徑符合 PP_i 的 prefix

步驟 2.2: 若此部分樣本句上所有的詞彙都已被搜尋到了，也就是此部分樣本句已經存在部分樣本樹中了，跳到步驟 2.4

步驟 2.3: 對於 PP_i 上還沒被搜尋到的詞彙，Ph_k^i

步驟 2.3.1：建立新的內部節點，IN_N

步驟 2.3.2：設定此 IN_N 內所代表的詞彙為 Ph_k^i

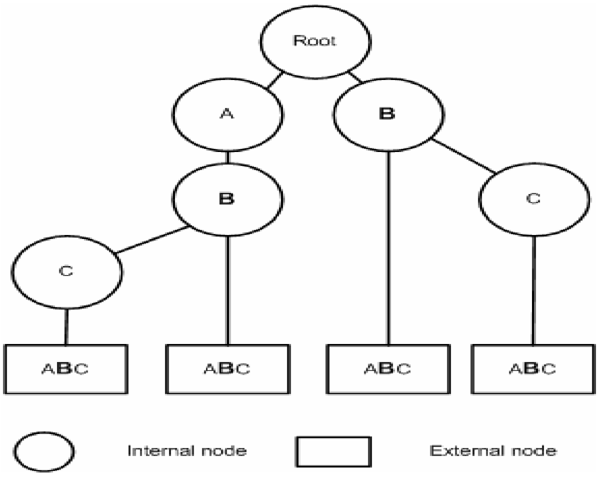
步驟 2.3.3：將 IN_N 加入 IN_s 的 Son 指標陣列中，並增加 IN_s 的 NS_s 個數

步驟 2.3.4：將 IN_s 設為 IN_N

步驟 2.4：若是此部分樣本句的外部節點不存在，則建立新的外部節點，

NE_N 並且設定 NE_N 所代表的部分樣本句為 PP_i

步驟 2.5：將在這條路徑上每一個內部節點 FP 參數加 1



圖六、建立好的部分樣本樹

圖六為訓練語句〝ABC〞所建立的部分樣本樹，其中〝A〞、〝C〞為功能性詞彙，〝B〞為主要語意詞，對上述部分樣本樹建構演算法說明，假設只考慮這句訓練語句〝ABC〞，首先對此句語句作前處理，先將句子〝ABC〞斷詞，並將主要語意詞標記為〝Semantic word〞代表這句子的語意，而在圖中粗體字 B 代表〝Semantic word〞的到結果「A，Semantic word，C」，再將斷好詞的句子拆解成所有的部分樣本句：「A，Semantic word，C」,「A，Semantic word」,「Semantic word，C」以及「Semantic word」，此時開始使用上述演算法建構部分樣本樹，第一步驟先設立好根節點 R 之後，步驟二對所有的部分樣本句一一將其加入部分樣本樹中，例如一開始先加入「A，Semantic word，C」這句，本論文利用 prefix 搜尋法，搜尋已建立好的部分樣本樹，因為一開使沒有任何路徑存在，所以搜尋結果停留在根節點 R，此時依據未被搜尋到的詞彙順序，依序為「A」,「Semantic word」以及「C」，對每一個詞彙新增對應的內部節點，並且設定好節點間的父子關係，以及每一個內部節點的資訊，接下來因為代表此部分樣本句的外部節點尚未存在，所以本論文建立一個新的外部節點，設定其代表部分樣本句為「A，Semantic word，C」，就將第一句部分樣本句「A，Semantic word，C」加入了，而接下來當加入「A，Semantic word」這句部分樣本句時，會發現此路徑已經存

在部分樣本樹中了，因此不會增加新的內部節點，但是對每一個被搜尋過的內部節點其 FP 參數都加一，並且此時根據搜尋演算法，會停留在剛剛建立好代表詞彙為 B 的內部節點上，並且會發現其子節點並沒有外部節點，因此新增一個對應到「A，Semantic word」的外部節點，依此類推，將所有的部分樣本句加入後，就完成了步驟二，也就完成的部分樣本樹。

完成部分樣本樹的建構後，再利用部分樣本樹來偵測出意圖，在此只需對使用者的輸入與部分樣本樹中的路徑(Path)，找尋最相近似的路徑，就可以知道這句話的意圖。本論文將考慮兩個句子間的文法結構和語意相似度後，以動態規劃的方式來算出兩個句子的最佳相似分數。

假設P_I是使用者輸入的語句，$P_I = \{a_1, a_2, \ldots, a_i\}$，$P_T$是部分樣本樹的其中一條路徑可以被表示成$P_T = \{b_1, b_2, \ldots, b_j\}$，接著計算兩個句子間的結構相似度，所謂的結構相似度，就是指兩個句子之間的詞序是否相同，此定義如式子(6)

$$sim_{syn}(a_i, b_j) = \begin{cases} 0 & \text{if } a_i \neq b_j \\ 1 & \text{if } a_i = b_j \end{cases} \tag{6}$$

當在考量兩個句子之間的相似度時，除了從文法結構上來分析外，語意亦佔有重要的地位，在語意上來分析便要考慮到詞與詞之間的相關性，因此我們採用之前所建立的醫療概念模型來計算詞與詞之間的相似度。測試句與樣本樹的一條路徑中的一個詞分別為a_i跟b_j透過醫療概念模型來計算他們的相似度，在相似度的定義如式子(7)

$$sim_{sem}(a_i, b_j) = \begin{cases} 1 & if\ a_i = b_j \\ \left(\frac{1}{2}\right)^l & if\ a_i\ and\ b_j \text{ are hypernyms} \\ 1-\left(\frac{1}{2}\right)^n & if\ a_i\ and\ b_j \text{ are synomyms} \\ 0 & others \end{cases} \tag{7}$$

在式(7)中l是指兩個具有上下位關係之概念在概念模型中距離多少層，n 是兩個同義詞間將其所有同意詞展開後，有多少個共同之同義詞。

在考慮到文法結構和語意的相似度後，將這兩項因素列入考慮來計算使用者的輸入句和部分樣本句之間的相似度。採用類似動態規劃的方法如式子(8)

$$sim_{Int}(0,0) = 0$$

$$sim_{Int}(i,j) = \max \begin{cases} sim_{Int}(i-1,j-1) + \left(sim_{sem}(a_{i-1},b_{j-1}) + sim_{syn}(a_{i-1},b_{j-1})\right) \\ sim_{Int}(i-1,j)) + \left(sim_{sem}(a_{i-1},b_{j}) + sim_{syn}(a_{i-1},b_{j})\right) \\ sim_{Int}(i,j-1) + \left(sim_{sem}(a_{i},b_{j-1}) + sim_{syn}(a_{i},b_{j-1})\right) \end{cases} \qquad (8)$$

$$sim_{Int}(P_I,P_J) = sim_{Int}(I,J)$$

在式(8)中 $P_I = \{a_1, a_2, \ldots, a_i\}$代表輸入的句子，$P_t$是部分樣本樹的其中一條路徑可以被表示成 $P_T = \{b_1, b_2, \ldots, b_j\}$，$sim_I(P_I,P_T)$代表它們之間的相似度，一開始先將其間的相似度設定為零，然後從兩個句子的第一個詞開始遞迴找尋最大相似度，$sim_{sem}(a_i,b_j)$是語意相似度，$sim_{syn}(a_i,b_j)$則是句法結構相似度。

以實例來說，系統輸入句為「我有點感冒」而樣本句有「我、有、發燒」，輸入句經過斷詞後，會得到「我、有、點、感冒」，首先第一個詞「我」對到「我」，相似度加二，第二個詞「有」對到「有」 相似度在加二變成四，第三個詞輸入句是「點」，和樣本句沒有相同的詞，因此本論文考慮「有」和「發燒」兩個詞，「點」和「有」的相似度為零，和「發燒」的相似度也為零，因此對到哪一個詞分數都一樣，因此本論文把「點」對到「有」，然後把相似度不變為四，再考慮下一個詞「感冒」，同樣的「感冒」和樣本句也沒有相同的詞，因此本論文考慮「有」和「發燒」兩個詞，「感冒」和「有」的相似度為零，和「發燒」的相似度為 1/2 因為「發燒」是「感冒」的下位詞它們之間相差一層，所以本論文就把「感冒」對到「發燒」，相似度加 1/2 變成 4 又 1/2 這就是此輸入和這句樣本句的相似度了，依此方法比對每句樣本句，找出最相近的樣本句後，就可以知道使用者這句話的意圖。

4. 對話控制模組

本節說明如何整合各項服務，及個別模組之功能說明，在前面已經解釋如何從使用者的輸入擷取出使用者的意圖，在有了使用者的意圖之後，系統便可偵測使用者是要使用那樣模組，再配合各項模組內的 Semantic Frame 產生相關的對應的回應。

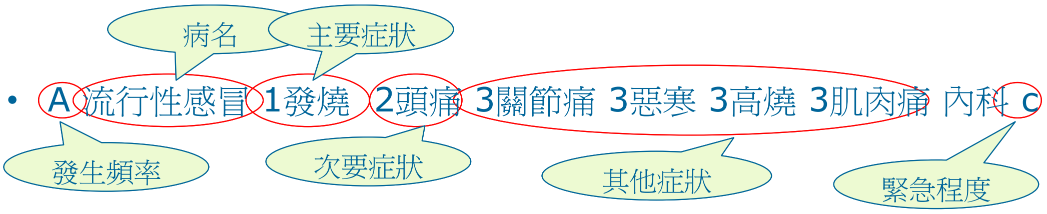
在各個模組使用相對應之語意框架(Semantic Frame)相配合之來控制對話流程[23]，以實例來說明，在掛號諮詢模組中的語意框架如圖七所示：

```
{Register:}-
        [Clinics]->()
        [Month]->()
        [date]->()
        [doctor’s name]->()
        [doctor’s professional skill]->()
```

圖七、 語意框架

(1)掛號諮詢模組，主要功能是讓為協助使用者完成線上掛號，因此需要一個可提供查詢的線上資料庫，本文是以成大醫院資訊做為系統的線上掛號資料庫，提供掛號所需的各項資訊。

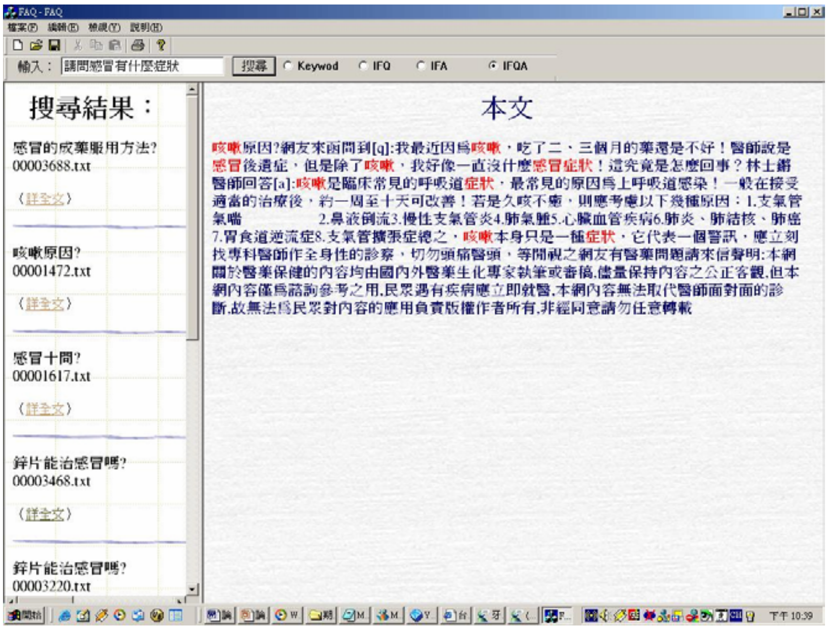
(2)此模組提供科別諮詢功能，所謂的科別諮詢功能即是引導使用者來掛號適當的科別，有病痛時，並不是每次都知道該去找尋什麼樣子的醫生或掛什麼科別，本模組提供之前建立好的推論原則，來找出最適合的科別，來節省使用者的時間和醫療資源。


病名
主要症狀
• A 流行性感冒 1發燒 2頭痛 3關節痛 3惡寒 3高燒 3肌肉痛 內科 c
發生頻率
次要症狀
其他症狀
緊急程度

圖八、推論規則詳細說明

疾病發生頻率本論文總共分為 A，B，C 三個等級分別代表常發生、可能發生以及少發生，主要症狀只有一個代表此疾病最主要之症狀，次要症狀則是伴隨疾病發生的機會次之，其他症狀則代表可能但隨個人體質不同而伴隨的症狀，看診科別則是此疾病應當看哪一種科別，最後一個欄位的疾病緊急程度分為 a，b，c 三個等級，a 代表很緊急應當急診，b 代表緊急應該趕緊就醫，c 則代表普通。

(3)FAQ 諮詢模組：本論文也提供相關醫療資訊給使用者查詢，語料是從網路上收集下來，其系統界面如圖九所示:



輸入：請問感冒有什麼症狀

搜尋結果：

感冒的成藥服用方法?
00003688.txt
〈詳全文〉

咳嗽原因?
00001472.txt
〈詳全文〉

感冒十問?
00001617.txt
〈詳全文〉

鋅片能治感冒嗎?
00003468.txt
〈詳全文〉

鋅片能治感冒嗎?
00003220.txt

本文

咳嗽原因?網友來函問到[q]:我最近因爲咳嗽，吃了二、三個月的藥還是不好！醫師說是感冒後遺症，但是除了咳嗽，我好像一直沒什麼感冒症狀！這究竟是怎麼回事？林士爵醫師回答[a]:咳嗽是臨床常見的呼吸道症狀，最常見的原因爲上呼吸道感染！一般在接受適當的治療後，約一周至十天可改善！若是久咳不癒，則應考慮以下幾種原因：1.支氣管氣喘 2.鼻液倒流3.慢性支氣管炎4.肺氣腫5.心臟血管疾病6.肺炎、肺結核、肺癌7.胃食道逆流症8.支氣管擴張症總之，咳嗽本身只是一種症狀，它代表一個警訊，應立刻找專科醫師作全身性的診察，切勿頭痛醫頭，等閒視之網友有醫藥問題請來信聲明:本網關於醫藥保健的內容均由國內外醫藥生化專家執筆或審稿.儘量保持內容之公正客觀.但本網內容僅爲諮詢參考之用.民眾遇有疾病應立即就醫.本網內容無法取代醫師面對面的診斷.故無法爲民眾對內容的應用負責版權作者所有.非經同意請勿任意轉載

圖九、 FAQ 系統介面圖

5. 實驗與討論

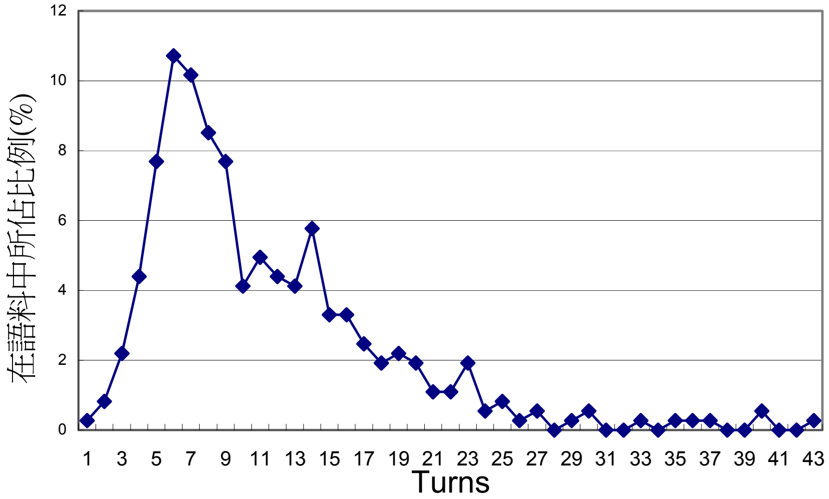
實驗所使用的機器為 Pentium Ⅳ 2G 的個人電腦 512MB RAM，開發的工具是 Microsoft Visual C++ 6.0，ASP 和 IIS5.0 版，在 Windows2000 的作業系統下進行開發與實驗 。

5.1 對話語料分析

在本節將對所收集的語料進行分析，本論文的語料共分為三類，一是實際對話語料，二是 WOZ 收集來的語料，三是系統實際測試時所收集回來的語料，首先分別對各個語料對使用者一次對話的長度作分析，即是一次對話需要來回多少次(turns)，其結果如下：

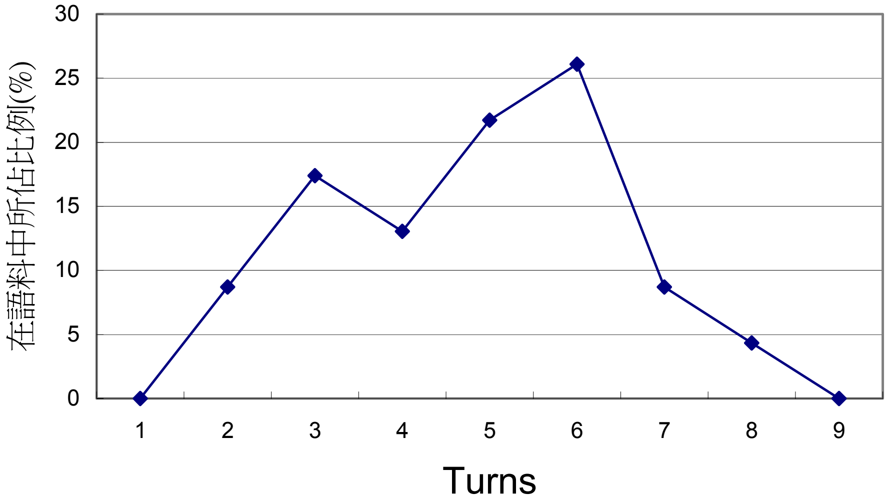
5.1.1. 電話語料

圖十中橫軸是一次對話的對話長度，縱軸是長度所佔語料的比例，在電話語料部分共收集了 4089 比對話資料(turns)、為 364 人次累計出來的，每人平均的對話次數為 11.235 次，其最長的一回對話為 47 回合，從圖十可以看出來對話長度約在 6~14 之間，平均對話長度拉長的原因是有不少特別長的對話。



圖十、電話語料長度分佈

5.1.2. WOZ 語料

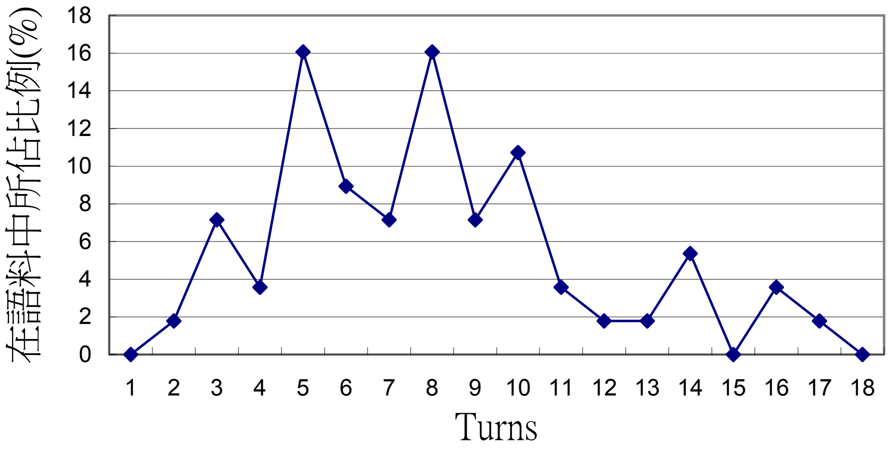


圖十一、 WOZ 語料長度分佈

在 WOZ 語料部分共收集了 234 比對話資料(turns)、為 34 人次累計出來的，每人平均的對話次數為 6.882 次，其最長的一回對話為 10 回合，從圖十一可以看出來大多對話次數約在 6~8 之間。

5.1.3. 實際系統語料

在系統測試語料部分共收集了 500 筆對話資料(turns)、為 56 人次累計出來的，每人平均的對話次數為 8.928 次，其最長的一回對話為 16 回合，從圖十二可以看出來大多對話次數約在 5~8 之間，但與圖十與圖十一不同的是，圖十二有兩個高峰點分別在對話長度為 5 和 8 時，測試語料會有這樣狀況的原因是系統提供三項服務，因此使用者使用不同服務時對話的長度也就不一樣，其功能與對話長度的關係如表一所示：



圖十二、系統測試語料長度分佈

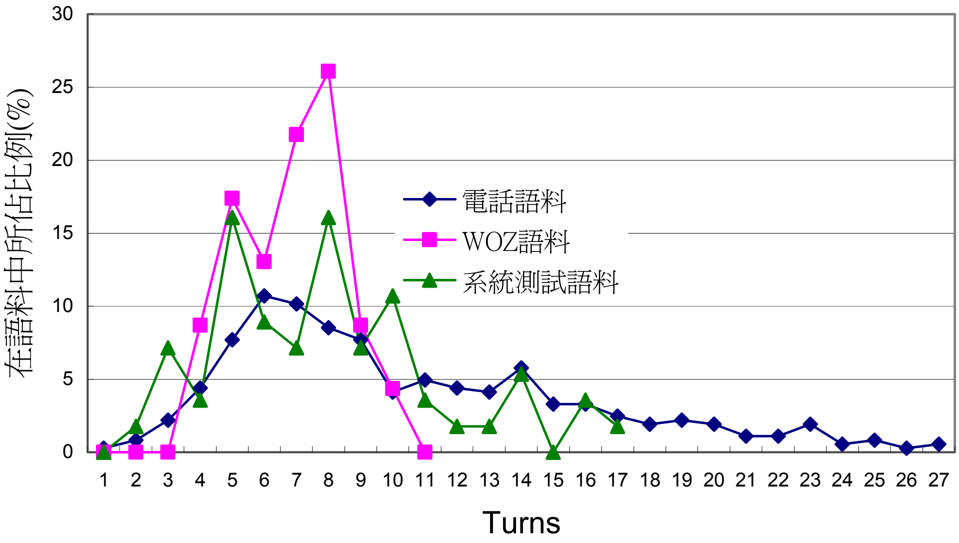
Module	Registration Module	Clinic Query Module	FAQ Module	Integrated System
Average Number of Turns	8.40	9.27	4.80	9.80

表一、各功能模組平均對話長度

從表一，我們可看出常見問答集模組的平均對話長度為 4.80，而掛號諮詢模組的平均對話長度為 8.4，因此我們可推斷圖十二中，第一個高峰是使用者使用常見問答集模組的狀況，第二個高峰則是使用其他模組或者混合的狀況。

將三種語料比較之後，我們可歸納出在電話語料的對話度最長為 11.235 次，因為使用者打電話進來後，常常有些贅詞，如：恩、喂、嘿…等等，就會形成一

次對話，在 WOZ 收集時因為目標明確且由人來回答，使用者可以快速達到目標因此平均對話長度為 6.882 次，系統測試時，因為實際對話流暢度不如人類之間的對話，因此需要較多次的嘗試後，使用者才能達成目標，平均對話長度為 8.928 次，其三種語料的分佈圖，如圖十三所示：



圖十三、對話長度分佈

4.2. 對話系統評估

在評估整個對話系統方面，我們請五十個未參與本研究之大專生來測試系統，並參考[1][24]的方法來評估整個對話系統，分別對各個模組計算對話成功率(Task Success Rate)、平均對話長度(Average Number of Turns)以及答句適切度(Contextual Appropriateness)，為了評估服務的整合，加上一個意圖偵測的正確率(Intention detection Rate)作為其指標，其結果如表二所示。

從表二中可得知，從對話次數來分析，在使用 FAQ 模組時所需對話次數最少，而混合使用時對話次數最多。意圖偵測部分由於掛號諮詢模組與科別建議模組意圖則較容易混淆，因此偵測正確率較低。而整體系統整合時亦然，對話成功率部分，在科別建議模組與整體系統的成功率較低，其原因為在科別建議時需要使用者提供症狀，而症狀的描述方式有非常多種如「頭痛」可以描述成「頭有一點痛」、「我左邊的太陽穴附近會痛」等等，因此導致處理複雜度提高。而整體系

Evaluation Parameters	Intention detection Rate(%)	Task Success Rate(%)	Average Number of Turns	Contextual Appropriate-ness(%)
Registration Module	87.3%	92%	8.40	82.%
Clinic Query Module	84.6%	80%	9.27	75.1%
FAQ Module	92.4%	88%	4.80	85.2%
Integrated System	80.6%	68%	9.80	74.3%

表二、系統效能表

統處理時成功率下降為68%的原因有二：一為自然語言的混淆度而導致，二為意圖偵測錯誤，但這所佔的影響度較小。因此我們可針對每次對話收集的語料再次進行系統改進而改進整體效能，如表三即為第二次改進後的實驗結果：

由表三可明顯看出整體系統意圖偵測正確率由80.6%提升到86.2%，對話長度由9.8次降為9.2次，系統成功率也由68%提升為77%，這主要的效能提升，是由意圖正確率的提升以及自然語言處理能力的增強所提升。

Evaluation Parameters	Intention detection Rate(%)	Task Success Rate(%)	Average Number of Turns	Contextual Appropriate-ness(%)
Registration Module	92.4%	95%	7.30	85%
Clinic Query Module	90.4%	82%	8.20	79.4%
FAQ Module	94.1%	92%	4.70	88.8%
Integrated System	86.2%	77%	9.20	78.5%

表三、系統效能表Evaluation2

6. 結論與未來展望

在本論文中，我們建立一套整合多項服務的醫療查詢對話系統，所整合的服務有掛號資訊諮詢、科別資訊諮詢以及常見問答集諮詢三項大服務。使用意圖偵測來整合服務，在意圖偵測部分則使用部分樣本樹作為判斷意圖之依據，建立部分樣本樹與幫助語言理解必須借重醫療概念模型之推論與概念的描述，透過實驗證明，系統服務成功率為 77%，充分說明了本文所提之方法是具體可行的，但仍有下列問題有待改進：

1.在醫療概念模型抽取部分，本論文使用醫療領域的語料來找出在整個概念模型結構上屬於醫療領域的節點，但是前端的斷詞系統，並不能對句子斷出專有名詞及新詞尤其是屬於領域內之概念詞，因此造成增加雜訊以及屬於醫療領域的節點沒有被找出，因此如能在這部分統計屬於醫療領域的節點時，先做新詞偵測再斷詞，會有效提升醫療概念模型的抽取結果。

2.對於科別建議模組部分，在症狀的描述上隨著使用者的口語化而難以讓系統理解，因此需要提出一套能漸進式的找出使用者可能症狀的方法，如此才能有效提升科別建議模組的效能。

3.對於整個對話系統演進的部分，目前本論文的系統再做過實驗後，除了意圖偵測那一部份，其他部分的效能改進仍然需要不少人力的介入，但對於對話系統而言不斷的演進是必要的，因此如何將演進所需要介入的人力降低就是另一個重要的課題。

致謝

本研究承蒙成功大學王獻章先生提供於成大醫院實際蒐集之語料以及經濟部工研院電通所前瞻技術中心計畫經費補助特此致謝。

參考文獻

[1] Michael F. McTEAR, “Spoken Dialogue Technology: Enabling the Conversational User Interface,” ACM Computer Surveys, Vol 34, No. 1, March 2002, pp.90-169.

[2] James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, Amanda Stent, “Towards Conversational Human-Computer

Interaction," AI Magazine, 2001.

[3] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, "Spoken Language Processing", Prentice-Hall Inc, 2001.

[4] James Allen, "Natural Language Understanding", The Benjamin/Cummings Publishing Company. 1994.

[5] Victor Zue, Stephanie Seneff, James Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen, Lee Hetherington, "JUPITER: A telephone-based conversational interface for weather information," IEEE Trans. on Speech and Audio Processing, vol. 8, no. 1, January 2000, pp.85-96.

[6] AT&T, "How may I help you?" http://www.research.att.com/~algot/hmihy/

[7] S. Bennacef and L.Lamel et al., "Dialogue in the RAILTEL Telephone-Based System," ICSLP'96 Vol. 1. pp. 550-553

[8] L. Lamel, S. Rosset, J. Gauvain, S. Bennacef, M. Garnier-Rizet , and B. Prouts, "The LIMSI Arise System," Speech Communication, 31(4), 2000, pp. 339-353.

[9] Kuei-Kuang Lin, Hsin-Hsi Chen," A Semiautomatic Knowledge Extraction Model for Dialogue Management", Master thesis of department of computer science and information engineering, National Taiwan University, 2002

[10] Bor-shen Lin, Hsin-min Wang, and Lin-shan Lee, "A Distributed Agent Architecture for Intelligent Multi-Domain Spoken Dialogue Systems," IEICE Trans. on Information and Systems, E84-D(9), pp. 1217-1230, Sept. 2001.

[11] Tung-Hui Chiang, Chung-Ming Peng, Yi-Chung Lin,Huei Ming Wang and Shih Chieh Chien, "The Design of a Mandarin Chinese Spoken Dialogue System," in Proceedings of COTEC'98, Taipei 1998, pp. E2-5.1~E2-5.7

[12] Chung-Hsien Wu, Gwo-Lang Yan, and Chien-Liang Lin, "Speech act modeling in a spoken dialogue system using a fuzzy fragment-class Markov model," Speech Communication, Vol. 38, 2002, pp183~199.

[13] Jhing-Fa Wang and Hsien-Chang Wang, "A Portable Auto Attendant System with Sophisticated Dailogue Structure", Journal of Information Science Engineering, Vol.18, No.4, July. 2002, pp627-636

[14] Jhing-Fa Wang and Hsien-Chang Wang" Experiences of Multi-Speaker Dialogue System for Mobile Information Retrieval," Workshop on DSP in Mobile and Vehicular Systems, April 3-4, 2003,Nagoya, Japan

[15] Jhing-Fa Wang and Hsien-Chang Wang, Chieh-Yi Huang, Chung-Hsien Yang" Multi-Speaker Dialogue for Mobile Information Retrieval," ISCSLP 2002, August 23-24, 2002, Taipei

[16] Steffen Staab, Rudi Studer, Hans-Peter Schnurr, York Sure, "Knowledge Processes and Ontologies," IEEE Intelligent Systems 16 (1) January February 2001 ,pp. 26-34

[17] J. F. Sowa, Knowledge Representation: Logical, Philosophical and Computational Foundations, Brooks Cole Publishing Co.1999.

[18] J.l, Chu-Carrol, "MIMIC" An adaptive mixed initiative spoken dialogue system for information queries," Proceedings of the 6th ACL article on Applied Language Processing, pp. 97-104

[19] Lee, S.H., Lee H. and Kim, J.H. 1995 On-line Cursive Script Recognition using an Island-Driven Search Technique, Proceedings of the Third International Conference on Document Analysis and Recognition, Volume: 2 , 14-16 Aug. 1995. 886 -889.

[20] Chung-Hsien Wu, Yeou-Jiunn Chen, and Cher-Yao Yang, "Error Recovery and Sentence Verification Using Statistical Partial Pattern Tree for Conversational Speech," in Proceedings of ICSLP2000, Beijing, China, 2000.

[21] A. Miller, "WordNet: An On-Line Lexical Resource," J. Lexicography, vol.3, no.4, Dec.1990.

[22] 董振東, 董強, "知網," http://www.keenage.com/

[23] N. Fraser and G. N. Gilbert, "Simulating speech system," Computer Speech and Language 5, 1991 pp.81-99.

[24] M. A. Walker, D. Litman, C. Kamm, and A. Abella, PARADISE: a general framework for evaluating spoken dialogue agents. In Proceedings of the 35th Annual General Meeting of the Association for Computational Linguistics, ACL/EACL (Madrid, Spain). ACL, 1997, pp.271–280.