International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI Core, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而
治後世聖人易之以書
契百官以治萬民以察
說文敘曰蓋文字者經
藝之本宣教明化之始
前人所以垂後後人所
以識古故曰本立而道
生知天下之至賾而不
可亂也教化既萌文心
雕龍則謂人之立言因
字而生句積句而成章
積章而成篇篇之彪炳

Vol.17 No.3 September 2012 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 17, no. 3, September 2012

Contents

Papers

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 1-28

Effects of Combining Bilingual and Collocational Information on Translation of English and Chinese Verb-Noun Pairs[1]

Yi-Hsuan Chuang[*], Chao-Lin Liu[*], and Jing-Shin Chang[+]

Abstract

We studied a special case of the translation of English verbs in verb-object pairs. Researchers have studied the effects of the linguistic information of the verbs being translated, and many have reported how considering the objects of the verbs will facilitate the quality of translation. In this study, we took an extreme approach - assuming the availability of the Chinese translation of the English object. In a related exploration, we examined how the availability of the Chinese translation of the English verb influences the translation quality of the English nouns in verb phrases with analogous procedures. We explored the issue with 35 thousand VN pairs that we extracted from the training data obtained from the 2011 NTCIR PatentMT workshop and with 4.8 thousand VN pairs that we extracted from a bilingual version of *Scientific American* magazine. The results indicated that, when the English verbs and objects were known, the additional information about the Chinese translations of the English verbs (or nouns) could improve the translation quality of the English nouns (or verbs) but not significantly. Further experiments were conducted to compare the quality of translation achieved by our programs and by human subjects. Given the same set of information for translation decisions, human subjects did not outperform our programs, reconfirming that good translations depend heavily on contextual information of wider ranges.

[1] This paper was converted from the Master’s thesis of the first author, which was partially published in (Chuang *et al*., 2011a) and (Chuang *et al*., 2011b).

[*] National Chengchi University, Taiwan,
E-mail: {98753004, chaolin}@nccu.edu.tw
The author for correspondence is Chao-Lin Liu.
[+] National Chi Nan University, Taiwan
E-mail: jshin@csie.ncnu.edu.tw

Keywords: Machine Translation, Feature Comparison, Near Synonyms in Chinese, E-HowNet, Human Judgments

1. Introduction

In general, the problem we are exploring is an instance of *translation of collocations* (Smadja *et al*., 1996). The collocations consist of the verbs and their direct objects, *i.e*., nouns, in verb phrases. Researchers have extensively studied the translation problems related to individual verbs/nouns (Dorr *et al*., 2002; Lapata & Brew, 2004) and verbs/nouns in phrases (Chuang *et al*., 2005; Koehn *et al*., 2003; Lü & Zhou, 2004). Some techniques have been developed for text of special domains (Seneff *et al*., 2006). The techniques are applicable in many real-world problems, including computer-assisted language learning (Chang *et al*., 2008) and cross-language information retrieval (Chen *et al*., 2000).

We work on the processing of patent documents (Lu *et al*., 2010; Yokoama & Okuyama, 2009), and present an experience in translating common verbs and their direct objects based on bilingual and collocational information. In this study, we took an extreme assumption of the availability of the Chinese translations of the English objects to examine whether the extra information would improve the quality of verbs' translations. The proposed methods are special in that we are crossing the boundary between translation models and language models by considering information of the target language in the translation task. The purpose of conducting such experiments was to investigate how the availability of such bilingual information might contribute to the translation quality. It is understood and expected by many that the Chinese translations of English words might not be directly available for all cases and that a good translator should consider a lot more features to achieve high translation quality. Nevertheless, we thought it would be interesting to know how the availability of such extraordinary information could influence the translation quality within the context that we present in this paper.

The experiments were conducted with the training data available to the participants of the 2011 NTCIR Patent MT task. The original corpus contains one million pairs of a Chinese word and its English translation. We explored four different methods to determine the verb's Chinese translation. These methods utilized the bilingual and contextual information of the English verbs in different ways. Effects of these methods were compared based on experimental evaluation that was conducted with 35 thousand verb-object pairs extracted from the NTCIR corpus. Additional experiments using materials in a bilingual version of *Scientific American*[2] magazine were also conducted. (Since objects are nouns, we will refer to verb-object pairs as verb-noun pairs or VN pairs to simplify the wording.)

[2] http://www.scientificamerican.com/

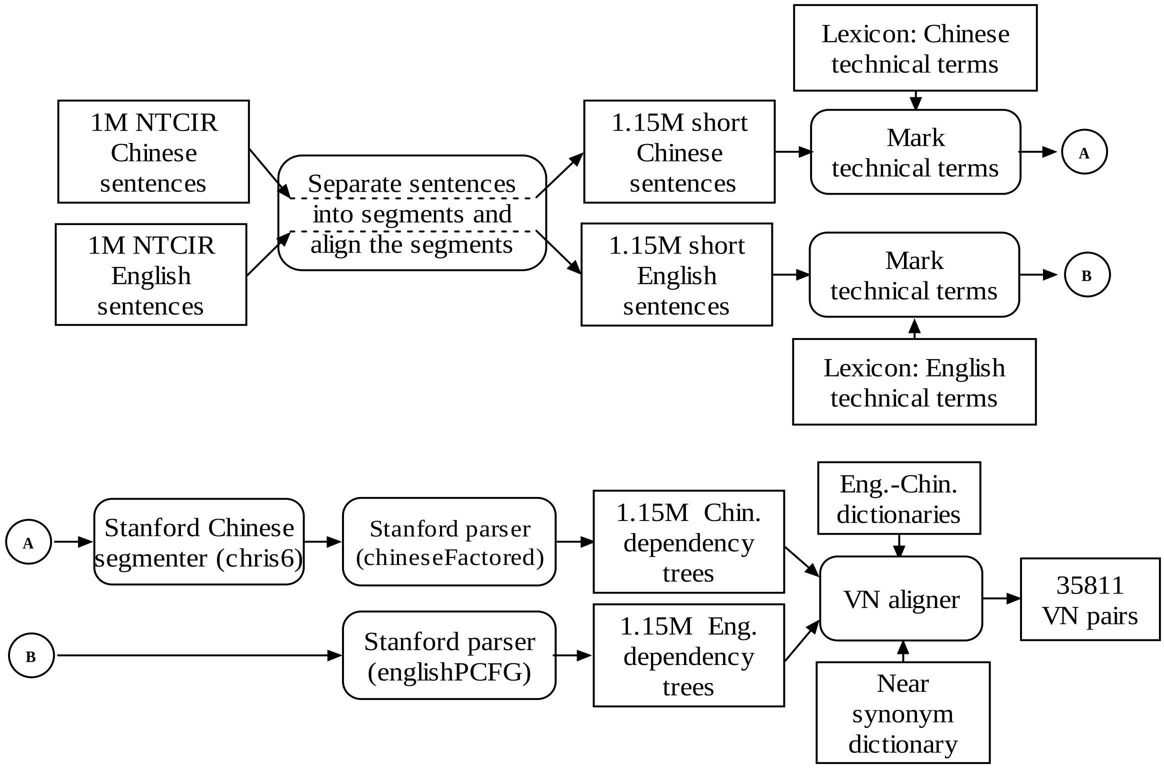

Lexicon: Chinese technical terms
1M NTCIR Chinese sentences
1M NTCIR English sentences
Separate sentences into segments and align the segments
1.15M short Chinese sentences
1.15M short English sentences
Mark technical terms
Mark technical terms
A
B
Lexicon: English technical terms
A
Stanford Chinese segmenter (chris6)
Stanford parser (chineseFactored)
1.15M Chin. dependency trees
Eng.-Chin. dictionaries
VN aligner
35811 VN pairs
B
Stanford parser (englishPCFG)
1.15M Eng. dependency trees
Near synonym dictionary

Figure 1. The procedure for extracting VN pairs from the original corpora

We provide a broad outline of our work in Section 2, and we present our methods for aligning the bilingual VN-pairs in Section 3. We explain how we build lexicons with information about synonyms to serve the needs of VN-pair alignment in Section 4 and delineate the design of our experiments in Section 5. We discuss the experimental results in Section 6, and we compare the translation quality achieved by human subjects in Section 7. Finally, we wrap up this paper in Section 8.

2. The Big Picture

Our work consisted of two major stages. We extracted the VN pairs from the original corpus. Then, we applied our translation methods to translate English words into Chinese and *vice-versa* before comparing the translation quality achieved through different combinations of collocational and bilingual information.

Figure 1 shows how the VN pairs were extracted from the 1 million parallel sentences, which we obtained from the NTCIR 9 PatentMT task in 2011.[3] The process started from the

[3] http://ntcir.nii.ac.jp/PatentMT/

upper left of the figure. Most of the original sentences were very long. A sentence had 34 words on average, and the longest sentence had 141 words. Since our goal was to extract VN pairs from the corpus, not doing a full-scale research project in machine translation, we chose to segment the sentences into shorter parts at commas and periods. Normally, VN pairs will not expand across punctuation; even if some VN pairs did, we could afford to neglect them because we had 1 million pairs of long sentences.

We then re-aligned the short English and Chinese segments with a sentence aligner (Tien *et al.*, 2009) that we implemented based on the concept of Champollion (Ma, 2006). We treated the original long sentence pairs as aligned paragraphs, and we ran our aligner on the sentences that originally belonged to a long sentence. Like the Champollion, we computed scores for the sentence pairs, so we could choose those pairs with higher scores to achieve higher confidence on the aligned pairs. More specifically, we kept only the leading 33% of the short sentence pairs, and obtained 1,148,632 short sentence pairs.

We employed the Stanford Chinese segmenter[4] (Chang *et al.*, 2008; Tseng *et al.*, 2005) to segment the Chinese text. This segmenter allows us to mark the technical terms, so the segmenter will treat the words belonging to technical terms as a unit, preventing them from being segmented again. In addition, currently, our technical terms are nouns, so they are annotated accordingly. When there were multiple ways to mark the technical terms in a string, we preferred the longer choices. English texts were tokenized by the Stanford parser[5] with the PCFG grammar (Klein & Manning, 2003). Technical phrases and compound words in English were also marked and would not be treated as individual words either. The special terms came from the glossary that will be explained in Section 4.1.

Based on these short sentence pairs, we aligned the VN pairs with the method in Section 3. This process employed an English-Chinese glossary for technical terms, which we will discuss in Section 4.1, and a bilingual dictionary enhanced with Chinese near synonyms, which we will discuss in Section 4.2. In the end, we accepted 35,811 VN pairs to be used experiments at the second stage.

During the second stage of our work, we split the VN pairs into training and test data. Useful statistics were collected from the training data and were applied to select Chinese translations for the English words in question. Details about the design and results of the experiments are provided in Sections 5 and 6.

[4] http://nlp.stanford.edu/software/segmenter.shtml, version 1.5

[5] http://nlp.stanford.edu/software/lex-parser.shtml, with the PCFG grammar, version 1.6.5

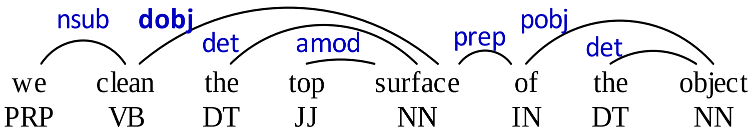


Figure 2. A sample dependency tree with POS tags

Remove the small bar.	移開小塊的木條
root(ROOT-0, Remove-1)	root(ROOT-0, 移開-1)
det(bar-4, the-2)	dep(的-3, 小塊-2)
amod(bar-4, small-3)	assmod(木條-4, 的-3)
dobj(Remove-1, bar-4)	dobj(移開-1, 木條-4)

Figure 3. A pair of aligned sentences and their dependency trees, where the `dobj` relationships can be aligned

3. VN Pair Alignment

We employed the Stanford parsers to compute the dependency trees for the parallel texts for English and Chinese. We extracted the `dobj` relations from the trees and aligned the VN pairs.

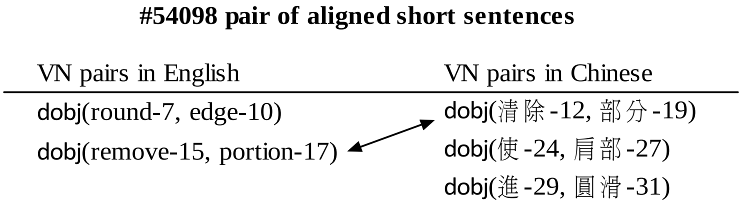
3.1 Dependency Trees

Based on the general recommendations on the Stanford site, we parsed English with the englishPCFG.ser.gz grammar, and parsed Chinese with the chineseFactored.ser.gz grammar.

Figure 2 shows the dependency tree for a simple English sentence, “we clean the top surface of the object.” Stanford parsers can provide the parts of speech (POSs) of words and recognize the relationships between the words. POSs are shown below the words, and the relationships are attached to the links between the words. The `dobj` link between “clean” and “surface” indicates that “surface” is a direct object of “clean,” and we could rely on such `dobj` links to identify VN pairs in the corpus.

3.2 VN Pair Alignment

We found 375,041 `dobj` links in the 1.15M short English sentences and 465,866 `dobj` links in the short Chinese part. Nevertheless, not all of the words participating in a `dobj` link were real words, and the tags for the English and the Chinese short sentences did not always agree. Figure 3 shows the dependency trees of a sample pair of short sentences containing two `dobj` relationships that would be aligned (English on the left; Chinese on the right).



#54098 pair of aligned short sentences

VN pairs in English	VN pairs in Chinese
dobj(round-7, edge-10)	dobj(清除-12, 部分-19)
dobj(remove-15, portion-17)	dobj(使-24, 肩部-27)
	dobj(進-29, 圓滑-31)

Figure 4. Aligning VN pairs within an aligned short sentence

Fill the hole with water.	將洞裝滿水
root(ROOT-0, Fill-1)	nn(洞-2, 將-1)
det(hole-3, the-2)	nsubj(裝滿-3, 洞-2)
dobj(Fill-1, hole-3)	root(Root-0, 裝滿-3)
prep(Fill-1, with-4)	dobj(裝滿-3, 水-4)
pobj(with-4, water-5)	

Figure 5. A pair of aligned sentences that we could not align via the dobj relationships

Hence, we took two steps to align the VN pairs. First, we looked up the English and Chinese words in our bilingual lexicon, which we will explain in Section 4.2. If the lexicon did not contain the words, we would not use the words in the corresponding dobj links as VN pairs. After this step, we had 254,091 and 249,591 VN pairs in English and Chinese, respectively. We then tried to align the remaining English and Chinese VN pairs, noting that only those VN pairs that originated from the same pair of long sentence pairs can be aligned.

The alignment is not as trivial as it might appear to be. Let (EV, EN) and (CV, CN) denote an English and a Chinese VN pair, respectively; let EV, EN, CV, and CN denote an English verb, an English noun, a Chinese verb, and a Chinese noun, respectively. We had to check whether CV is a possible translation of EV and whether CN is a possible translation of EN. If both answers are positive, then we aligned the VN pairs. An illustration of this basic idea is shown in Figure 4, where the English and the Chinese short sentences contained multiple dobj relationships and only one pair could be aligned.

Nevertheless, even when an English verb can carry only one sense, there can be multiple ways to translate it into Chinese, and there is no telling whether a dictionary will include all of the possible translations and contain the Chinese translations actually used in the Patent MT corpus. For instance, (improve, quality) can be translated to (改善(gai3 shan4), 品質(pin3 zhi2)) or (改進(gai3 jin4), 品質). If an English-Chinese dictionary only lists “改善” as the translation for “improve” but does not include “改進” as a possible translation, then we could not use that dictionary to align (improve, quality) and (改進, 品質). We need a way to tell that “改進” and “改善” are interchangeable.

Therefore, we expanded the set of possible Chinese translations in a given dictionary with near synonyms, and employed the expanded dictionary to enhance the quality of VN pair alignment. The process of constructing the expanded dictionary is provided in Section 4.2.

After completing the VN pair alignment, we obtained 35,811 aligned VN pairs, *cf*. Figure 1. Note that, although we started with 1 million pairs of long sentences, we identified less than 36 thousand VN pairs. Many problems contributed to the small number of extracted pairs. We have mentioned that translators might not use the words in dictionaries available to us when they translated. We removed `dobj` relationships that contained words not in our dictionaries. Also, the parser might not parse sentences as one might expect, and we show an example of this in Figure 5. The parser considered the "hole" as the object in the English sentence and considered "water" (水 (shui3)) as the object of the "fill" (裝滿 (zhuang1 man3)) in the Chinese sentence. Hence, the two `dobj` relationships could not be aligned.

4. Lexicon Constructions

We explain (1) how we built the glossary of technical terms and (2) how we constructed a bilingual dictionary that contains information about near synonyms in this section.

4.1 Creating a Glossary of Technical Terms

As explained in Section 2, we built a glossary of technical terms to distinguish technical terms from normal text, thereby achieving higher quality of parsing.

We downloaded 138 different kinds of domain-dependent dictionaries from Taiwan National Academy for Educational Research.[6] The files contained technical term pairs in the form of (English word(s), Chinese word(s)) that were stored in Excel format. The total file size is 177MB.

The format of English-Chinese technical term pairs is not always a one-to-one relationship; some English technical terms have more than one translation in Chinese. We converted such pairs into multiple one-to-one pairs, and acquired 804,068 English-Chinese technical one-to-one term pairs.

To validate the reliability of the glossary, we conducted a small experiment; that is, to segment patent sentences with the glossary. The results showed that the coverage of these "technical term" pairs was too broad, and a plethora of ordinary words were considered technical terms.

We alleviated this problem with E-HowNet[7] (Chen *et al*., 2005) and WordNet.[8] Treating

[6] http://terms.nict.gov.tw/

[7] http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-edoc.htm

the words listed in E-HowNet and WordNet as ordinary words, we used them to identify ordinary words in our technical term pairs. If the English or the Chinese parts of the original pairs were also listed in E-HowNet or WordNet, then the pairs would be removed.

As a result, we removed 14% of the original pairs and kept 690,640 technical term pairs. The English and Chinese parts of the term pairs then were used as two dictionaries of "technical terms," shown in Figure 1.

4.2 The English-Chinese Dictionary and Near Synonyms

As announced in Section 3.2, we built a bilingual dictionary and enhanced it with information about near synonyms to improve the recall rates of the VN pair alignment.

A good English-Chinese dictionary is the basis for the task of VN pair alignment. We collected and combined the Chinese translations of English words in the Concise Oxford English Dictionary and the Dr.Eye online dictionary[9] to acquire 99,805 pairs of English words and their translations.

As we explained in Section 3.2, the Chinese translations listed in the dictionaries might not be complete, so we enhanced the merged dictionary with information about near synonyms. We employed two sources of relevant information to obtain near synonyms in this study.

The Web-based service of Word-Focused Extensive Reading System[10] (Cheng, 2004) is maintained by the Institute of Linguistics of the Academia Sinica in Taiwan. The service allows us to submit queries for the near synonyms of Chinese words for free, so we collected the near synonyms from the web site. Given an entry in our bilingual dictionary, we queried the near synonyms for each of the Chinese translations of an English word and added the results to the Chinese translations of the English word.

E-HowNet is another source of computing and obtaining near synonyms. E-HowNet is a lexicon for Chinese. Each entry in E-HowNet provides the information about a sense of a Chinese word. If a word can carry multiple senses, the word will have an entry for each of its senses. Among other items, an entry contains two levels of detailed semantic information for a word: TopLevelDefinition and BottomLevelExpansion. The TopLevelDefinition item in a lexical entry records the higher semantic information in the E-HowNet Ontology[11] (Chen *et al*, 2005). In contrast, the BottomLevelExpansion item in a lexical entry records the semantic information at the lowest level in the E-HowNet Ontology. The TopLevelDefinition may not contain any information when the TopLevelDefinition is the same as the same as the

[8] http://wordnet.princeton.edu/

[9] http://www.dreye.com/index_en.html

[10] http://elearning.ling.sinica.edu.tw/c_help.html

[11] http://ckip.iis.sinica.edu.tw/taxonomy/taxonomy-edoc.htm

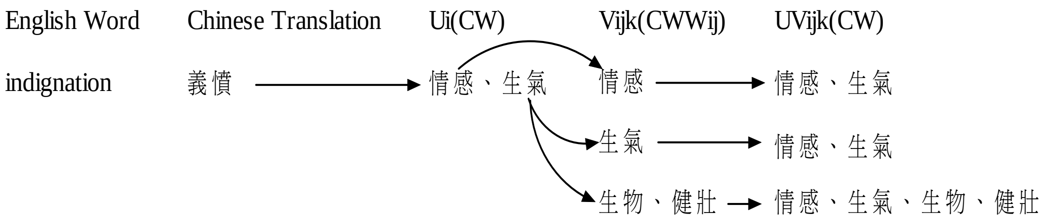


Figure 6. Expanding the Chinese translations of an English word with near synonyms

BottomLevelExpansion. The semantic definitions provided in these two entries can be used to compute similarity scores between word senses.

We determine whether two Chinese words are near synonyms by the following procedure. Given a Chinese word, CW, we looked in E-HowNet for its senses. Let S_i(CW) be one of CW’s senses. We combined the semantic definitions listed in the TopLevelDefinition and BottomLevelExpansion of S_i(CW), which might include multiple Chinese words. Denote this set of Chinese words by U_i(CW), and let CWW_{ij} be a word in U_i(CW). We looked in E-HowNet for the senses of CWW_{ij}. Let $S_k(CWW_{ij})$ denote one of the senses of the CWW_{ij}, and let $V_{ijk}(CWW_{ij})$ denote the set of Chinese words in the combined semantic definitions listed in the TopLevelDefinition and BottomLevelExpansion of $S_k(CWW_{ij})$. Finally, we computed the union of U_i(CW) and $V_{ijk}(CWW_{ij})$ as a sense vector UV_{ijk} of S_i(CW). Note that, due to lexical ambiguity, a Chinese word might have multiple such vectors.

Figure 6 shows an illustration of the process of finding near synonyms for “義憤” (yi4 fen4), which is a possible translation for “indignation”. In this illustration, we assume (1) that there is only one sense for “義憤” and (2) that its semantic information contains two Chinese words: “情感” (qing2 gan3) and “生氣” (sheng1 qi4). Namely, we have CW=“義憤”, U_1(CW)={“情感”, “生氣”}, CWW_{11}=“情感”, and CWW_{12}= “生氣”. There is only one sense for CWW_{11}, and its combined semantic information contains only one Chinese word “情感”. Hence, $V_{111}(CWW_{11})$={“情感”}. There are two senses for CWW_{12}. The combined semantic information for $S_1(CWW_{12})$ contains only “生氣,” and the combined semantic information for $S_2(CWW_{12})$ contains only “生物” (sheng1 wu4) and “健壯” (jian4 jhuang4). Therefore, $V_{121}(CWW_{12})$={“生氣”} and $V_{122}(CWW_{12})$={“生物”, “健壯”}. Finally, we compute the unions of U and V sets to acquire UV_{111}(CW)={“情感” , “生氣”}, UV_{121}(CW)={“情感”, “生氣”}, and UV_{122}(CW)={“情感”, “生氣”, “生物”, “健壯”}. Although we have three sets, only two of them are different. Similar to how we compute the sense vectors for “義憤” in Figure 6, we can compute the sense vectors for any Chinese words.

We treated two Chinese words as near synonyms if the cosine value of any of their sense vectors exceeded 0.7.[12] To compute the cosine value of two sense vectors, we first computed the union of the words in two vectors, treated each different word as a different dimension, and converted the word vector into a Boolean vector. Therefore, if a word in a vector did not appear in another vector, a "0" would be used in its place. Assume that we were to compute the cosine of UV_{121}(CW) and UV_{122}(CW) in the preceding paragraph, we would create a 4-dimension space of {"情感," "生氣," "生物," "健壯"}, UV_{121}(CW) would become {1, 1, 0, 0}, and UV_{122}(CW) would become {1, 1, 1, 1}.

Given an entry in our bilingual dictionary, we computed the near synonyms of the Chinese translations of each English word. This was carried out by comparing the sense vectors of Chinese translations in every English-Chinese pair with the sense vectors of 88,074 Chinese words in E-HowNet. The qualified words were added to the Chinese translations of the English words in our dictionary.

Thus, an entry for an English word in our English-Chinese dictionary includes four parts. The first part is the English word itself. The second part is the Chinese translations that we found in our dictionaries (Oxford and Dr.Eye). The third part is the synonyms, obtained from Cheng's (2004) system, for the words in the second part. The fourth part is the near synonyms that we computed with the aforementioned procedure (with E-HowNet).

The purpose of adding information about near synonyms into our bilingual dictionary was to increase the recall rates of VN-pair alignment. Having not-very-good Chinese near synonyms may not hurt our performance, unless the translators of the PatentMT corpus happened to use the same erroneous translations. Nevertheless, more complex methods for identifying synonyms, *e.g.* Bundanitsky and Hirst (2006) and Chang and Chiou (2010), may be instrumental for the study.

5. Design of the Experiments

We conducted experiments to translate from English to Chinese and from Chinese to English. In addition, in separate experiments, we tried to find the best translations of verbs, and tried to find the best translations of objects of the verbs given appropriate contexts. Nevertheless, we present the design of our experiments only with the experiments of translating English verbs to Chinese verbs in this section. Other experiments were conducted with the same procedure.

[12] Given that we did not have the context to do word sense disambiguation at this stage, we have to consider two words synonymous to each other if any of their senses are close enough. This threshold of 0.7 was chosen based on observed results of small-scale experiments and was not chosen scientifically.

5.1 Statistics about the Aligned VN pairs

We calculated the frequencies of the verbs in the 35,811 aligned VN pairs and ranked the verbs based on the observed frequencies. Table 1 shows the 20 most frequent English verbs and their frequencies. We identified the 100 most frequent English verbs and the corresponding aligned VN pairs in our experiments. In total, there were 30,376 such aligned VN pairs. The most frequent English verb appeared 4,530 times, as shown in Table 1. The 100th most frequent English verb is “lack,” and it appeared 47 times.

Table 1. 20 most frequent English verbs in the aligned VN pairs

Verb	have	provide	use	include	comprise	contain	form	receive	reduce	perform
Freq.	4530	3345	1993	1954	1588	1080	914	863	774	616
Verb	increase	produce	maintain	determine	represent	show	obtain	achieve	improve	allow
Freq.	465	453	397	382	373	352	329	329	322	287

Some of the English verbs are easier to translate than others. We can calculate the frequencies of the Chinese translations of verbs to verify the differences. For instance, “add” was translated in five different ways: “增加” (zeng1 jia1) 48 times, “添加” (tian1 jia1) 44 times, “加入” (jia1 ru4) 43 times, “加上” (jia1 shang4) 2 times, and “增添” (zeng1 tian1) 1 time. The distribution, (48, 44, 43, 2, 1), is not very skewed, and the frequencies of the most frequent translation and the second most frequent translation are close. Therefore, we would not achieve very good results if we should choose to use the most frequent translation for all occurrences of “add”.

Table 2. 22 most “challenging” English verbs and their indices

Verb	make	exhibit	add	represent	retain	leave	enhance	reduce	lack	improve	achieve
	1.00	1.09	1.09	1.21	1.21	1.22	1.25	1.26	1.27	1.33	1.39
Verb	employ	reach	create	give	replace	take	apply	adjust	obtain	carry	explain
	1.41	1.43	1.50	1.54	1.69	1.69	1.69	1.72	1.76	1.82	2.00

Based on this observation, we defined the ***challenging index*** of a word as the ratios of the frequency of their most frequent translation against the frequency of their second most frequent translation. The challenging index of “add” mentioned in the previous paragraph is 1.09.

This challenging index is not a scientifically-proven index for difficulty for translation, but could serve as a heuristic. Intuitively, larger challenging indices imply that it is easier to achieve good translations via the most frequent translations. Table 2 lists the 22 verbs that had the smallest challenging indices.

5.2 Translation Decisions

Given the aligned VN pairs, we could compute conditional probabilities and apply the conditional probabilities to determine the Chinese translation of English words.

Table 3. Translation decisions

$\arg\max_{CV_i} \Pr(CV_i \mid EV)$	(1)
$\arg\max_{CV_i} \Pr(CV_i \mid EV, EN)$	(2)
$\arg\max_{CV_i} \Pr(CV_i \mid EV, EN, CN)$	(3)
$\arg\max_{CV_i} \Pr(CV_i \mid EV, CN)$	(4)

Table 3 lists four possible ways to choose a Chinese translation for an English verb in a VN pair. Equation (1) is the most simplistic. Let EV denote a specific English verb and CV_i be one of EV's translations observed in the training data. Given the English verb, the equation chooses the CV_i that maximizes the conditional probability. Namely, at the test stage, Equation (1) prefers the most frequent Chinese translation of EV in the training data.

We could obtain the conditional probability $\Pr(CV_i|EV)$ by dividing the frequency of observing the VN pair (EV, CV_i) in the training data by the frequency of observing EV in any VN pairs. Using the data for "add" that we mentioned in Section 5.1 as an example, we observed 135 occurrences of "add". Therefore, Pr("增加" | "add") = 48/135=0.356 and Pr("加上" | "add") = 2/135=0.015.

Let EN be a specific English noun. Equation (2) considers the object of the verb when choosing the verb's translation. Let $C(\bullet)$ denote the frequency of a given event. The conditional probability in Equation (2) is defined in Equation (5). C(EV, EN) denotes the frequency that we observed the occurrences of EV and EN in the training data, and C(EV, EN, CVi) denotes the frequency that we observed the occurrence of EV, EN, and CVi in the training data.

$$\Pr(CV_i \mid EV, EN) = \frac{C(EV, EN, CV_i)}{C(EV, EN)} \tag{5}$$

The remaining equations, (3) and (4), take extreme assumptions. We assumed the availability of the Chinese translation of the English object at the time of translation and used this special information in different ways. Equation (3) considers the words EV, EN, and CN. In a strong contrast, Equation (4) considers only EV and CN to determine the translation of the English verb. The conditional probabilities in Equations (3) and (4) were calculated using Equation (6) and (7), respectively, based on the training data.

$$\Pr(CV_i \mid EV, EN, CN) = \frac{C(EV, EN, CV_i, CN)}{C(EV, EN, CN)} \tag{6}$$

$$\Pr(CV_i \mid EV, CN) = \frac{C(EV, CN, CV_i)}{C(EV, CN)} \tag{7}$$

We felt that the exploration of using the information about the Chinese translation of the English noun would be interesting. Would the information about CN provide more information, assuming we had information about EV and EN? What would we achieve when we had information about only EV and CN but not EN?

In all of the experiments, we used 80% of the available aligned VN pairs as the training data and the remaining 20% as the test data. The training data were randomly sampled from the available data.

As a consequence, it was possible for us to encounter the zero probability problems. Take Equation (6) for example. If, for a training case, we needed C(EV, EN, CN) in Equation (6), but we happened not to have observed any instances of (EV, EN, CN) in the aligned VN pairs in the training data, then we would not be able to compute Equation (6) for the test case. When such situations occurred, we chose to allow our system to admit that it was not able to recommend a translation, rather than resorting to smoothing techniques.

6. Experimental Results

Using the formulas listed in Table 3 would allow our systems to recommend only one Chinese translation. In fact, we relaxed this unnecessary constraint by allowing our systems to consider the largest *k* conditional probabilities and to recommend *k* translations.

Although we have been presenting this paper with the 1 million parallel sentences in NTCIR PatentMT data as the example, we also have run our experiments with the English-Chinese bilingual version of *Scientific American*. Moreover, we ran experiments that aimed at finding the best Chinese translations of English objects. The formulas were defined analogously with those listed in Table 3.

6.1 Basic Results for the Top 100 Verbs in Patent Documents

When we conducted experiments for the top 100 verbs (*cf.* Section 5.1), we had 24,300 instances of aligned VN pairs for training and 6,076 instances of aligned VN pairs for testing.

We measured four rates as the indication of the performance of using a particular formula in Table 3: rejection rates, inclusion rates, average number of actual recommendations, and average ranks of the answers.

The ***rejection rate*** is the percentage of not being able to respond to the test cases. This is due to our choosing not to smooth the probability distributions, as we explained at the end of Section 5.2.

The rejection rates were 0, 0.201, 0.262, and 0.218 when we applied Equations (1) through (4) in the experiments. It is not surprising that the rejection rates increased as we considered more information in the formulas. As expected, we encountered the highest rejection rate when using Equation (3), when we essentially collected information about four grams at the training stage. Note that using Equation (4) resulted in higher rejection rates than using Equation (2). To have to reject a test instance when we used Equation (2), we must have had no prior experience with the EN in our training data. In contrast, to have to reject a test instance when we used Equation (4), we must have had no prior experience with the CN in our training data. In reality, it was much likely not to have observed a CN for the EN in our training data than not to have observed the EN at all. Hence, it is more likely for Pr(CV_i | EV CN) to be zero than Pr(CV_i | EV EN), and the rejection rates for Equation (4) were higher.

Table 4. Inclusion rates for the top 100 verbs

Inclusion	*k*=1	*k*=3	*k*=5
Eq(1)	0.768	0.953	0.975
Eq(2)	0.786	0.913	0.918
Eq(3)	0.795	0.911	0.916
Eq(4)	0.791	0.910	0.916

Table 4 shows the ***inclusion rates***: rates of the correct answers included in the recommended *k* translations. We did not consider the cases where our systems could not answer in computing the statistics in Table 4. Hence, the data show the average inclusion rates when our systems could respond. As one may have expected, when we increased *k*, the inclusion rates also increased.

The comparison between the results for using Equations (3) and (4) and the results of using Equation (2) show that using the bilingual information about CN improved the translation quality when *k*=1, but the changes in the inclusion rates were marginal.

It may also be surprising that the inclusion rates for Equations (2) through (4) seem to be saturated when we increase *k* from 3 to 5. This was because our systems actually could not recommend 5 possible translations when they were allowed to. Although we had hundreds or thousands of aligned VN pairs for an English verb, *cf*. Table 1, including more conditioning information in Equations (2) through (4) still reduced the number of VN pairs qualified for training and testing, consequently limiting the actual numbers of available translations to recommend. Table 5 shows the average number of actual recommendations in the tests. Even when we allowed 5 recommendations (*k*=5), using Equations (2) through (4) produced only about 2 recommendations on average. This phenomenon limited the chances to increase the inclusion rates when we increased *k*.

Table 5. Average number of actual recommendations

Recommend	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	2.919	4.614
Eq(2)	1.000	1.923	2.225
Eq(3)	1.000	1.847	2.107
Eq(4)	1.000	1.920	2.244

Table 6. Average ranks of the answers

Ranking	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	1.241	1.310
Eq(2)	1.000	1.166	1.185
Eq(3)	1.000	1.151	1.168
Eq(4)	1.000	1.153	1.173

The main advantage of using Equations (2) through (4) is that they were more precise when they could answer. Table 6 shows the average ranks of the correct translations in the recommended translations. The first word in the recommendation list is considered Rank 1, the second word is Rank 2, *etc*. Hence, we preferred to have smaller average ranks. The average ranks improved as we considered more information from Equation (1) to Equation (2) and to Equation (3). Using Equation (2) achieved almost the same quality of translations as using Equation (4). Equation (2) achieved better inclusion rates, but Equation (4) offered better average ranks.

6.2 Improving Results for the Top 100 Verbs in Patent Documents

Results reported in the previous subsection indicated that Equation (1) is robust in that it could offer candidate answers all the time. Methods that employed more information could recommend translations more precisely, but were less likely to respond to test cases. Hence, a natural question is whether we could combine these methods to achieve better responsiveness while maintaining the translation quality. To this end, we examined all of the combinations of the basic methods listed in Table 3.

Table 7. Inclusion rates (combined methods)

Inclusion	*k*=1	*k*=3	*k*=5
Eq1	0.768	0.953	0.975
Eq2+Eq1	0.772	0.960	0.979
Eq3+Eq1	0.778	0.960	0.979
Eq4+Eq1	0.776	0.959	0.978

Table 8. Average ranks of the correct answers (combined methods)

Ranking	*k*=1	*k*=3	*k*=5
Eq1	1.000	1.241	1.310
Eq3	1.000	1.151	1.168
Eq2+Eq1	1.000	1.240	1.301
Eq3+Eq1	1.000	1.234	1.294
Eq4+Eq1	1.000	1.233	1.296

In Tables 7 and 8, we use the notation EqX+EqY to indicate that we used Equation (X) to find as many candidate translations as possible before we reached a total of *k* recommendations. If applying Equation (X) could not offer sufficient candidate translations, we applied Equation (Y) to recommend more candidate translations until we acquired *k* recommendations.

Using Equation (1) is sufficiently robust in that the conditional probabilities would not be zero, unless the training data did not contain any instances that included the English verb. Nevertheless, using Equation (1) is relatively less precise (*cf.* Table 6). Hence, we used Equation (2) through Equation (4) before using Equation (1) as a backup. Naturally, in these experiments, the rejection rates for "Eq2+Eq1," "Eq3+Eq1," and "Eq4+Eq1" became zero. In other words, our systems responded to all test cases when we used these combined methods to recommend *k* candidates.

In Tables 7 and 8, we compare the performance of these combined methods. We copy the inclusion rates of Equation (1) from Table 4 to Table 7 to facilitate the comparison, because Equation (1) was the best performer, on average, in Table 4. The combined methods improved the inclusion rates, although the improvement was marginal.

Moreover, we copy the average ranks for Equation (1) and Equation (3) from Table 6 to Table 8. Using Equation (1) and using Equation (3) led to the worst and the best average ranks, respectively, in Table 6. Again, using the combined methods, we improved the average ranks marginally over the results of using Eq. 1.

Table 9. Inclusion rates for the 22 challenging verbs

Inclusion	*k*=1	*k*=3	*k*=5
Eq(1)	0.449	0.865	0.923
Eq(2)	0.561	0.818	0.820
Eq(3)	0.564	0.827	0.829
Eq(4)	0.550	0.827	0.829

Table 10. Average number of recommendations

Recommend	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	2.977	4.756
Eq(2)	1.000	2.090	2.364
Eq(3)	1.000	2.022	2.230
Eq(4)	1.000	2.106	2.411

Statistics in Table 7 suggest that using this machine-assisted approach to translate verbs in common VN pairs in the PatentMT data is feasible. Providing the top five candidates to a human translator to choose will allow the translator to find the recorded answers nearly 98% of the time. Statistics in Table 7 and Table 8 show that the combined methods were able to improve the inclusion rates and the ranks of the correct answers at the same time.

It is interesting to find that using Equation (2) and Equation (4) did not lead to significantly different results in Tables (4) through (8). The results suggest that using either the English nouns or the Chinese nouns as a condition in the translation decisions (*cf*. Table 3) contributed similarly to the translation quality of the English verbs.

6.3 Results for the Most Challenging 22 Verbs in Patent Documents

We repeated the experiments that we conducted for the top 100 verbs for the most challenging 22 verbs (*cf*. Section 5.1). Tables 9 through 13 correspond to Tables 4 through 8, respectively. The most noticeable difference between Table 9 and Table 4 is the reduction of the inclusion rates achieved by Equation (1) when *k*=1. Although the inclusion rates reduced noticeably when we used Equation (2), Equation (3), and Equation (4) as well, the drop in the inclusion rate for Equation (1) (when *k*=1) was the most significant. The 22 verbs have small challenging indices (Section 5.1), so providing only one candidate allowed considerably fewer chances to include the correct answers.

Although we did not define the challenging index of verbs based on their numbers of possible translations, comparing the corresponding numbers in Table 10 and Table 5 suggest that the challenging verbs also have more possible translations in the NTCIR data. (Having more possible ways to translate the word made it relatively difficult for computer algorithms to translate correctly.)

Table 11. Average ranks of the answers

Ranking	*k*=1	*k*=3	*k*=5
Eq(1)	1.000	1.607	1.773
Eq(2)	1.000	1.365	1.373
Eq(3)	1.000	1.374	1.383
Eq(4)	1.000	1.394	1.400

Table 12. Inclusion rates (combined methods)

Inclusion	*k*=1	*k*=3	*k*=5
Eq1	0.449	0.865	0.923
Eq2+Eq1	0.512	0.896	0.940
Eq3+Eq1	0.503	0.894	0.940
Eq4+Eq1	0.508	0.900	0.942

Table 13. Average ranks of the correct answers (combined methods)

Ranking	*k*=1	*k*=3	*k*=5
Eq1	1.000	1.607	1.773
Eq3	1.000	1.374	1.383
Eq2+Eq1	1.000	1.537	1.662
Eq3+Eq1	1.000	1.546	1.677
Eq4+Eq1	1.000	1.547	1.664

Corresponding numbers in Table 6 and Table 11 support the claim that translating the 22 challenging words is more difficult. The average ranks of the answers became worse in Table 11.

Data in Tables 12 and 13 repeat the trends that we observed in Tables 7 and 8. Using the combined methods allowed us to answer all test cases and improved both the inclusion rates and the average ranks of the answers.

If we built a computer-assisted translation system that recommends the top *k* possible translations for these 22 verbs, the performance would not be as good as what we could achieve by building a system for the top 100 verbs. When the system suggested the leading 3 translations (*k*=3), the inclusion rates dropped to around 0.90 in Table 12 from 0.96 in Table 7.

Again, using either the English nouns or the Chinese nouns, along with the English verbs, in the conditions of the methods listed in Table 3 did not result in significant differences. When we replaced Equation (2) with Equation (4), or *vice-versa*, in the experiments, we observed very similar results in Tables 12 and 13 most of the time.

6.4 Translating English Nouns

We repeated the experiments that we discussed in Sections 6.1, 6.2, and 6.3 for the top 100 nouns in the PatentMT data. The top 100 nouns appeared in 19,756 VN pairs. The word “method” was the most frequent object in the VN pairs, and it appeared 982 times. For experiments with these nouns, we had 15,804 training instances and 3,952 test instances.

Table 14. Translation decisions for nouns

$\arg\max_{CN_i} \Pr(CN_i \mid EN)$	(8)
$\arg\max_{CN_i} \Pr(CN_i \mid EV, EN)$	(9)
$\arg\max_{CN_i} \Pr(CN_i \mid EV, EN, CV)$	(10)
$\arg\max_{CN_i} \Pr(CN_i \mid EN, CV)$	(11)

Table 15. Average ranks of the answers for translating the nouns

Ranking	*k*=1	*k*=3	*k*=5
Eq(8)	1.000	1.171	1.223
Eq(9)	1.000	1.118	1.138
Eq(10)	1.000	1.104	1.125
Eq(11)	1.000	1.116	1.142

The goal was to find the best Chinese translation of the English objects, given its collocational and bilingual information. The structure of the experiments was analogous to what we have reported for the experiments for finding the best translations of English verbs. More specifically, in addition to the English verbs and the English nouns, we were interested in whether providing the Chinese translations of the English verbs would help us improve the translation quality of the English objects. Hence, the translation decisions that we listed in Table 3 became those in Table 14.

The statistics showed analogous trends that we discussed in the previous sections. Namely, the availability of the Chinese translations of the English verbs was useful but did not help significantly when we already considered the English verbs and objects in the translation decisions, so we do not show all of the tables for the results in this paper. The rejection rates observed when we used Equations (8) through (11) were 0, 0.126, 0.184, and 0.128, respectively. The average ranks of the correct answers for the English nouns are listed in Table 15.

6.5 Experiments using Aligned Sentences in *Scientific American*

Scientific American is a magazine for introducing scientific findings to the general public. The writing style is close to ordinary life. We ran our sentence aligner (Tien *et al*., 2009) to extract aligned sentences from 1,745 articles that were published between 2002 and 2009 in the bilingual version of *Scientific American*[13]. We extracted 63,256 pairs of sentence pairs and ran the procedure depicted in Figure 1 over this set of sentence pairs to obtain 4,814 VN pairs. This scale of experiment is smaller than with the PatentMT corpus.

Since we had only 4,814 VN pairs, we chose only the 25 most frequent verbs in the experiments. This selection further reduced available VN pairs to only 1,885 pairs. With an 8:2 split for training and test data, we had only 1,508 training instances and 377 test instances. The procedure for the experiments was the same as reported in Sections 6.1 and 6.2. Again, the observed statistics indicated that using the Chinese translations of the English objects helped the translation quality of the English verbs, but the improvement was not significant. An incidental observation was that it was harder to find good translations of English verbs in *Scientific American* than in the PatentMT corpus. When providing five recommendations (*k*=5), only about 88% of the time the recommendations of our system could include the correct translations. In contrast, we had achieved inclusion rates well above 90% in Tables 7 and 12 in the experiments that used PatentMT corpus.

7. A Comparison with Human Performance

Using equations listed in Table 3 and Table 14 to make translation decisions posed a serious constraint on the available information for achieving good translations. A good translator would check a larger context to select the best translations. What would ordinary people achieve if they were provided the same limited information that our systems were provided?

To explore this interesting question, we recruited 52 human subjects who were Computer Science majors at the time of testing. Some of them were undergraduates, and some were graduate students. We placed them into three groups for three different tests: 17, 19, and 16 subjects in Test 1, Test 2, and Test 3, respectively. No human subject participated in different tests because the test questions were similar.

We chose 10 instances of verb translations from our *Scientific American* corpus, and converted each of them into three different formats for different tests. These 10 verbs were among the 25 most frequent verbs in the aligned VN pairs in our *Scientific American* corpus.

[13] http://sa.ylib.com/

Table 16. A sample question for Test 1 and Test 2

English sentence	Investigators are, of course, also exploring additional avenues for **improving** efficiency; as far as we know, though, those other approaches generally extend existing methods.
Chinese sentence	當然，研究人員也在尋找其他可______效率的方法，但就我們目前所知，其他方法一般只是延伸現有的途徑罷了。
Available choices	(1) 增進 (2) 提高 (3) 改進 (4) 改善
Possible translations and their frequencies for “improve” in *Scientific American*	improve={利用=1, 增加=1, 改良=1, 運用=1, 使=2, 加強=3, 提高=4, 改進=4, 增進=11, 改善=22}

Table 17. A sample question for Test 3

Test question	**improve** efficiency: ______ 效率
Available choices	(1) 增進 (2) 提高 (3) 改進 (4) 改善

The formats varied in the information available to the translators. Table 16 shows a test instance for Test 1. In this test, the human translators were provided 10 test instances. In each test instance, there was (1) a complete English sentence with a highlighted verb; (2) a partially translated Chinese sentence for the English sentence, with the translation for the highlighted English verb removed; and (3) four candidate Chinese verbs to be used to translate the highlighted English verb. The candidate Chinese verbs, listed in the row of “Available choices,” were selected from the translations of the highlighted English verbs in our corpus. The very last row shows the complete list of the translations for “improve” in our corpus, but this list was not provided to the human subjects.

In Test 2, the human subjects had to respond to 10 test instances. The format was the same as that for Test 1, except that the candidate Chinese verbs were not provided. The human subjects had to fill in the blanks in the Chinese sentences in Test 2.

Table 17 shows a test question for Test 3. In Test 3, the human subjects would also have to respond to 10 test questions, and they only saw the English verb, the English object, and the Chinese translation of the English object. The subjects had to choose the best translation from the list of candidate translations.

The human subjects could take their time to respond to 10 questions in the tests. There were no time limits. They usually turned in their responses within a short time, but they did not always respond to all questions. Correctness of their responses was judged based on the actual translations in *Scientific American*, even when other alternatives were also reasonable for the test questions. The sample question shown in Table 17 is an obvious example. In this example, all four translations are reasonable Chinese verbs to go with the Chinese noun. That was because there was no contextual information in Table 17 to distinguish the subtle differences between the candidate translations. Nevertheless, the original sentence pairs, shown in Table 16, were translated in exactly one way among the alternatives. Therefore, only one of the choices was considered correct.

Table 18. Average correct rates of human subjects and Equation (3)

	Human Subjects	Equation (3)
Test 1	0.524	0.600
Test 2	0.342	0.600
Test 3	0.395	0.600

We applied Equation (3), *k*=1, in Table 3 in this experiment. The average correct rates achieved by the human subjects and our programs in three tests are collected in Table 18. The correct rate is the portion of test questions with correct responses. More specifically, questions that were not answered were considered incorrect responses, and this principle applied to both human translators and our programs. Our programs made decisions only based on the English verbs, the English nouns, and the Chinese nouns in all tests. Hence, its performance was 0.6 and remained the same in all of the tests. In contrast, the average correct rates achieved by the human subjects varied with the difficulty of the tests. The human subjects performed best in Test 1, partially because they were offered more information to make decisions. Test 2 was the most difficult one, because the subjects had to provide Chinese translations themselves on the fly. The difficulty of the test questions in Test 3 was similar to those in Test 2, but the human subjects were provided with candidate translations, so the average correct rate was higher.

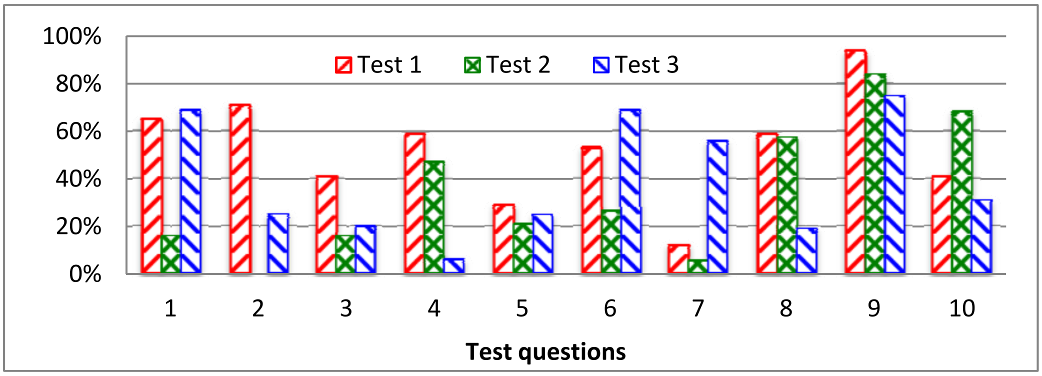


Figure 7. Average correct rates of the human translators

Figure 7 shows the average correct rates for individual questions in the three tests. The averages were computed based on the responses of the human subjects who participated in the tests. Although the average correct rates listed in Table 18 corresponded approximately to the average difficulty levels of the test formats, the performance of human subjects varied with the individual test questions. In Table 18, the average correct rate for Test 1 is the highest. In Figure 7, we can see that the correct rates for questions used in Test 1 did not always exceed those for the corresponding questions used in Test 2 and Test 3.

We do not mean to interpret results of these simple tests as a competition between human beings and computers. The results, however, suggest that translating English verbs based on partial information, *i.e.*, the English verb, the English noun, and the Chinese noun can be difficult for human subjects. The average correct rates can be seriously impacted when we insisted that there was exactly one correct answer for a test question, where the answer was defined based on the original corpus.

A previous reviewer of our work contended that we should treat all of the candidate Chinese translations in Table 16 as correct answers. Although that is a reasonable consideration, when we evaluate a system with a considerable number of test questions, doing so would require a non-negligible amount of human intervention. One possible approach might be to create an evaluation system that considers “acceptable answers” while comparing the outputs of a decoder and the expected translations.

8. More Discussion

We discuss some issues raised by anonymous reviewers in this section.

One reviewer questioned the use of the Stanford parser for both English and Chinese material, and wondered whether we should have used the CKIP parser[14] for Chinese. The point was brought up because the CKIP parser may be more reliable than the Stanford parser for Chinese.

While we agree with the reviewer about the reliability of the CKIP parser, we chose to employ the Stanford parser for both languages for two reasons at the time of our implementation. The first reason was that we needed the parsers to provide not just parse trees but also dependency relationships between words, *i.e.*, the `dobj` relationship. Using the same parser for both languages made our processing more efficient. The second reason was that the Stanford parser is an open system, so we can download the parser and parse our text on our computers. In contrast, we have to submit text material to the CKIP server for services. For copyrighted material, we were not sure that it was appropriate to rely on the CKIP services.

A concern was about how we deal with the forms of English words, *e.g.*, the tenses of verbs, in the translation of the VN pairs. The tenses of English verbs carry information about when the actions were taken, so are crucial for quality translation. Nevertheless, when we generated the VN pairs from the NTCIR corpus (Figure 1), we lemmatized the English words.

[14] http://godel.iis.sinica.edu.tw/CKIP/parser.htm

Hence, the current work, as the reviewers have noticed, did not aim at choosing the correct morphological forms for the English verbs. Similarly, we did not attempt to choose the singular and plural forms for nouns either. This issue should be tackled in further studies.

Table 19. Frequencies of 22 most "challenging" English verbs

Verb	make	exhibit	add	represent	retain	leave	enhance	reduce	lack	improve	achieve
	114	103	138	373	131	61	178	774	47	322	329
Verb	employ	reach	create	give	replace	take	apply	adjust	obtain	carry	explain
	135	119	201	70	53	210	50	69	329	241	54

Another question was about how the selection of verbs (or nouns) influences the general implication of our experimental results. Namely, how general are our results? Table 19 shows the frequencies of the 22 most challenging verbs. Evidently, the sample sizes of these verbs were not as large as those of the 20 most frequent English verbs in our dataset (*cf.* Table 1). Nevertheless, most of them were frequent enough for conducting experiments.

The resulting differences between choosing the most frequent verbs and the most challenging verbs were discussed in Section 6.3. When using the most challenging ones, the most noticeable changes were that it became more difficult to recommend the best translations of the verbs with the same number, *i.e.*, *k*, of recommendations. The inclusion rates dropped, *cf.* Table 4 and Table 9, especially when we recommended only one candidate translation. The ranks of the true answers worsened as well, *cf.* Table 6 and Table 11.

We believe that the changes observed in the experimental results are general because of the definition of degrees of challenging index (*cf.* Section 5.1). A word is more challenging if its most frequent translation is not significantly more frequent than its second frequent translation. Hence, using the challenging words made it more difficult to achieve good translations, given the same contextual information and the same number of recommended translations.

The presentation of the human performance triggered some questions. The first one was about the answers to the tests. The test item in Table 17 shows a confusing example, in which some distractors are acceptable to native speakers. Hence, a natural question is about how a "correct" answer was defined.

We touched upon this question at the end of Section 7. Apparently, some distractors are acceptable to native speakers, and some of them should have been considered correct. Nevertheless, when we evaluated a computer program, we normally had one correct answer in the test data. Even though the computer program "knew" a lot of acceptable synonyms of the correct answer, it still has to find "the" answer to be considered "correct" in the evaluation. The example shown in Table 17 is such an example. To make the computers and human subjects be evaluated on the same basis, we allowed only one answer from the available choices. The available choices came from the training data, and the answer for a test item was based on the original English and Chinese sentence pair.

When the human subjects were given contextual information in Test 1 in Section 7, they did not perform very well on average. One obvious reason was because of the multiple attractive candidates, which we discussed in the last paragraph. One may also challenge the language ability of the human subjects. Indeed, we chose the human subjects from engineering majors at the levels of undergraduates and graduate students, but we did not test their language ability before the experiments. If we were to investigate machine translation problems in this line of concern, we would probably have to ask whether all available bilingual textual material were produced by qualified linguistic and domain-dependent experts. This line of work should be important for the research community.

A reviewer stated that the human subjects did not always perform better in “easier” tasks in Test 1. For instance, in the seventh question in Figure 7, the human subjects performed much better in Test 3 than in Test 1. This may be possible for a variety of reasons. For instance, without context, the “correct” answer happened to be the most frequently collocating words, and, with context, the human subjects were distracted by confusing information in the context. As a consequence, it became easier to guess the correct answer without context.

Although we believe it is informative to compare the performance of our methods and the performance of human subjects, we did not intend to design a waterproof psycholinguistic experiment in Section 7. Hence, we chose the test instances arbitrarily from the dataset, and we compared the average performance of just 52 human subjects. A more carefully-designed psycholinguistic investigation may reveal more serious details about human performance in language translation.

9. Concluding Remarks

We designed a procedure to extract and align VN pairs in bilingual corpora. The PatentMT corpus contains 1 million pairs of English and Chinese sentences, and we aligned 35,811 VN pairs. We employed the VN pairs to investigate whether the availability of the Chinese translations for nouns in English VN pairs would improve the translation quality of the English verbs. Experimental results suggest that the information about the Chinese translation of the English noun is marginally helpful when both the English verbs and English nouns are already available. Choosing the best Chinese translation of the English verb based on the constraint of its English object or based on the information about the object’s Chinese translation achieved similar results in the experiments.

Additional and analogous experiments were conducted with the PatentMT data. In these new experiments, we aimed at the translating the nouns in the English VN pairs, given different combinations of the bilingual and contextual information. Again, we observed that, after putting the English verb and the English noun in the conditions in the formulas for translation decisions (partially shown in Table 14), the Chinese translations of the English verbs did not offer much extra help.

Acknowledgments

The work was supported in part by the funding from the National Science Council in Taiwan under the contracts NSC-99-2221-E-004-007 and NSC-100-2221-E-004-014. The authors are obliged to the anonymous reviewers for their valuable comments, and are grateful to Ms. Chia-Chi Tsai and Mr. Jui-Ping Wang for their help in some important steps of data processing. We also thank Mr. Sun for his help in English.

References

Bundanitsky, A. & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1), 14-47.

Carpuat, M., Fung, P. & Ngai, G. (2006). Aligning word senses using bilingual corpora. *ACM Transaction on Asian Language Information Processing*, 5(2), 89-120.

Chang, J.-S. & Chiou, S.-J. (2010). An EM algorithm for context-based searching and disambiguation with application to synonym term alignment. *Proceedings of the Twenty Third Pacific Asia Conference on Language, Information and Computation*, 2, 630-637.

Chang, P.-C., Galley, M., & Manning, C. D. (2008). Optimizing Chinese word segmentation for machine translation Performance. *Proceedings of the ACL Third Workshop on Statistical Machine Translation*, 224-232.

Chang, Y. C., Chang, J. S., Chen, H. J., & Liou, H. C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3), 283-299.

Cheng, C. C. (2004). Word-focused extensive reading with guidance. *Selected Papers from the Thirteenth International Symposium and Book Fair on English Teaching*, 24-32. http://elearning.ling.sinica.edu.tw/WordFocused%20Extensive%20Reading%20with%20Guidance.pdf

Chuang, T. C., Jian, J.-Y., Chang, Y.-C. & Chang, J. S. (2005). Collocational translation memory extraction based on statistical and linguistic information. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(3), 329-346.

Chuang, Y.-H., Liu, C.-L., & Chang, J.-S. (2011a). Translating common English and Chinese verb-noun pairs in technical documents with collocational and bilingual information.

Proceedings of the Twenty Fifth Pacific Asia Conference on Language, Information and Computation, 493-502.

Chuang, Y.-H., Wang, J.-P., Tsai, C.-C., & Liu, C.-L. (2011b). Collocational influences on the Chinese translation of non-technical English verbs and their objects in technical documents. *Proceedings of the Twenty Third Conference on Computational Linguistics and Speech Processing*, 94-108. (in Chinese)

Chen, A., Jiang, H., & Gey, F. (2000). Combining multiple sources for short query translation in Chinese-English cross-language information retrieval. *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, 17-23.

Chen, K.-J., Huang, S.-L., Shih, Y.-Y., & Chen, Y.-J. (2005). Extended-HowNet: A representational framework for concepts. *Proceedings of the 2005 IJCNLP Workshop on Ontologies and Lexical Resources*, 1-6.

Dorr, B. J., Levow, G.-A., & Lin, D. (2002). Construction of a Chinese-English verb lexicon for machine translation and embedded multilingual applications. *Machine Translation*, 17, 99-137.

Klein, D. & Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the Forty First Meeting of the Association for Computational Linguistics*, 423-430.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 48-54.

Lapata, M. & Brew, C. (2004). Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1), 45-73.

Lu, B., Tsou, B. K., Jiang, T., Kwong, O. Y., & Zhu, J. (2010). Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT. *Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 79-86.

Lü, Y. & Zhou, M. (2004). Collocation translation acquisition using monolingual corpora. *Proceedings of the Forty Second Annual Meeting on Association for Computational Linguistics*, 167-174.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. *Proceedings of the Fifth International Conference of the Language Resources and Evaluation*, 489-492.

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

Seneff, S., Wang, C., & Lee, J. (2006). Combining linguistic and statistical methods for bi-directional English Chinese translation in the flight domain. *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*, 213-222.

Tien, K.-W., Tseng, Y.-H., & Liu, C.-L. (2009). Sentence alignment of English and Chinese patent documents. *Proceedings of the Twenty First Conference on Computational Linguistics and Speech Processing*, 85-99. (in Chinese)

Tseng, H., Chang, P.-C., Andrew, G., Jurafsky, D., & Manning, C. D. (2005). A conditional random field word segmenter. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 168-171.

Yokoama, S. & Okuyama, M. (2009). Translation disambiguation of patent sentences using case frames. *Proceedings of the Third Workshop on Patent Translation*, in Machine Translation Summit XII, 33-36.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 29-44

聲符部件排序與形聲字發音規則探勘

Phonetic Component Ranking and Pronunciation Rules Discovery for Picto-Phonetic Chinese Characters

張嘉惠*、林書彥*、蔡孟峰*、李淑萍+、廖湘美+、黃鍔#

Chia-Hui Chang, Shu-Yen Lin, Meng-Feng Tsai, Shu-Ping Li, Hsiang-Mei Liao, and Norden E. Huang

摘要

近年來台灣有相當多的新移民的加入，這些新移民在口語的學習上雖然有地利之便，但是在漢字的認識上則是相當弱勢。由於漢字乃是圖形文字，學習單一字的成本相對的高。如果可以讓漢字教一個字，可以學到十個字，對於漢字教學的成效應有相當的助益。本文從部件教學的概念出發，考慮聲符的發音強度、出現頻率、及筆劃數，做爲聲符部件教學順序的準則。我們利用部件發音強度(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010)，以線性加總、幾合乘積、及調和平均三種方法對部件排序。根據此部件排序學習，前五個部件便可延伸學習多達 140 個相似發音的漢字。進一步，我們應用中研院文獻處理實驗室所建立的「漢字構形資料庫」，以及標記所得之形聲字，拆解形聲字組成的部件，挖掘串連漢字之間關係的形音關聯規則。我們從 600 萬條發音規則中篩選與分群出 3 組高信賴度與 5 組高支持度的規則，並藉由這些規則來輔助漢語發音的學習，提高學習效率。

關鍵詞：形聲字、聲符強度、部件教學、學習曲線、關聯規則

[*] 國立中央大學資訊工程所 Dept. of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: chia@csie.ncu.edu.tw
The author for correspondence is Chia-Hui Chang.

[+] 國立中央大學中文系 Dept. of Chinese Literature, National Central University, Taiwan

[#] 國立中央大學數據中心 Research Center for Adaptive Data Analysis, National Central University, Taiwan

Abstract

In recent years, there are a considerable number of new immigrants in Taiwan. Although these people are in the good position to learn Chinese, the advantages are limited to speaking and listening. Recognizing Chinese characters is a tough task since one has to memorize the shape, meaning and pronunciation at the same time. Therefore, the cost of learning a single character is relatively high compared with other languages in alphabet system. The goal of this study is to make the 80% pictophonetic characters to be organized more systematically such that the pronunciation of most pictophonetic characters can be inferred automatically. We evaluate the importance of Chinese components by considering the pronunciation strength, occurring frequency, and number of strokes using linear sum, product, and harmonic mean, respectively. Furthermore, we discover pronunciation rules by association mining with priority grouping. Three groups of high reliability rules and five groups of high support rules are demonstrated in this paper to show the effectiveness of pronunciation rule discovery.

Keywords: Picto-phonetic Character, Pronuciation Strength of Phonetic Component, Component-based Teaching Method, Learning Curve, Association Rule

1. 簡介

漢字是世界上最古老的文字之一，也是至今仍廣爲使用一種形系文字。近年來由於中國市場的興起，以華語做爲第二外語的學習也連帶地愈來愈受到重視，華語學習者的人數也倍數成長，據China Daily 2010的文章指出，目前全世界超過四千萬的非華裔人士正在學習華語文。由此可見未來華語文學習市場的龐大需求；再者，台灣近年來外籍與大陸配偶的人數從2002年的二十三萬人成長至今四十四萬人，其中外籍配偶約十四萬六千多人，已取得國籍者約九萬人，在在顯示了漢語學習的重要性。

過去學習漢語只能靠資深的中文老師的教導或是學習者慢慢累積經驗，不僅對於海外華語師資的培育緩不濟急，對於學習者而言更是一條漫長的路。然而，漢語字形讀音繁複，初學者並不易掌握學習要訣，尤其漢語的發音更是複雜多變。事實上華語作爲第二語言的學習，比起英文作爲第二語言的學習更是難上許多，因爲漢語的字形與音調相較拼音文字複雜，學習者要同時進行形、音、義三者的連結，如果沒有適當的學習方法，個別漢字的學習成本相當高。比起傳統的拼音拉丁文字，即使會說華語的海外華人對於漢字的認識也可能相當有限。其最主要的原因在於漢字是圖形文字(pictograph system)，無法像英文等拼音文字(alphabet system)一樣，一旦學會拼音方法(phonetic representation)，即有基本的閱讀能力。相較之下，一般漢字學習者讀寫的學習進展則會比較緩慢，而且必須搭配注音符號(Chinese phonetic symbols)或是其他拼音方法，才可知道每個漢字的發音。這樣的限制，對於漢字的學習相當不利，這也是爲什麼二十世紀初期中國大陸欲將

漢字拉丁化的主要原因。

漢字的構成包含象形、指事、會意、形聲、轉注、假借(總稱六書(許慎，1999))。據統計資料，7000 個現代漢語通用字中，屬於「形聲」結構的有 5631 個，約佔總字數的 80.5%，這麼多的形聲字在整字的組合上，多數採用「1+1」的方式，也就是一個意符加上一個聲符。基於這樣一個語言事實，我們可以借助部件教學，充分發揮部件的組合關係強化學習者對於漢字的識記。但如何折衷構字能力強度與發音強度，篩選或排序聲符部件則是本文主要探討的研究議題。

本篇論文中，我們應用(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010)，以部件發音分佈的集中性計算聲符強度，加以部件延伸字數及筆劃數的考量，提出線性加總、幾合乘積、及調和平均三種結合方法，對部件加以排序。利用此排序做爲漢字部件教學的順序，可以幫助學習者在短時間內提高閱讀效率。我們以累計延伸字個數做爲學習成效的比較，發現有效的排序，可以在學習完前五個部件，便可藉此延伸學習多達 140 個具有高度相似發音的漢字，同時累計筆劃數也是可以接受的範圍，顯示適當排序的重要性。

除了考量聲符部件學習順序之外，我們也試圖分析漢字發音規則，做爲學習發音的參考。爲了要產出易懂的發音規則，讓中文的學習者可以應用形聲字的特性來推測漢字的發音，在本文中我們應用關聯規則探勘挖掘形聲字發音所存在的規則。我們應用中研院文獻處理實驗室所建立的「漢字構形資料庫」，拆解其組成的部件，挖掘串連漢字發音關係的形音關聯規則，來輔助學習者學習，讓漢字不是教一個字才學到一個字，而能搭配關聯規則「一舉數字」，發揮數位學習的優點。我們從 600 萬條發音規則中篩選與分群出 8 條高信賴度與兩組各約 10 條高支持度的規則，並藉由這些規則來輔助漢語發音的學習效率。

2. 相關研究

最早有關漢字構造的研究，應屬中央研究院資訊科學研究所文獻處理實驗室，從 1993 年開始，陸續建構古今文字的源流演變、字形結構及異體字表，做爲記錄漢字形體知識的資料庫，也就是漢字構形資料庫(中研院文獻處理實驗室)。漢字構形資料庫不僅銜接古今文字以反映字形源流演，也記錄了不同歷史時期的文字結構。另外也由於開發漢字部件檢字系統，得以解決缺字問題。然而漢字構形資料庫過去的研究著重在字形知識的整理，尚未涉及字音與字義的處理；因此文獻處理實驗室近年來開始文字學入口網站建置計畫(莊德明、謝清俊，2005；莊德明、鄧賢瑛，2008)。一如其文所述：“漢字構形資料庫目前只著重在字形知識的整理，尚未涉及字音與字義；建立一個形、音、義俱備的漢字知識庫，仍是我們長遠的目標”。因此本論文的目的即是以挑戰漢字的發音規則知識庫爲出發，除了了解漢字發音規則外，也希望藉由此項研究找出一套形聲字發音轉換規則，讓華語學習者可以在聲符與規則的輔助下，順利讀出字的發音出來。

與本研究最爲相關的研究計畫是淡江大學中文系高柏園、郭經華、胡映雪等教授所

主持之“字詞教學模式與學習歷程研究”。其概念是藉由即時回饋的寫字練習(學文Easy Go!)，比較部件拆解做爲漢字教學策略成效(洪文斌，2010)，輔以線上教學平台「IWiLL Campus」（郭經華，2010），進行「以字帶詞」之詞彙學習策略（高柏園，2010）。此計畫在美國加州地區 Saratoga High School 針對 26 名修習 AP 中文課程之學生，實施四週約八堂之主題課程，用以評估漢字部件教學之學習策略對於海外華語文學習者之成效。從其國科會期中報告顯示，採用多媒體自習一組的學生在認字、書寫、及字的結構上，比傳統標示筆劃順序的習字方法呈現較佳的成果，顯示以部件拆解做爲漢字教學策略的可行性。

張嘉惠等人於 2010 年提出了兩種自動化判定形聲字聲符的方法(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010) ：其一是藉由聲符構件與原字的發音相似度高於非聲符構件與原字的發音相似度的概念，與語言學專家的所制訂聲母與聲母、韻母與韻母之間發音相似度，做爲第一種形聲字聲符的方法。同時也比較採用限制性最佳化技術，求得發音相似度分數。第二種方法則爲構件發聲分佈比較法，藉由聲符構件其衍生字的發聲分佈比非聲符構件的漢字發聲分佈較爲集中的概念，來計算每個構件的發聲分佈與所有漢字的發聲分佈 KL 值，做爲構件做爲聲符的強度。實驗結果顯示，發音相似度比較法在 7340 個形聲字中的判定聲符準確率爲 93.35%，而構件發聲分佈比較法則可達到 98.66%的準確率。雖然形聲字聲符的判定只是過渡性的需求，但是構件發聲強度卻可做爲學習漢字順序的重要參考準則，這也是本篇論文的重點之一。

3. 部件重要性排序

首先我們從部件教學的概念出發，希望對於聲符的教學順序，提出一個考慮聲符發音強度、出現頻率、及筆劃數的排序方法，做爲聲符部件教學順序的準則。由於構件發聲分佈比較法對於判定形聲字聲符有高達九成八的準確率，因此我們此處即採用做爲聲符發音強度。根據(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010)的定義，每一個部件的聲母發音強度、韻母的發音強度、及調號的發音強度可由下列三式計算而得：

$$I(w) = KL(P_I(W) \| P_I(A)) \quad (1)$$

$$F(w) = KL(P_F(W) \| P_F(A)) \quad (2)$$

$$T(w) = KL(P_T(W) \| P_T(A)) \quad (3)$$

其中 A 表示所有漢字所成的集合，W 則表示部件 w 所延伸的字所成的集合。函數 $P_I(A)$、$P_f(A)$、$P_T(A)$分別表示 A 集合中漢字的聲母、韻母及調的分佈機率。KL(P||Q)則代表兩個機率分佈的 KL-divergence：

$$KL(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (4)$$

對於聲符而言，由於發音集中度較高，因此 w 的聲母分佈 $P_I(W)$與所有漢字的聲母分佈 $P_I(A)$會有較大的差異。同理韻母分佈 $P_f(W)$與 $P_f(A)$差異，以及聲調分佈 $P_T(W)$與

$P_T(A)$差異也會較大。因此我們即可以 KL-divergence 公式對此差異值計算出其程度，換句話說我們利用公式 1, 2, 3 分別計算一個部件的聲母、韻母、及調號的 KL 值，這三種數值分別反應出此部件的聲母、韻母、及調號的發音強度。

除了部件的發音強度，在部件學習排序上，我們也必須考慮部件的頻率。因爲對於漢字學習者來說，發音強的部件，也要有一定的出現頻率，才能發揮其做爲聲符的功能。因此若單純以發音強度來決定教學順序，並不是非常適當的選擇。再者，對於學習者來說，漢字的筆畫數多寡也會影響學習的效率。因此如何將三者同時考慮於部件教學的順序，是此處最主要的挑戰。常見的結合方式是以線性加總，然而在此處並非最佳的結合方法，如圖一部件發音強度與頻率散佈圖顯示，若以線性加總發音強度與部件的頻率（部件頻率定義爲包含部件 w 的形聲字字數|W|除以全部字數），可能先找到的是頻率高但發音強度較弱的部件，或是發音強的部件但是頻率較低的部件，而非同時擁有高頻及高發音強度的部件。

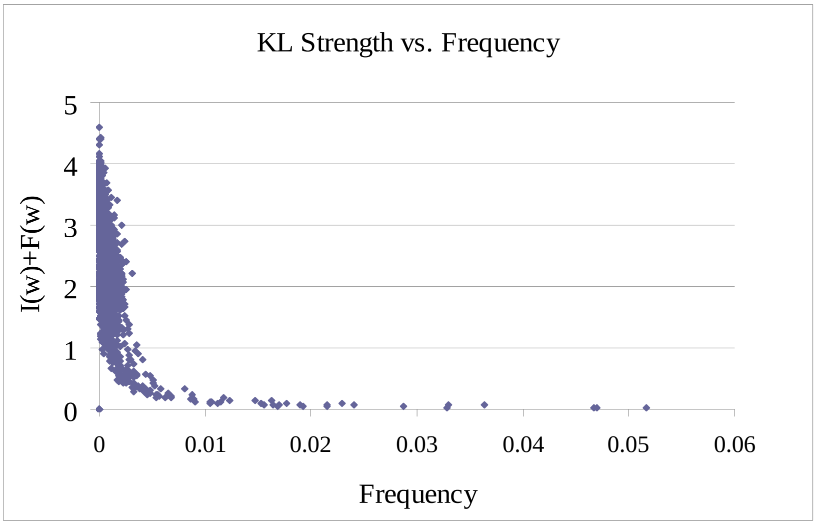


圖 1. 部件發音強度與頻率散佈圖

爲了找出頻率高且發音強度強的部件，且同時也希望能將筆劃數較少的部件優先排序。我們提出三種排序部件的依據：

1. 線性加總：$ScoreA(w)=a*Freq(w)+ I(w)+ F(w) + b*Strokes(w)$
2. 幾何乘積：$ScoreG(w)= Freq(w)*(I(w)+F(w))/\sqrt{Strokes(w)}$
3. 調和平均：$ScoreH(w)=ScoreG(w)/ScoreA(w)$

其中 *Freq(w)*代表部件 *w* 的頻率，*Strokes(w)*爲部件 *w* 的筆畫數；*a* 與 *b* 則是線性加總的權重。由圖 1 可知發音強度約爲頻率的 a=90 倍，同理，我們求得筆畫數的權重 *b=0.01*，可使線性加總的三個因素間取得平衡。第二種結合方法則是三個因素的幾何乘積，最後調和平均則是取線性加總與幾何乘積的調和平均做爲部件排序的評估。加法與乘法是結合不同因數最直接的方法，而調和平均則是取兩者的結合。

3.1 實驗評估

爲了評估三個部件排序是否能有效率地提昇學習效率，我們繪製出以幾何乘積做爲部件排序，與其累積延伸字數的關係[1]。如圖 2 所示，橫軸表示排序過的部件，從左而右依序是：分令丁方干包…等字，縱軸淺色代表累積延伸字的個數 Y_1，縱軸深色則代表聲符能正確預測聲母個數與韻母個數的總和 Y_2，兩者分別定義如下：

$$Y_1 = \Sigma_i|W_i|，\tag{5}$$

$$Y_2 = \Sigma_i\ (\mathrm{Imatch}(w_i,W_i)+\mathrm{Fmatch}(w_i,W_i))\tag{6}$$

其中 $Imatch(w_i,W_i)$代表部件 w_i 延伸字集合 W_i 中與部件 w_i 具有相同聲母的字數，$Fmatch(w_i, W_i)$代表部件 w_i 延伸字集合 W_i 中與部件 w_i 具有相同韻母的字數。舉例來說若聲符 w_i 爲包(ㄅㄠ)，若其延伸字集 W_i 爲{炮(ㄆㄠ)、胞(ㄅㄠ)、苞(ㄅㄠ)}，那麼$|W_i|=3$，而 W_i 中與 w_i 有相同聲母的字爲{胞、苞}，因此 $Imatch(w_i, W_i)=2$；而 W_i 中與 w_i 有相同韻母爲的字有{炮、胞、苞}，因此 $Fmatch(w_i, W_i)=3$。因此兩者相加後可得正確預測聲母個數與韻母個數的總和=5。

正確預測聲母個數與韻母個數的總和（Y_2）愈接近兩倍累積延伸字的個數（$2Y_1$），表示預測正確的準確率愈高，將上述兩值相除，可得準確發音比例。從圖 2 可以看出排序在前面的字即有相當多的延伸字，同時準確發音的比例也相當的高。表 1 列出排序前十個部件及其可延伸學習的形聲字，如表 1 所示，這些部件都具有延伸字發音高度相似、出現頻率高、筆數少的特性，益於先行學習。

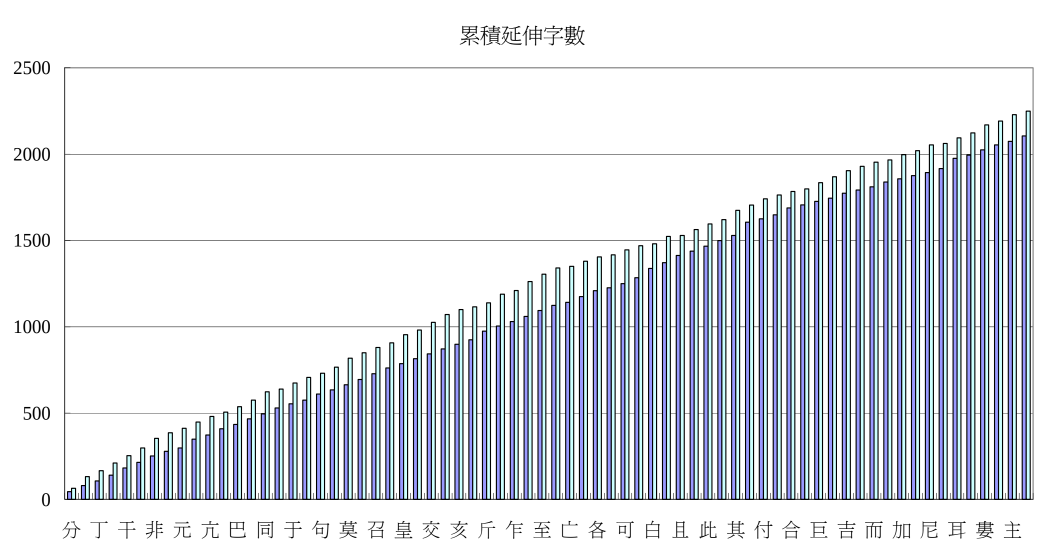


圖 2. 幾何乘積排序與累積延伸字關係

[1] 所有漢字相關資料來源則是使用中研院所開發的漢字構形資料庫。

表 1. 幾何乘積排序之部件

部件 w_i	延伸字 $\|W_i\|$	Y_1	Y_2	準確發音比例	筆劃數	累積筆劃數	延伸字
分	45	45	64	0.71	4	4	份坋坌弅吩玢枌楞棼…
令	35	80	132	0.83	5	9	伶冷玲呤囹岭狑昤泠…
丁	27	107	167	0.78	2	11	仃亭圢可叮朾宁寧玎…
方	33	140	211	0.75	4	15	仿坊彷妨枋瓬放昉防…
干	42	182	253	0.70	3	18	刊平幹杆犴旰旱汗扞…
包	32	214	298	0.70	5	23	抱胞炮砲刨匏咆庖怉…
非	38	252	353	0.70	8	31	菲啡扉緋婓腓翡徘排…
屯	26	278	386	0.69	4	35	沌肫囤鈍坉伅炖飩忳…
元	20	298	412	0.69	4	39	刓岏完妧玩杬沅忨芫…
工	51	349	448	0.64	3	42	巨仜功左巧巫差式攻…

接著我們比較三種排序方法的學習曲線如圖 3，同樣地橫軸爲部件排序，縱軸爲正確預測聲母個數與韻母個數的總和。從圖 3 中可看出幾何乘積排序較線性加總法來的有效，在學到 1000 字以前幾何乘積排序呈現大幅度的成長，也就是說若我們依照乘積排序的部件順序來學習，一開始便能達到快速學習到大量的延伸字。調和平均排序採用幾何乘積與線性加總算數平均法的調和，不過其走勢幾乎與幾何乘積排序相同，這點也顯示出幾何乘積排序明顯優於線性加總。

最後我們以累積筆畫數的學習曲線來看(圖 4)，幾何乘積排序的累積筆畫數學習曲線也較線性加總排序所得來的優異。圖 3 的收斂點與圖 4 的筆劃數大增的轉折點也顯示了在學習了 2200 個部件後，累積延伸字數已呈飽和狀態，顯示接續其後的部件已是複合部件。另外圖 4 顯示 2200 部件之後筆畫數增加速度較快，可判斷排序大於 2200 後的漢字多是較複雜的字，並不是迫切的學習對象。

4. 形聲字發音規則探勘

本文第二個重點在於形聲字發音規則的探勘，藉由已標記的形聲字聲符，找出聲符與延伸的形聲字之間是否有常見的發音規則。爲了要產出易懂的發音規則，讓漢字的學習可以應用形聲字的特性來推測字的發音，在本文中我們將應用關聯規則探勘 Apriori 演算法做爲探勘形聲字發音規則的方法。每一條關聯規則必須符合最小支持度(support)及最小信賴度(confidence)，對於學習者才算有用。以下我們首先介紹如何準備形聲字成爲關聯規則探勘所需要的交易資料，以及規則的篩選與分群，以及最終所得的發音規則。

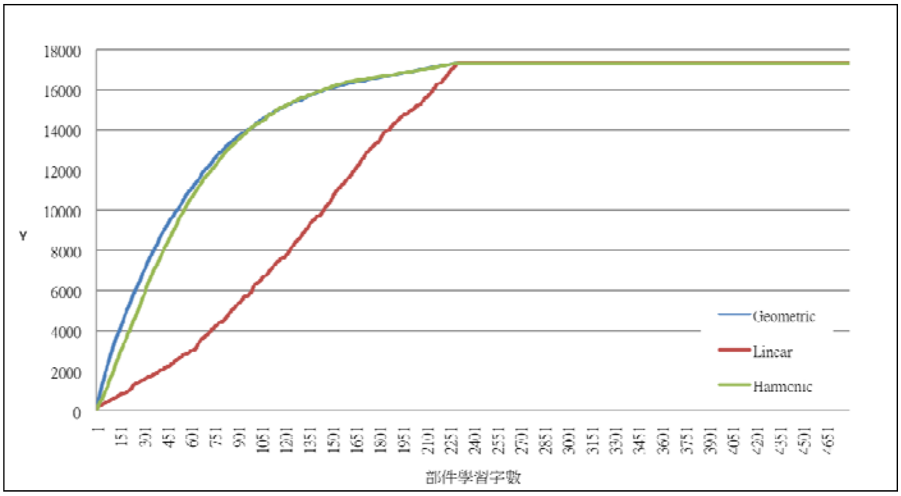


圖 3. 部件排序學習曲線比較圖

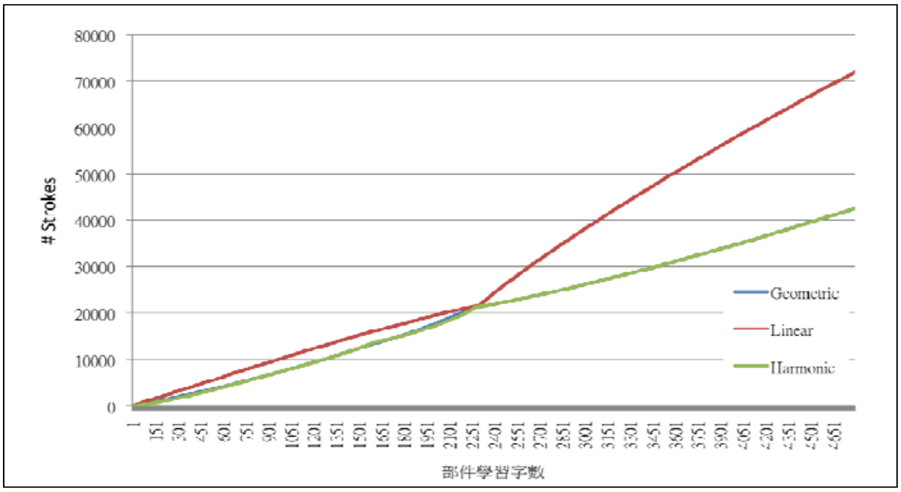


圖 4. 部件排序與筆畫數學習曲線比較圖

4.1 形聲字交易資料

關聯規則探勘原本的目的是從超市購買交易記錄的資料庫中，找出產品之間被購買的關聯程度，其主要依據爲支持度(support)及信賴度(confidence)。其中支持度代表一個規則的涵蓋率（全部交易資料中有多少百分比讓規則爲真），而信賴度則代表一個規則的準確率（前提爲真的情況下，有多少百分比資料讓結果也同時爲真）。關聯規則探勘是資料探勘領域最爲廣泛使用的工具，許多資料探勘軟體都提供此項功能，Weka[2]即是眾多資料探勘軟體其中之一。爲了推測發音規則，我們以常用字中的 3000 個形聲字準備成 3000 筆交易資料。

[2] http://www.cs.waikato.ac.nz/ml/weka/

形聲字的發音分成三個部份：聲母、韻母、以及調號，分別將其記爲INITIAL、FINAL、TONE。另將形聲字的聲符(Phonetic component)，以及聲符的發音以 PC_INITIAL、PC_FINAL、PC_TONE 三個屬性標記。其次漢字的部首(Radical component)、形聲字排列方式(單體字、左右連接、上下連接、包圍式、其他)、形聲字筆劃(Stroke)、聲符筆劃(PC_Stroke)、兩者差值(diff_STROKE)等特徵都列爲表達發音規則的探勘項目之一。最後，形聲字的發音若與其聲符的發音相同，則標記成聲母發音不變(IU)、韻母不變(FU)、音調不變(TU)等項目，做爲交易資料的一部份。

表 2. 漢字特徵對照表及”炮”的交易範例

符號	意義	數值範圍	範例:炮
INITIAL	聲母發音	{ø ,ㄅ,ㄆ,…,ㄙ}	ㄆ
FINAL	韻母發音	{ø,ㄧ,ㄨ,…,ㄦ}	ㄠ
TONE	調號	{1,2,3,4,5}	4
CONNECT	形聲字的連接方法	{單體字,左右連接,上下連接,包圍式,其他}	左右
PC	聲符	形聲字	包
PC_LOCATION	聲符所在形聲字之位置	{左,右,上,下,內,其他}	右
PC_INITIAL	聲符的聲母	{ø,ㄅ,ㄆ,…,ㄙ}	ㄅ
PC_FINAL	聲符的韻母	{ø,ㄧ,ㄨ,…,ㄦ}	ㄠ
PC_TONE	聲符的調號	{1,2,3,4,5}	1
STROKE	形聲字筆劃數 L16 表示>=16 b12-15 表示介於 12 與 15 s11 表<=11	{s11, b12-15, L16}	s11
PC_STROKE	聲符筆劃數	{s11, b12-15, L16 }	s11
Diff_STROKE	形聲字與其聲符筆劃差值	{s3, b4-5, L6}	4-5
INITIAL_UNCHANGED(IU)	形聲字聲母發音不變	{false , true}	IU=false
FINAL_UNCHANGED(FU)	形聲字韻母發音不變	{false , true}	FU=true
TONE_UNCHANGED(TU)	形聲字聲調不變	{false , true}	TU=false

值得一提的部份是，由於筆劃數乃數值性屬性，爲能運用關聯資料探勘技術，我們統計了漢字構形資料庫中所有的漢字的筆劃數將其平分爲三類，分別是筆劃小於等於 11、介於 12-15、大於等於 16。同時形聲字與其聲符筆劃差值，也就是部首的筆劃數也分爲三類，分別是筆劃小於等於 3、介於 4-5、大於等於 6。每筆形聲字交易資料所包含的項目屬性如表 2 所示。

我們取最小支持度 0.3%、 最小信賴度 60%來進行形聲字發音規則探勘。針對最小支持度取 0.3%、0.5%與 1%對應各種不同的最小信賴度 60%~100%，進行 Apriori 運算後，得到不同數量的規則數如表 3。在常見 3000 筆形聲字中，支持度 0.3% 相當於符合 9 個形聲字。更小支持度的規則由於使用率不高，因此最小信賴度設爲 60%。雖然在最小支持度 1%及最小信賴度 100%時，即可探勘出 50,054 條發音規則，然許多高支持度的規則並不具有高信賴度，爲避免錯失重要的發音規則，以上各項參數設定中，我們取最多規則數的參數組合(最小支持度 0.3%，最小信賴度 60%情形下)，共 6,625,518 條規則，做爲進一步的篩選過濾。

表 3. 關聯規則探勘後規則數

conf sup	60%	70%	80%	90%	100%
0.3%	6,625,518	5,144,742	3,879,619	2,809,951	1,810,585
0.5%	1,573,613	1,149,779	802,029	500,708	314,523
1%	304,330	217,346	143,301	87,324	50,054

4.2 規則篩選

每條關聯規則皆是由“左邊條件[左支持度]➔右邊結果[右支持度,信賴度]”組成。雖然關聯規則探勘可以取得爲數不少的發音規則，但其中有許多是不符合我們預期的規則。舉例來說：

PC_LOCATION=右 (sup=2054) ➔ CONNECT =左右 (sup=2054, conf=1)

上述這條規則表示“若聲符位置在右，則形聲字連接方式爲左右連接”。像這樣的規則對發音的推測其實並沒有幫助。又如以下規則：“若形聲字聲母發音爲ㄅ，則其聲符聲母發音爲ㄅ”，像這樣的規則也無助於推測發音。由於我們的本意是讓學習者在具備基礎聲符的閱讀能力下，利用對聲符的相關認知，來推測出更多尙未認識的形聲字發音。因此合法的規則應該具備:“聲符條件或形聲字筆劃數”➔“形聲字發音或形聲字發音之變化”。根據此一篩選原則，我們統計出最小支持度與最小信賴度不同參數下合法的規則數如表 4。最後我們存入 368,810 條規則於資料庫中。

表 4. 篩選後規則數

conf sup	60%	70%	80%	90%	100%
0.3%	368,810	272,957	195,735	152,152	106,740
0.5%	61,171	32,089	15,243	7,561	5,190
1%	13,470	6,340	1,889	505	42

4.3 規則分群

雖然在最小支持度 0.3%，最小信賴度 60%情形下，規則篩選已將的規則數減少至 368,810 筆規則，但由於規則中有許多同質性的規則散佈在資料庫中，我們需要有系統地將它們分群。以圖 5 條件集爲例，可以發現 1、2、3 具有相同條件「聲符的聲母=ㄌ」，且這些規則均具有相近的支持度。仔細深入查看符合這些條件的字後發現，支持這些規則的字組也相當程度的重疊（如「老」、「呂」、「里」等聲符的延伸字），所以聲符的聲母條件可以是分群的重要參考因素。同理聲符的韻母也多涉及相同性質的規則，因此規則中若有指定相同的聲符韻母，也是我們分群的依據之一。

另外我們也發現相同部首的規則具有相近的支持度及相同的延伸字集，因此可再結集成群。而形聲字的連接方法在具有相同部首的狀況下，通常也會有特定的連接方法如上規則{4、5}；{6、7}，因此相同部首及形聲字的連接方法也在分群條件之一。完整分群條件優先權如下：聲符、聲符聲母、聲符韻母、部首、形聲字的連接方法。根據這些分群優先條件，可將相同性質規則分爲同群。

1. 聲符的聲母=ㄌ，聲符的調=2，聲符所在位置=右，形聲字筆劃數=12-15 (sup=17)
2. 聲符的聲母=ㄌ，聲符的筆劃數=L16，漢字與其聲符筆劃差值=4-5 (sup=16)
3. 聲符的聲母=ㄌ，聲符的調=3，漢字與其聲符筆劃差值=s3 (sup=16)
4. 部首=艸，形聲字的連接方法=上下連接，support=22
5. 部首=艸，聲符所在位置=下，support=22
6. 部首=女，形聲字的連接方法=左右連接，support=15
7. 部首=女，聲符所在位置=右，support=15

圖 5. 發音規則條件範例

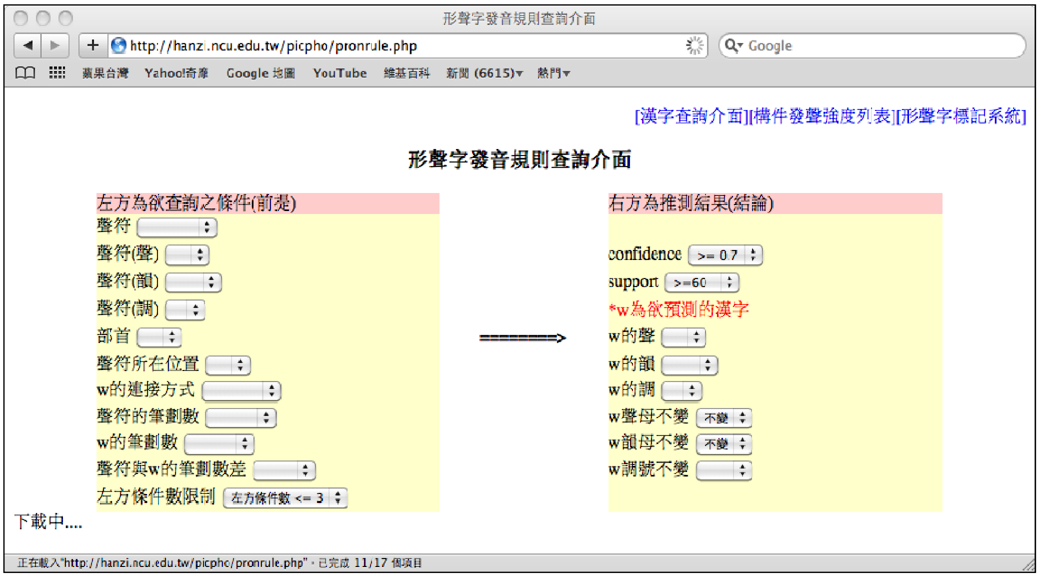
針對聲母不變(IU)及韻母不變(FU)的條件下，我們將查詢所得規則經過分群之後的所得結果呈現於表 5。如表所示，分群之後，發音關聯規則即可大幅減少，有助於規則的觀察與了解。

表 5. 符合聲母不變或韻母不變的規則數及分群後規則數

Confidence Condition	60%	70%	80%	90%	100%
IU, FU	3454/332	2004/225	1097/139	597/73	264/39
IU	9002/486	5383/383	3067/262	1758/161	809/91
FU	12171/690	8373/608	4855/470	2673/325	1392/189

4.4 關聯規則查詢介面

關聯規則的查詢介面主要是爲了教材編排者所設計，得讓使用者能根據不同條件快速篩選發音規則。「形聲字發音規則查詢系統」的設計，是採用動態呈現條件選單內容，因此第一次載入網頁時等待時間較長（約 20 秒），而後選擇條件時系統會透過 Ajax 的方式傳送搜尋條件至伺服器端擷取相關規則做爲呈現動態分群結果。查詢過程中，左下角的「下載中…」字樣會表示資料正在回傳中，右上角兩個選項則是開啓「形聲字標記系統」及「構件發聲強度列表」的連結。



[漢字查詢介面][構件發聲強度列表][形聲字標記系統]

形聲字發音規則查詢介面

左方為欲查詢之條件(前提)
聲符
聲符(聲)
聲符(韻)
聲符(調)
部首
聲符所在位置
w的連接方式
聲符的筆劃數
w的筆劃數
聲符與w的筆劃數差
左方條件數限制 左方條件數 <= 3

=========>

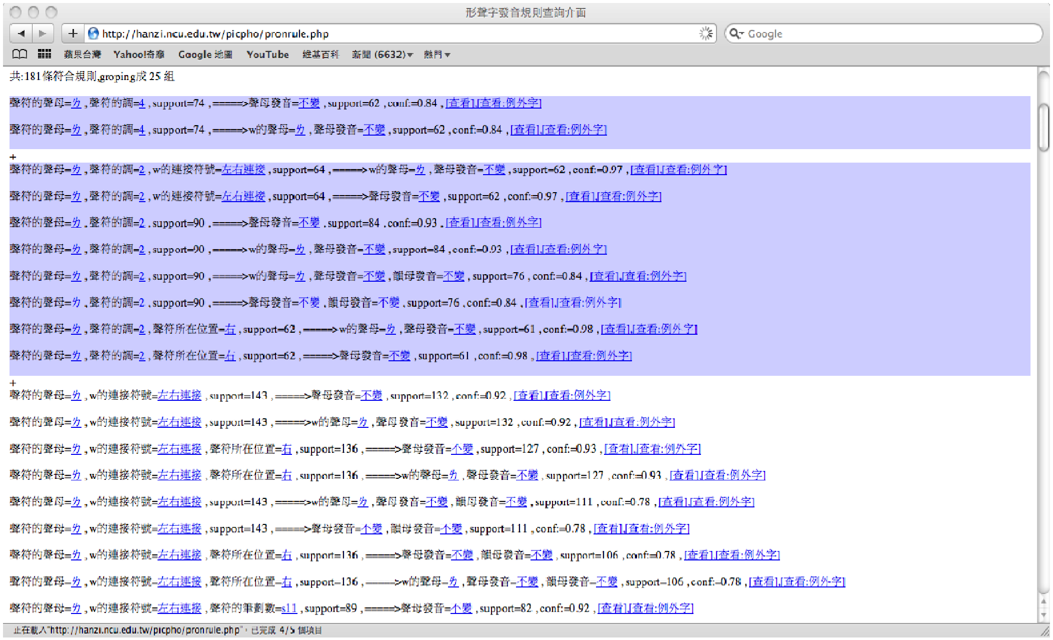
右方為推測結果(結論)
confidence >= 0.7
support >=60
*w為欲預測的漢字
w的聲
w的韻
w的調
w聲母不變 不變
w韻母不變 不變
w調號不變

下載中....

正在載入"http://hanzi.ncu.edu.tw/picpho/pronrule.php"，已完成 11/17 個項目

圖 6. 形聲字發音規則查詢介面(http://hanzi.ncu.edu.tw/picpho/pronrule.php)

查詢系統主要依據前述「形聲字發音規則探勘項目集」的特徵爲查詢條件，左邊是已知條件，右邊是推測結果，w 代表欲推測發音之形聲字。「左方條件限制」的功能則可篩選過長的規則。由於太長的規則通常不利於人們記誦，因此本系統預設條件數小於等於三。其他預設查詢條件爲「信賴度 >= 70%」，「支持度 >= 60」。當「下載完成」出現後，所有符合條件篩選的規則，經由分群後會顯示在介面的最下方，不同性質的規則會用不同底色做區隔（如圖 7）。每條規則中都有可供細項查詢的連結。當我們選擇[查看]連結，則可顯示滿足整條規則（包含左方前提及右方結果）的形聲字；除此之外，[例外字]則可查出究竟有哪些形聲字是符合左方前提，但是不符合右方結果的形聲字。其他連結則可顯示符合單一條件的形聲字，如[部首=木]連結，可顯示出所有部首爲木的形聲字，[聲符的聲母=ㄉ]可找出所有聲符聲母是ㄉ的形聲字。



```
共:181 條符合規則,groping成 25 組

聲符的聲母=ㄌ ,聲符的調=4 ,support=74 ,=====>聲母發音=不變 ,support=62 ,conf:=0.84 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=4 ,support=74 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,support=62 ,conf:=0.84 ,[查看],[查看:例外字]

+
聲符的聲母=ㄌ ,聲符的調=2 ,w的連接符號=左右連接 ,support=64 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,support=62 ,conf:=0.97 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=2 ,w的連接符號=左右連接 ,support=64 ,=====>聲母發音=不變 ,support=62 ,conf:=0.97 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=2 ,support=90 ,=====>聲母發音=不變 ,support=84 ,conf:=0.93 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=2 ,support=90 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,support=84 ,conf:=0.93 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=2 ,support=90 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,韻母發音=不變 ,support=76 ,conf:=0.84 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=2 ,support=90 ,=====>聲母發音=不變 ,韻母發音=不變 ,support=76 ,conf:=0.84 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=2 ,聲符所在位置=右 ,support=62 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,support=61 ,conf:=0.98 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,聲符的調=2 ,聲符所在位置=右 ,support=62 ,=====>聲母發音=不變 ,support=61 ,conf:=0.98 ,[查看],[查看:例外字]

+
聲符的聲母=ㄌ ,w的連接符號=左右連接 ,support=143 ,=====>聲母發音=不變 ,support=132 ,conf:=0.92 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,support=143 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,support=132 ,conf:=0.92 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,聲符所在位置=右 ,support=136 ,=====>聲母發音=不變 ,support=127 ,conf:=0.93 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,聲符所在位置=右 ,support=136 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,support=127 ,conf:=0.93 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,support=143 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,韻母發音=不變 ,support=111 ,conf:=0.78 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,support=143 ,=====>聲母發音=不變 ,韻母發音=不變 ,support=111 ,conf:=0.78 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,聲符所在位置=右 ,support=136 ,=====>聲母發音=不變 ,韻母發音=不變 ,support=106 ,conf:=0.78 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,聲符所在位置=右 ,support=136 ,=====>w的聲母=ㄌ ,聲母發音=不變 ,韻母發音=不變 ,support=106 ,conf:=0.78 ,[查看],[查看:例外字]

聲符的聲母=ㄌ ,w的連接符號=左右連接 ,聲符的筆劃數=s11 ,support=89 ,=====>聲母發音=不變 ,support=82 ,conf:=0.92 ,[查看],[查看:例外字]
```

正在載入"http://hanzi.ncu.edu.tw/picpho/pronrule.php"，已完成 4/5 個項目

圖 7. 形聲字發音規則分群結果

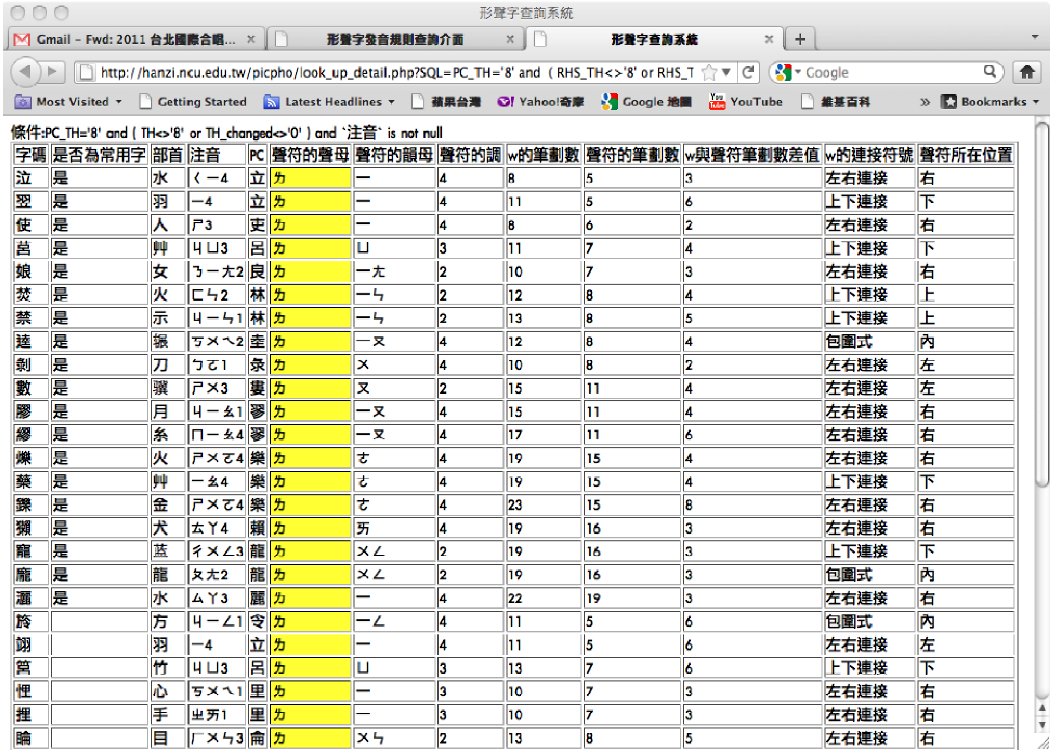
有了以上形聲字發音規則查詢系統，我們即可設定所需條件，找出相關發音規則。舉例來說，高支持度 3%、信賴度 80%且聲母發音不變的條件下的規則共有 15 條，共分成 3 組，如 R1-R3 所示。

(R1) 聲符的聲母=ㄌ (supp:197) ➔聲母發音=不變 (supp:178, conf:0.9)

(R2) 聲符的聲母=ㄇ (supp:128) ➔聲母發音=不變 (supp:105, conf:0.82)

(R3) w 的筆劃數=L16 , 聲符的筆劃數=L16 (supp:123) ➔聲母發音=不變 (supp:98, conf: 0.8)

規則一(R1)說明聲符的聲母若爲ㄌ的前提下，形聲字的聲母發音將維持ㄌ（聲符例字：力令立列老利呂良里來侖兩戾拉林坴夌彔剌郎栗留婁累連勞量廉虜雷豊劉閭厲慮樂閻魯晶歷盧賴龍闌羅麗蘭覽）；規則二(R2)顯示聲符的聲母若爲ㄇ的前提下，形聲字的聲母發音將維持 ㄇ（聲符例字：木末毛母民冊目矛名牟米免每孟明門冒某眉眇美苗面冥迷莽莫麻悶買閔滿蒙貌麼磨瞢彌）;規則三(R3)則敘述聲符的筆劃數若大或等於 16 以上，則形聲字的發音也多維持原本聲符的聲母發音（聲符例字：冀嬴歷燕盧磨穌縣羲翰蕭謁賴頻龍瞢褱嬰彌龜爵襄闌隱霜鮮鐵瞿聶轉離魏萑羅藝贊顛麗嚴藺蘇覺矍籣覽霸彎），不過第三個規則，由於筆劃數高，對於初學者來說幫助不大。除了查看例字之外，使用者也可查看例外字，了解符合前提(聲符的聲母=ㄌ)但是聲母發音卻改變的形聲字〔見圖 8〕。



條件:PC_TH='8' and (TH<>'8' or TH_changed<>'0') and `注音` is not null

字碼	是否為常用字	部首	注音	PC	聲符的聲母	聲符的韻母	聲符的調	w的筆劃數	聲符的筆劃數	w與聲符筆劃數差值	w的連接符號	聲符所在位置
泣	是	水	ㄑㄧ4	立	ㄌ	ㄧ	4	8	5	3	左右連接	右
翌	是	羽	ㄧ4	立	ㄌ	ㄧ	4	11	5	6	上下連接	下
使	是	人	ㄕ3	吏	ㄌ	ㄧ	4	8	6	2	左右連接	右
莒	是	艸	ㄐㄩ3	呂	ㄌ	ㄩ	3	11	7	4	上下連接	下
娘	是	女	ㄋㄧㄤ2	良	ㄌ	ㄧㄤ	2	10	7	3	左右連接	右
焚	是	火	ㄈㄣ2	林	ㄌ	ㄧㄣ	2	12	8	4	上下連接	上
禁	是	示	ㄐㄧㄣ1	林	ㄌ	ㄧㄣ	2	13	8	5	上下連接	上
逵	是	辵	ㄎㄨㄟ2	坴	ㄌ	ㄧㄡ	4	12	8	4	包圍式	內
剝	是	刀	ㄅㄛ1	彔	ㄌ	ㄨ	4	10	8	2	左右連接	左
數	是	攴	ㄕㄨ3	婁	ㄌ	ㄡ	2	15	11	4	左右連接	左
膠	是	月	ㄐㄧㄠ1	翏	ㄌ	ㄧㄡ	4	15	11	4	左右連接	右
繆	是	糸	ㄇㄧㄠ4	翏	ㄌ	ㄧㄡ	4	17	11	6	左右連接	右
爍	是	火	ㄕㄨㄛ4	樂	ㄌ	ㄜ	4	19	15	4	左右連接	右
藥	是	艸	ㄧㄠ4	樂	ㄌ	ㄜ	4	19	15	4	上下連接	下
鑠	是	金	ㄕㄨㄛ4	樂	ㄌ	ㄜ	4	23	15	8	左右連接	右
獺	是	犬	ㄊㄚ4	賴	ㄌ	ㄞ	4	19	16	3	左右連接	右
寵	是	宀	ㄔㄨㄥ3	龍	ㄌ	ㄨㄥ	2	19	16	3	上下連接	下
龐	是	龍	ㄆㄤ2	龍	ㄌ	ㄨㄥ	2	19	16	3	包圍式	內
灑	是	水	ㄙㄚ3	麗	ㄌ	ㄧ	4	22	19	3	左右連接	右
旍		方	ㄐㄧㄥ1	令	ㄌ	ㄧㄥ	4	11	5	6	包圍式	內
翊		羽	ㄧ4	立	ㄌ	ㄧ	4	11	5	6	左右連接	左
筥		竹	ㄐㄩ3	呂	ㄌ	ㄩ	3	13	7	6	上下連接	下
悝		心	ㄎㄨㄟ1	里	ㄌ	ㄧ	3	10	7	3	左右連接	右
捚		手	ㄓㄞ1	里	ㄌ	ㄧ	3	10	7	3	左右連接	右
睔		目	ㄏㄨㄣ3	侖	ㄌ	ㄨㄣ	2	13	8	5	左右連接	右

*圖 8. 查看發音規則例外字（不符合 **R1** 的例外字）*

又如查詢高信賴度 100%、支持度 0.5%、且聲母與韻母均未改變的規則，可得 34 條符合條件的規則，分成 5 組，如 R4– R8 所示。規則左方的支持度表示滿足左方條件的常用形聲字，規則右方的支持度則爲滿足整個規則的常用形聲字。舉例來說規則七(R7)說明聲符的聲母爲ㄒ、聲調爲一聲且聲符筆劃數小於等於 11 的時候，則衍生形聲字的聲母與韻母均不改變；符合這條規則的形聲字中包含的聲符包括「希」、「析」、「宣」、「星」、「相」、「胥」、「奚」等衍生的 16 個常用形聲字。不過使用者若是查看符合規則的形聲字，則同時可以看到其他符合條件的非常用形聲字，如「心」、「先」、「西」、「析」、「欣」、「香」、「悉」、「脩」等聲符所衍生的形聲字。規則八(R8)則說明當聲符的韻母爲ㄤ、聲調爲一聲、聲符筆劃數小於等於 11 且聲符與部首爲左右連接的時候，則衍生形聲字的聲母與韻母均不改變；符合這條規則的形聲字中包含的聲符包括「方」、「邦」、「岡」、「昌」等衍生的 17 個常用形聲字。

(R4) 聲符的聲母=ㄌ, 聲符的調=3, w與聲符筆劃數差值=s3 (supp:16)

➔聲母發音=不變, 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]

(R5) 聲符的聲母=ㄌ, 聲符的調=2, 聲符所在位置=右, w的筆劃數=12-15 (supp:17)

➔聲母發音=不變, 韻母發音=不變 (supp:17, conf:1) [查看],[例外字]

(R6) 聲符的聲母=ㄌ, 聲符的筆劃數=L16, w與聲符筆劃數差值=4-5 (supp:16)

➔聲母發音=不變, 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]

(R7) 聲符的聲母=ㄒ, 聲符的調=1, 聲符的筆劃數=s11, w的筆劃數=12-15 (supp:16)

➔聲母發音=不變, 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]

(R8) 聲符的韻母=ㄤ, 聲符的調=1, w的連接符號=左右連接, w的筆劃數=s11 (supp:17)

➔聲母發音=不變, 韻母發音=不變 (supp:17, conf:1) [查看], [例外字]

5. 結論及未來研究

本文的研究目標係提出一套以聲符爲主的部件教學策略，將構詞能力很強的部件放在課程的前面，發揮「以簡馭繁」、「快速掌握形聲字的結構」等部件教學的優點，加強學習者利用部件線索來學習新的生字的觀念，提升其於漢字識字學習上的能力。

在本篇論文中，我們延續機率分佈比較法，考慮到發音一致性強、出現頻率高且部件筆劃數少等三種因素，我們提出三種部件排序方法，其中幾何乘積法在延伸學習字數及筆劃數曲線圖的表現上較爲出色。本論文的第二部份則是藉由形聲字的特徵，運用關聯探勘法則挖掘出許多發音規則。而發音規則經由我們歸納後可分爲，高支持度與高信賴度兩大類。藉由這兩大類的規則能幫助不同程度的初學者更易於推測未知漢字的發音。

目前有關部件發音強度的計算，以及形聲字發音的關聯規則雖已完成，但是對於輔助以聲符爲主的部件教學教材編輯，仍有不足之處。舉例來說，由於關聯規則探勘可能找到相當多的規則，而且某些規則可由其他規則涵蓋，因此如何找出一組最重要的規則涵蓋愈多的常用字及將發音規則排序，則是此處我們必須要解決的問題。再者，漢字教學步驟通常爲先教獨體字，再教簡單合體字，最後教複雜合體字。但並非每個部首和任何聲符都可組成合體字，對初學者而言，可能出現偏旁部首張冠李戴的情形。如何幫助學習者釐清這些差異，也是挑戰之一。

參考文獻

許慎撰，段玉裁注(1999)。《說文解字注》，台北藝文印書館。

莊德明、謝清俊(2005)。《漢字構形資料庫的建置與應用》，漢字與全球化國際學術研討會，台北 。

莊德明、鄧賢瑛(2008)。《文字學入口網站的規畫》，第四屆中國文字學國際學術研討會，山東煙台。

董鵬程(2007)。《台灣華語文教學的過去、現在與未來展望》，多元文化與族群和諧國際研討會，台北教育大學。http://r9.ntue.edu.tw/activity/multiculture_conference/memoirs.html。

許聞廉、呂明蓁、胡志偉、柯華葳、辜玉旻、呂菁菁、張智凱、莊宗嚴(2009-2011)。《構建一個新移民者有機成長的多元認同平台的整合研究（期中進度報告）》。

高柏園、郭經華、胡映雪(2009-2010)。《華語文作爲第二語言之字詞教學模式與學習歷程研究》。

洪文斌(2010)。《華語文作爲第二語言之字詞教學模式與學習歷程研究－ － 子計畫一：中文字部件拆解教學模式與電腦輔助學習系統之研發（期中進度報告）》。

張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黄鍔(2010)。《以最佳化及機率分佈判斷漢字聲符之研究》, Computational Linguistic and Chinese Language Processing, 15(2), 145-160。

萬雲英，《兒童學習漢字的心理特徵與教學》，載於楊中芳、高尚仁主編，中國人、中國心－發展與教學篇，403-448。台北：遠流。

盛繼豔，《華文教學中漢語的部件教學》。

梁彥民(2004)。《漢字部件區別特徵與對外漢字教學》，語言教學與研究。

李思維、王昌茂編著(2000)。《漢字形音學》，武漢：華中師範大學出版社。

中研院文獻處理實驗室，「漢字構形資料庫」，http://cdp.sinica.edu.tw/cdphanzi/。

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 45-86

Enhancement of Feature Engineering for Conditional Random Field Learning in Chinese Word Segmentation Using Unlabeled Data

Mike Tian-Jian Jiang[∗+], Cheng-Wei Shih[†+], Ting-Hao Yang[+],

Chan-Hung Kuo[+], Richard Tzong-Han Tsai[‡] and Wen-Lian Hsu[∗†+]

Abstract

This work proposes a unified view of several features based on frequent strings extracted from unlabeled data that improve the conditional random fields (CRF) model for Chinese word segmentation (CWS). These features include character-based *n*-gram (CNG), accessor variety based string (AVS) and its variation of left-right co-existed feature (LRAVS), term-contributed frequency (TCF), and term-contributed boundary (TCB) with a specific manner of boundary overlapping. For the experiments, the baseline is the *6-tag*, a state-of-the-art labeling scheme of CRF-based CWS, and the data set is acquired from the 2005 CWS Bakeoff of Special Interest Group on Chinese Language Processing (SIGHAN) of the Association for Computational Linguistics (ACL) and SIGHAN CWS Bakeoff 2010. The experimental results show that all of these features improve the performance of the baseline system in terms of *recall*, *precision*, and their harmonic average as F_1 *measure score,* on both accuracy (F) and out-of-vocabulary recognition (F_{OOV}). In particular, this work presents compound features involving LRAVS/AVS and TCF/TCB that are competitive with other types of features for CRF-based CWS in terms of F and F_{OOV}, respectively.

Keywords: Conditional Random Fields, Word Segmentation, Accessor Variety, Term-contributed Frequency, Term-contributed Boundary.

[∗] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan.

[†] Institute of Information System and Application, National Tsing Hua University, Hsinchu, Taiwan.

[‡] Department of Computer Science & Engineering, Yuan Ze University, Taoyuan, Taiwan.
E-mail: thtsai@saturn.yzu.edu.tw

[+] Institute of Information Science, Academia Sinica, Taipei, Taiwan.
E-mail: {tmjiang, dapi, tinghaoyang, laybow, hsu}@iis.sinica.edu.tw

1. Introduction

Background

Many intelligent text processing tasks, such as information retrieval, text-to-speech, and machine translation assume the ready availability of a tokenization into words, which is relatively straightforward in languages with word delimiters (*e.g.*, space) but is a little difficult for Asian languages, such as Chinese and Japanese.

Chinese word segmentation (CWS) has been an active area of research in computational linguistics for two decades. SIGHAN, the Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics, has conducted five word segmentation bakeoffs (Emerson, 2005; Jin & Chen, 2007; Levow, 2006; Sproat & Emerson, 2003; Zhao & Liu, 2010). After years of intensive research, CWS has achieved high accuracy, but the issue of out-of-vocabulary (OOV) word recognition remains.

The State of the Art of CWS

Traditional approaches for CWS adopt a dictionary and rules to segment unlabeled texts, such as the work of Ma and Chen (2003). In recent years, there has been a potent trend of using statistical machine learning models, especially the conditional random fields (CRF) (Lafferty *et al.*, 2001), which displays moderate performance for the sequential labeling problem and achieves competitive results with character-position based methods(Zhao *et al.*, 2010).

Unsupervised Feature Selection for CWS

In this work, unsupervised feature selection for CWS is based on frequent strings that are extracted automatically from unlabeled corpora. For convenience, these features are referred to as *unsupervised features* in the rest of this paper. Unsupervised features are suitable for closed training evaluation where external resources or extra information is not allowed, especially for cross-domain tasks, such as SIGHAN CWS bakeoff 2010(Zhao & Liu, 2010). Without proper knowledge, the closed training evaluation of word segmentation can be difficult with OOV words, where frequent strings collected from the test data may help. For incorporating unsupervised features into character-position based CRF for CWS, Zhao and Kit (2007) tried strings based on *accessor variety* (AV), which was developed by Feng *et al.* (2004), and based on *co-occurrence strings* (COS). Jiang *et al.* (2010) applied a feature similar to COS, called *term-contributed boundary* (TCB).

According to Zhao and Kit (2007), AV-based string (AVS) is one of the most effective unsupervised features for CWS by character-position based CRF. One motivation here is to seek deeper understanding of AVS's success. This work suspects that, since AVS is designed to keep overlapping substrings via the outer structure of a string while COS/TCB is usually selected via the inner structure of a string with its longest-first (*i.e.*, non-overlapping) nature before integration into CRF, combining overlapping and outer information with non-overlapping and inner information may enhance CRF-based CWS. Hence, a series of experiments is conducted to examine this hypothesis.

The remainder of the article is organized as follows. Section 2 briefly introduces CRF. Common unsupervised features based on the concept of frequent strings are explained in Section 3. Section 4 discusses related works. Section 5 describes the design of the labeling scheme and feature templates, along with a framework that is able to encode those overlapping features in a unified way. Details about the experiment are reported in Section 6. Finally, the conclusion is presented in Section 7.

2. Conditional Random Fields

Conditional random fields (CRF) are undirected graphical models trained to maximize a conditional probability of random variables X and Y, and the concept is well established for the sequential labeling problem (Lafferty *et al.*, 2001). Given an input sequence (or observation sequence) $X = x_1 ... x_T$ and a label sequence $Y = y_1 ... y_T$, a conditional probability of linear-chain CRF with parameters $\Lambda = \lambda_1 ... \lambda_n$ can be defined as:

$$P_\lambda(Y \mid X) = \frac{1}{Z_\mathrm{X}} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, X, t) \right) . \quad (1)$$

where Z_X is the normalization constant that makes probability of all label sequences sum to one; $f_k(y_{t-1}, y_t, X, t)$ is a feature function which is often binary valued, but can be real valued; and λ_k is a learned weight associated with feature f_k.

The feature functions can measure any aspect of state transition $y_{t-1} \rightarrow y_t$, and the entire observation sequence *X* is centered at the current position *t*.

Given the model defined in (1), the most probable labeling sequence for an input sequence *X* is as follows:

$$y^* = \underset{Y}{\operatorname{argmax}}\, P_\Lambda(Y \mid X) . \quad (2)$$

Equation (2) can be efficiently calculated by dynamic programming using the Viterbi algorithm. More details about the concepts of CRF and learning parameters could be found in Wallach (2004). For sequential labeling tasks, like CWS, a linear-chain CRF is currently one of the most popular choices.

3. Unified View via Frequent String

3.1 Character-based *N-gram*

The word boundary and the word frequency are the standard notions of frequency in corpus-based natural language processing. Word-based *n*-gram is an intuitive and effective solution of language modeling. For languages without explicit word boundaries, such as Chinese, character-based *n*-gram (CNG) is usually insufficient. For example, consider some sample texts in Chinese:

- "自然科學的重要性" (the importance of natural science), and
- "自然科學的研究是唯一的途徑" (natural science research is the only way),

where many character-based *n*-grams can be extracted, but some of them are out of context, such as "然科" (so; discipline) and "學的" (study; of), even when they are relatively frequent. For the purpose of interpreting overlapping behavior of frequent strings, however, character-based *n*-grams could still be useful for baseline analysis and implementation.

3.2 Reduced *N*-gram

The lack of correct information about the actual boundary and frequency of a multi-character/word expression's occurrence has been researched in different languages. The distortion of phrase boundaries and frequencies was first observed in the Vodis Corpus, where the word-based bigram "RAIL ENQUIRIES" and word-based trigram "BRITISH RAIL ENQUIRIES" were estimated and reported by O'Boyle (1993) and Ha *et al.* (2005). Both of them occur 73 times, which is a large number for such a small corpus. "ENQUIRIES" follows "RAIL" with a very high probability when "BRITISH" precedes it. When "RAIL" is preceded by words other than "BRITISH," however, "ENQUIRIES" does not occur, but words like "TICKET" or "JOURNEY" may. Thus, the bigram "RAIL ENQUIRIES" gives a misleading probability that "RAIL" is followed by "ENQUIRIES" irrespective of what precedes it.

A common solution to this problem is that, if some *n*-grams consist of others, then the frequencies of the shorter ones have to be discounted with the frequencies of the longer ones. For Chinese, Lin & Yu (2011) reported a similar problem and its corresponding solution in the sense of *reduced n-gram* of Chinese characters. By excluding *n*-grams with their numbers of appearance that fully depend on other superstrings, "然科" and "學的" from the sample texts in the previous sub-section are no longer candidates of the string. Zhao and Kit (2007) described the same concept briefly as *co-occurrence string* (COS). Sung *et al.* (2008) invented a specific data structure for suffix array algorithm to calculate exact boundaries of phrase-alike string and their frequencies called *term-contributed boundaries* (TCB) and *term-contributed frequencies* (TCF), respectively, to analogize similarities and differences with the term frequencies. Since this work uses the program of TCB and TCF (namely YASA, yet another suffix array) for experiments, the family of *reduced n-gram* will be referred as TCB hereafter for convenience.

3.3 Uncertainty of Succeeding Character

Feng *et al.* (2004) proposed *accessor variety* (AV) to measure the likelihood a substring is a Chinese word. Another measurement, called *boundary entropy* or *branching entropy* (BE), exists in some works (Chang & Su, 1997; Cohen *et al.*, 2007; Huang & Powers, 2003; Tanaka-Ishii, 2005; Tung & Lee, 1994). The basic idea behind those measurements is closely related to one particular perspective of *n*-gram and information theory, *cross-entropy* or *perplexity*. According to Zhao and Kit (2007), AV and BE both assume that the border of a potential Chinese word is located where the uncertainty of successive character increases. They believe that AV and BE are the discrete and continuous version, respectively, of a fundamental work of Harris (1970), and they decided to adopt AVS as an unsupervised feature for CRF-based CWS. This work follows their choice in hope of producing a comparable study. AV of a string s is defined as:

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\} \tag{3}$$

In (3), $L_{av}(s)$ and $R_{av}(s)$ are defined as the number of distinct preceding and succeeding characters, respectively, except, when the adjacent character is absent because of a sentence boundary, the pseudo-character of sentence beginning or sentence ending will be accumulated. Feng *et al.* (2004) also developed more heuristic rules to remove strings that contain known words or adhesive characters. For the strict meaning of unsupervised feature and for the sake of simplicity, these additional rules are dropped in this study.

Since a recent work of Sun and Xu (2011) used both $L_{av}(s)$ and $R_{av}(s)$ as features of CRF, this work will apply a similar approach, which is denoted as LRAVS, to make a thorough comparison.

4. Other Related Works

4.1 Frequent String Extraction Algorithm

Besides previous works of TCB and TCF extraction (Sung *et al.*, 2008), Chinese frequent strings (Lin & Yu, 2001), and *reduced n-gram* (Ha *et al.*, 2005), which have already been mentioned, the article about a linear algorithm for *frequency of substring with reduction* (Lü & Zhang, 2005) also falls into this category. Most of these projects focused on the computational complexity of algorithms. Broader algorithms for frequent string extraction are suffix array (Manber & Myers, 1993) and PAT-tree (Chien, 1997).

4.2 Unsupervised Word Segmentation Method

Zhao and Kit have explored several unsupervised strategies with their unified goodness measurement of logarithm ranking (Zhao & Kit, 2007), including *frequency of substring with reduction* (Lü & Zhang, 2005), *description length gain* (Kit & Wilks, 1999), *accessor variety* (Feng *et al*., 2004), and *boundary/branching entropy* (Chang & Su, 1997; Cohen *et al.*, 2007; Huang & Powers, 2003; Tanaka-Ishii, 2005; Tung & Lee, 1994). Unlike the technique described in this paper for incorporating unsupervised features into supervised CRF learning, those methods usually filter out word-alike candidates using their own scoring mechanism directly as unsupervised word segmentation.

4.3 Overlapping Ambiguity Resolution

Subword based tagging of Zhang *et al.* (2006) utilizes confidence measurement. Other overlapping ambiguity resolution approaches are Naïve Bayesian classifiers (Li *et al.*, 2003), mutual information, difference of *t*-test (Sun *et al.*, 1997), and sorted table look-up (Qiao *et al.*, 2008). These works concentrate on overlapping of words according to some (supervised) standard, rather than overlapping of substrings from unsupervised selection.

5. CRF Labeling Scheme

5.1 Character Position Based Labels

In this study, the CRF label set for CWS prediction adopts the *6-tag* approach of Zhao *et al.* (2010), which achieves very competitive performance and is one of the most fine-grained character position based labeling schemes. According to Zhao *et al.* (2010), since less than 1% of Chinese words are longer than five characters in most corpora from SIGHAN CWS bakeoffs 2003, 2005, 2006, and 2008, the coverage of a *6-tag* approach should be sufficient. This configuration of CRF without additional unsupervised features is also the control group of the experiment. Table 1 provides a sample of labeled training data.

Table 1. Sample of the 6-tag labels.

Character	Label
反	*B*
而	*E*
會	*S*
欲	*B*
速	*C*
則	*D*
不	*I*
達	*E*

For the sample text “反而 (contrarily) / 會 (make) / 欲速則不達 (more haste, less speed)” (on the contrary, haste makes waste), the tag *B* stands for the beginning character of a word, while *C* and *D* represent the second character and the third character of a word, respectively. The ending character of a word is tagged as *E*. Once a word consists of more than four characters, the tag for all of the middle characters between *D* and *E* is *I*. Finally, the tag *S* is reserved specifically for single-character words.

5.2 Feature Templates

Feature instances are generated from templates based on the work of Ratnaparkhi (1996). Table 2 explains their abilities. C_{-1}, C_0, and C_1 stand for the input tokens individually bound to the prediction label at the current position. For example, in Table 1, if the current position is at the label *I*, features generated by C_{-1}, C_0, and C_1 are “則,” “不,” and “達,” respectively. Meanwhile, for window size 2, $C_{-1}C_0$, C_0C_1, and $C_{-1}C_1$ expands features of the label *I* to “則不,” “不達,” and “則達,” respectively. One may argue that the feature template should expand to five tokens to cover the whole range of the 6-tag approach; however, according to Zhao *et al.* (2010), the context window size in three tokens is effective to catch parameters of the *6-tag* approach for most strings that do not exceed five characters. Our pilot test for this case also showed that context window size in two tokens would be sufficient without a significant decrease in performance (Jiang *et al.*, 2010).

Unsupervised features that will be introduced in the next subsection are generated by the same template, except the binding target moves column by column, as listed in tables of the next subsection.

Table 2. Feature template

Feature	**Function**
C_{-1}, C_0, C_1	Previous, current, or next token
$C_{-1}C_0$	Previous and current tokens
C_0C_1	Current and next tokens
$C_{-1}C_1$	Previous and next tokens

5.3 Unified Feature Representation of CNG/AVS/TCF/TCB

To our knowledge, TCF, which is designed to fulfill a symmetrical comparison between the properties of inner pattern (CNG, TCF, or COS/TCB) vs. outer pattern (AVS) and between overlapping string (CNG, AVS, or TCF) vs. maximally matched string (COS/TCB), has not been evaluated in any previous work. In short, while the original version of COS/TCB selects the maximally matched string (*i.e.*, *non-overlapping* string) as the feature (Feng *et al.*, 2004; Jiang *et al.*, 2010; Zhao & Kit, 2007), TCF collects features of *reduced n-gram* from every character position with additional rank of likelihood converted from *term-contributed frequency*, as its name implies. To compare different types of overlapping strings as unsupervised features systematically, this work extends the previous work of Zhao and Kit (2007) into a unified representation of features. The representation accommodates both character position of a string and the string's likelihood ranked in the logarithm. Formally, the ranking function for a string *s* with a score *x* counted by CNG, AVS, or TCF is defined as:

$$f(s) = r, if\, 2^r \le x < 2^{r+1} \quad (4)$$

.

The logarithm ranking mechanism in (4) is inspired by Zipf's law with the intention to alleviate the potential data sparseness problem of infrequent strings. The rank *r* and the corresponding character positions of a string then are concatenated as feature tokens. To give the reader a clearer picture about what feature tokens look like, a sample representation, which is denoted in regex as "[*0-9*]+[*B*|*C*|*D*|*I*|*E*|*S*]" for rank and character position, of CNG, AVS, or TCF is demonstrated and explained by Figure 1 and Table 3.

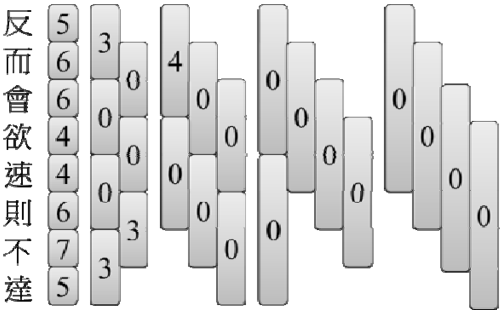


Figure 1. Example of overlapping strings with ranks.

Table 3. Sample of the unified feature representation for overlapping strings.

Input	**Unsupervised Feature**					**Label**
	1 char	2 char	3 char	4 char	5 char	
反	*5S*	*3B*	*4B*	*0B*	*0B*	*B*
而	*6S*	*3E*	*4C*	*0C*	*0C*	*E*
會	*6S*	*0E*	*4E*	*0D*	*0D*	*S*
欲	*4S*	*0E*	*0E*	*0E*	*0I*	*B*
速	*4S*	*0E*	*0E*	*0E*	*0E*	*C*
則	*6S*	*3B*	*0E*	*0E*	*0E*	*D*
不	*7S*	*3E*	*0E*	*0E*	*0E*	*I*
達	*5S*	*3E*	*0E*	*0E*	*0E*	*E*

For example, judging by strings with two characters, one of the strings “反而” gets rank *r* = 3; therefore, the column of two-character feature tokens has “反” denoted as *3B* and “而” denoted as *3E*. If another two-character string “而會” competes with “反而” at the position of “而” with a lower rank *r* = 0, then *3E* is selected for feature representation of the token at a certain position.

Note that, when the string “則不” conflicts with the string “不達” at the position of “不” with the same rank *r* = 3, the corresponding character position with rank of the leftmost string, which is *3E* in this case, is applied arbitrarily.

Although those are indeed common situations of overlapping strings, this work simply implements the above rules by Zhao and Kit (2007) for the sake of compatibility. In fact, pilot tests have been done with a more complicated representation, like *3E-0B* for “而” and *3E-3B* for “不,” to keep the overlapping information within each column, but the test result shows no significant differences in terms of accuracy and OOV recognition. Since the statistics of the pilot tests could be redundant, they are omitted in this paper.

To make an informative comparison, this work also applies the original version of *non-overlapping* COS/TCB features that is without ranks and is selected by the forward maximum matching algorithm (Feng *et al.*, 2004; Jiang *et al.*, 2010; Zhao & Kit, 2007). Table 4 illustrates a sample representation of features in this case. Notably, there are several features encoded as *-1* individually to represent that the desired string is unseen. For the *non-overlapping* siblings of the *reduced n-grams* family, such as COS/TCB, either the string is always occupied by other superstrings or it simply does not appear more than once.

Table 4. Sample of the unified feature representation for Non-overlapping COS/TCB strings.

Input	Original COS/TCB Feature	Label
反	*B*	*B*
而	*C*	*E*
會	*E*	*S*
欲	*-1*	*B*
速	*-1*	*C*
則	*-1*	*D*
不	*-1*	*I*
達	*-1*	*E*

The length of a string is limited to five characters for the sake of efficiency and consistency with the *6-tag* approach.

6. Experiments

CRF++ 0.54 (http://crfpp.sourceforge.net/) employs L-BFGS optimization and the tunable hyper-parameter (CRF++ training function argument "-c"), *i.e.*, the Gaussian prior, set to 100 throughout the whole experiment.

6.1 Data Set

The corpora used for the experiment are from the SIGHAN CWS bakeoff 2005 (Emerson, 2005) and SIGHAN CWS bakeoff 2010 (Zhao & Liu, 2010). SIGHAN 2005 comes with four different standards, including Academia Sinica (AS), City University of Hong Kong (CityU), Microsoft Research (MSR), and Peking University (PKU). SIGHAN 2010 provides a Traditional Chinese corpus and a Simplified Chinese corpus. Each corpus has training/test sets of four domains, including literature, computers, medicine, and finance, that are denoted as domains A, B, C, and D, respectively. For comparison, statistics on most corpora of SIGHAN 2003, 2006, and 2008 that have been obtained are listed in the appendix.

6.2 Unsupervised Feature Selection

Unsupervised features are collected according to pairs of corresponding training/test corpora. CNG and AVS are arranged with the help from SRILM (Stolcke, 2002). TCB strings and their ranks converted from TCF are calculated by YASA (Sung *et al.*, 2008). To distinguish the ranked and overlapping features of TCB/TCF from those of the original version of non-overlapping COS/TCB-based features, the former are denoted as TCF to indicate the score source of frequency for ranking, and the abbreviation of the later remains as TCB.

6.3 Evaluation Metrics

The evaluation metrics of CWS task are adopted from SIGHAN bakeoffs, including test *precision* (*P*), test *recall* (*R*), and their harmonic average *F_1 measure score* (*F*), as (5), (6), and (7), respectively. For performance of OOV, formulae that are similar to P/R/F are employed. To estimate the differences of performance between configurations of CWS experiments, this work uses the confidence level, which has been applied since SIGHAN CWS bakeoff 2003 (Sproat & Emerson, 2003). The confidence level assumes that the *recall (*or *precision) X of accuracy (or OOV recognition)* represents the probability that a word (or OOV word) will be identified from N words in total and that a binomial distribution is appropriate for the experiment. Confidence levels of P, R, P_{OOV}, and R_{OOV} appear in Tables 5-10 under the columns C_P, C_R, C_{Poov}, and C_{Roov}, respectively, and they are calculated at the 95% confidence interval with the formula $\pm 2 \sqrt{([X(1-X)] / N)}$. Two configurations of CWS experiments then are considered to be statistically different at a 95% confidence level if ***one of*** their C_P, C_R,

C_{Poov}, or C_{Roov} is different.

$$P = \frac{\text{the number of words thatare correctly segmented}}{\text{the number of words that are segmented}} \times 100\% \quad (5)$$

$$R = \frac{\text{the number of words that are correctly segmented}}{\text{the number of words in the gold standard}} \times 100\% \quad (6)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (7)$$

6.4 Experimental Results

The most significant type of error is unintentionally segmented alphanumeric sequences, such as English words or factoids in Arabic numerals. Rather than developing another set of feature templates for non-Chinese characters that may violate the rules of closed training evaluation, post-processing, which is mentioned in the official report of SIGHAN CWS bakeoff 2005 (Emerson, 2005), has been applied to remove spaces between non-Chinese characters in the gold standard data of the AS corpus manually, since there are no urgent expectations of correct segmentation on non-Chinese text. In SIGHAN 2005 and 2006, however, some participants used character types, such as digits, date/time specific Chinese characters, English letters, punctuation, and others (Chinese characters) as extra features, which triggered a debate of closed training criteria (Zhao *et al.*, 2010). Consequently, SIGHAN 2010 decided to allow four types of characters, distinguished as Chinese characters, English letters, digits, and punctuation. This work provides preliminary tests on non-Chinese patterns extracted from SIGHAN 2010 unlabeled training corpora A and B, extra features of character types (in character based trigram, $T_{-1}T_0T_1$, where T can be E, D, P, or C for alphabets, digits, punctuations, or Chinese characters, respectively), and their combinations to verify the performance impact of these special treatments, as shown in Table 5 –Table 8. On the one hand, the statistics indicate that the character types perform well and stably on most of the corpora. On the other hand, the features, such as AVS and TCF, may still need help from non-Chinese patterns of unlabeled training corpora A and B. As a matter of fact, our other preliminary test suggests that SIGHAN 2010 test corpora contain a lot of OOV and inconsistent segments from non-Chinese text (for example, inconsistency of usage on full-width or half-width non-Chinese characters, some English words and factoids being segmented but some of them not, *etc.*), which only can be memorized from the non-Chinese patterns. Consequently, the experimental results of SIGHAN 2010 corpora involve non-Chinese treatment based on the combination of the extra character type features and the non-Chinese patterns, but the experimental results of SIGHAN 2005 corpora do not.

Table 5. Non-Chinese treatment on SIGHAN'10 simplified Chinese corpora.

Domain	**Feature**	***P***	***C_P***	***R***	***C_R***	***F***
A	Original 6-tag	92.16	±0.002869	91.63	±0.002956	91.89
	+(Non-Chinese Pattern)	92.32	±0.002842	91.27	±0.003013	91.79
	+(Character Type)	*92.70*	±0.002777	**92.33**	±0.002840	*92.51*
	+(Non-Chinese Pattern, Character Type)	**92.71**	±0.002775	**92.33**	±0.002841	**92.52**
B	Original 6-tag	77.44	±0.004558	86.72	±0.003701	81.82
	+(Non-Chinese Pattern)	89.85	±0.003294	83.62	±0.004036	86.62
	+(Character Type)	91.68	±0.003013	**93.58**	±0.002673	**92.62**
	+(Non-Chinese Pattern, Character Type)	**92.93**	±0.002795	*91.19*	±0.003091	*92.05*
C	Original 6-tag	89.61	±0.003466	90.64	±0.003309	90.12
	+(Non-Chinese Pattern)	90.87	±0.003272	89.77	±0.003443	90.32
	+(Character Type)	*91.11*	±0.003233	**92.02**	±0.003078	**91.56**
	+(Non-Chinese Pattern, Character Type)	**91.54**	±0.003161	*91.29*	±0.003203	*91.42*
D	Original 6-tag	89.82	±0.003367	91.24	±0.003148	90.52
	+(Non-Chinese Pattern)	*93.48*	±0.002749	91.06	±0.003176	92.25
	+(Character Type)	92.35	±0.002960	**93.99**	±0.002646	*93.16*
	+(Non-Chinese Pattern, Character Type)	**93.97**	±0.002650	*93.61*	±0.002723	**93.79**

Table 6. Non-Chinese treatment OOV on SIGHAN'10 simplified Chinese corpora.

Domain	**Feature**	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
A	Original 6-tag	55.52	±0.019647	52.00	±0.019752	53.71
	+(Non-Chinese Pattern)	53.71	±0.019714	52.34	±0.019746	53.01
	+(Character Type)	**62.42**	±0.019149	*58.86*	±0.019455	**60.59**
	+(Non-Chinese Pattern, Character Type)	*61.77*	±0.019212	**59.24**	±0.019427	*60.48*
B	Original 6-tag	36.06	±0.014105	20.49	±0.011855	26.13
	+(Non-Chinese Pattern)	41.38	±0.014467	52.17	±0.014673	46.16
	+(Character Type)	**76.27**	±0.012496	*71.40*	±0.013274	**73.76**
	+(Non-Chinese Pattern, Character Type)	*67.49*	±0.013759	**76.28**	±0.012495	*71.62*
C	Original 6-tag	59.69	±0.016736	49.40	±0.017059	54.06
	+(Non-Chinese Pattern)	58.80	±0.016793	54.76	±0.016982	56.71
	+(Character Type)	**68.14**	±0.015898	*59.69*	±0.016736	**63.64**
	+(Non-Chinese Pattern, Character Type)	*66.03*	±0.016159	**60.54**	±0.016677	*63.17*
D	Original 6-tag	48.79	±0.018869	35.90	±0.018109	41.36
	+(Non-Chinese Pattern)	53.98	±0.018815	55.56	±0.018757	54.76
	+(Character Type)	**68.81**	±0.017487	*57.73*	±0.018648	*62.79*
	+(Non-Chinese Pattern, Character Type)	*68.64*	±0.017514	**66.30**	±0.017844	**67.45**

Table 7. Non-Chinese treatment on SIGHAN'10 traditional Chinese corpora.

Domain	**Feature**	***P***	***C_P***	***R***	***C_R***	***F***
A	Original 6-tag	90.63	±0.003065	88.72	±0.003326	89.66
	+(Non-Chinese Pattern)	90.73	±0.003049	88.58	±0.003344	89.64
	+(Character Type)	**92.95**	±0.002691	*92.16*	±0.002826	*92.55*
	+(Non-Chinese Pattern, Character Type)	*92.94*	±0.002693	**92.20**	±0.002819	**92.57**
B	Original 6-tag	94.52	±0.002248	93.28	±0.002474	93.90
	+(Non-Chinese Pattern)	94.12	±0.002325	91.32	±0.002781	92.70
	+(Character Type)	**96.15**	±0.001902	**95.53**	±0.002042	**95.84**
	+(Non-Chinese Pattern, Character Type)	*95.63*	±0.002019	*94.22*	±0.002307	*94.92*
C	Original 6-tag	92.95	±0.002479	91.42	±0.002712	92.18
	+(Non-Chinese Pattern)	92.69	±0.002521	90.77	±0.002803	91.72
	+(Character Type)	**94.72**	±0.002167	**93.95**	±0.002308	**94.33**
	+(Non-Chinese Pattern, Character Type)	*94.62*	±0.002186	*93.77*	±0.002341	*94.19*
D	Original 6-tag	94.06	±0.002199	93.39	±0.002312	93.72
	+(Non-Chinese Pattern)	93.85	±0.002236	92.73	±0.002416	93.28
	+(Character Type)	**95.50**	±0.001928	**95.51**	±0.001926	**95.51**
	+(Non-Chinese Patter, Character Type)	*95.48*	±0.001933	*95.34*	±0.001961	*95.41*

Table 8. Non-Chinese treatment OOV on SIGHAN'10 traditional Chinese corpora.

Domain	**Feature**	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
A	Original 6-tag	72.50	±0.015297	57.20	±0.016951	63.95
	+(Non-Chinese Pattern)	71.62	±0.015446	57.04	±0.016959	63.50
	+(Character Type)	*75.45*	±0.014745	*67.72*	±0.016017	*71.38*
	+(Non-Chinese Pattern, Character Type)	**75.60**	±0.014715	**68.44**	±0.015923	**71.84**
B	Original 6-tag	*76.46*	±0.014455	71.38	±0.015399	73.83
	+(Non-Chinese Pattern)	68.49	±0.015828	65.20	±0.016229	66.80
	+(Character Type)	**80.44**	±0.013514	**81.81**	±0.013143	**81.12**
	+(Non-Chinese Pattern, Character Type)	74.07	±0.014931	*76.40*	±0.014466	*75.22*
C	Original 6-tag	73.48	±0.015336	58.33	±0.017128	65.03
	+(Non-Chinese Pattern)	69.69	±0.015968	56.31	±0.017232	62.29
	+(Character Type)	**76.91**	±0.014641	**68.87**	±0.016087	**72.67**
	+(Non-Chinese Pattern, Character Type)	*75.97*	±0.014843	*68.18*	±0.016181	*71.87*
D	Original 6-tag	78.54	±0.013963	66.01	±0.016110	71.73
	+(Non-Chinese Pattern)	75.53	±0.014622	63.69	±0.016355	69.11
	+(Character Type)	**81.58**	±0.013184	**76.99**	±0.014315	**79.22**
	+(Non-Chinese Pattern, Character Type)	*80.64*	±0.013438	*76.22*	±0.014481	*78.37*

This empirical decision implies that CWS benchmarking corpus should be prepared more carefully to avoid unpredictable side effects from non-Chinese text. Note that the treatment does not use unlabeled training corpora A and B separately. Further discussions are mainly based on this treatment, hopefully without loss of generality and of interest for comparative studies. Numbers in bold face and italic style indicate the best and the second best results of a certain evaluation metric, respectively, except for the topline and the best record from each year of SIGHAN bakeoffs. Configurations with the same values of confidence level on *P* or *R* are underlined, but only records that have the same confidence level on ***both*** *P* and *R* should be considered as statistically insignificant, and this phenomenon did not occur in our experiment results.

Unlike the previous work, which showed a relatively clearer trend of feature selection (Jiang *et al.*, 2011), CWS performance may vary between different CWS standards and domains in this study. Considering either the best or second best records in terms of F, feature combinations consisting of LRAVS or AVS usually outperform, except on MSR of SIGHAN 2005 corpora. Nevertheless, in terms of F_{OOV}, feature combinations consisting of TCF or TCB consistently increase in performance on every corpus. Similar situations also can be recognized from the experiments on some of the SIGHAN 2003, 2006, and 2008 corpora; please refer to the appendix for details. This complicated phenomenon indicates that, since CWS studies usually struggle with incremental and small improvements, different CWS standards and/or domains can make comparative research difficult and cause experimental results of related works to be incompatible. For equipping supervised CWS with unsupervised feature selection from unlabeled data, the experimental results of this work suggests that using LRAVS+TCF with more careful non-Chinese text treatments and CRF parameter tuning (*e.g.*, more cross-validations to find a specific hyper-parameter of Gaussian prior) would be a very good choice. Nevertheless, it is still worth noting that the best performance of this work in terms of *F* is found on the best official records on traditional Chinese domain B (Computer) of SIGHAN 2010 corpora and all of the SIGHAN 2005 corpora except the PKU corpus. This is especially true when this work does not apply any special treatment of character type and non-Chinese text that many other related works do on SIGHAN 2005 corpora. Note that “Our Baseline/Topline” in the following tables indicates where official baseline/topline suffered from official release script for maximum matching malfunctions on data in UTF-8 encoding and/or some uncertain incompatibilities between obtained corpora and official ones that caused inconsistent statistics during experiment reproductions.

Table 9. Performance comparison of accuracy on SIGHAN 2005 AS corpus.

Configuration	P	C_P	R	C_R	F
6-tag	94.50	±0.001308	95.74	±0.001159	95.12
CNG	95.12	±0.001236	95.53	±0.001186	95.32
AVS	95.14	±0.001234	95.86	±0.001143	95.50
TCB	94.48	±0.001311	95.73	±0.001160	95.10
TCF	94.86	±0.001267	95.92	±0.001135	95.39
AVS+TCB	*95.21*	±0.001226	95.96	±0.001130	*95.58*
AVS+TCF	**95.27**	±0.001218	**96.02**	±0.001121	**95.65**
LRAVS	94.88	±0.001265	95.91	±0.001136	95.39
LRAVS+TCB	95.03	±0.001247	**96.02**	±0.001122	95.52
LRAVS+TCF	95.00	±0.001251	96.01	±0.001124	95.50
2005 Best	95.10	±0.001230	*95.20*	±0.001220	95.20
2005 Baseline	85.70	±0.002000	90.90	±0.001643	88.20
Our Baseline	86.40	±0.001967	91.15	±0.001629	88.71
2005 Topline	98.50	±0.000694	97.90	±0.000819	98.20
Our Topline	98.64	±0.000665	97.97	±0.000809	98.30

Table 10. Performance comparison of OOV on SIGHAN 2005 AS corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	66.09	±0.012356	61.85	±0.012678	63.90
CNG	67.39	±0.012235	66.81	±0.01229	67.10
AVS	68.93	±0.012078	70.73	±0.011875	69.82
TCB	66.16	±0.012349	64.02	±0.012668	64.02
TCF	**70.27**	±0.011929	63.89	±0.012536	66.93
AVS+TCB	69.31	±0.012037	**71.49**	±0.011783	**70.38**
AVS+TCF	69.59	±0.012006	*70.94*	±0.011850	*70.26*
LRAVS	66.31	±0.012336	67.07	±0.012266	66.69
LRAVS+TCB	67.33	±0.012241	67.91	±0.012184	67.62
LRAVS+TCF	*69.82*	±0.011981	66.15	±0.012350	67.94
2005 Best	69.60	±0.012005	N/A	N/A	N/A
2005 Baseline	0.40	±0.001647	N/A	N/A	N/A
Our Baseline	1.41	±0.003080	3.08	±0.004512	1.94
2005 Topline	99.60	±0.001647	N/A	N/A	N/A
Our Topline	99.59	±0.001677	95.48	±0.005420	97.49

Table 11. Performance comparison of accuracy on SIGHAN 2005 CityU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	94.82	±0.002207	94.64	±0.002245	94.73
CNG	**95.55**	±0.002055	94.39	±0.002292	94.97
AVS	95.27	±0.002115	94.93	±0.002185	95.10
TCB	95.21	±0.002129	94.93	±0.002186	95.07
TCF	95.30	±0.002107	94.96	±0.002180	95.13
AVS+TCB	95.34	±0.002100	95.13	±0.002145	95.23
AVS+TCF	95.39	±0.002088	95.15	±0.002140	95.27
LRAVS	95.35	±0.002099	95.08	±0.002155	95.21
LRAVS+TCB	*95.45*	±0.002077	**95.21**	±0.002127	**95.33**
LRAVS+TCF	95.41	±0.002085	*95.20*	±0.002130	*95.30*
2005 Best	94.60	±0.002230	94.10	±0.002330	94.30
2005 Baseline	79.00	±0.004026	88.20	±0.003189	83.30
Our Baseline	83.84	±0.003667	90.81	±0.002877	87.19
2005 Topline	99.10	±0.000934	98.80	±0.001076	98.20
Our Topline	99.24	±0.000867	98.90	±0.001040	99.07

Table 12. Performance comparison of OOV on SIGHAN 2005 CityU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	69.15	±0.016141	65.54	±0.016609	67.30
CNG	69.68	±0.016063	69.41	±0.016104	69.55
AVS	70.48	±0.015942	71.90	±0.015709	71.18
TCB	*71.83*	±0.015721	70.12	±0.016236	70.12
TCF	**72.39**	±0.015624	68.76	±0.016198	70.53
AVS+TCB	71.14	±0.015836	72.70	±0.01557	*71.91*
AVS+TCF	70.97	±0.015863	72.77	±0.015556	71.86
LRAVS	69.78	±0.016048	72.09	±0.015676	70.92
LRAVS+TCB	70.57	±0.015926	*73.06*	±0.015505	71.80
LRAVS+TCF	71.17	±0.015831	**73.22**	±0.015475	**72.18**
2005 Best	69.80	±0.016046	N/A	N/A	N/A
2005 Baseline	0.00	±0.000000	N/A	N/A	N/A
Our Baseline	16.22	±0.012882	33.91	±0.016544	21.94
2005 Topline	99.70	±0.001911	N/A	N/A	N/A
Our Topline	99.74	±0.001794	98.82	±0.003771	99.28

Table 13. *Performance comparison of accuracy on SIGHAN 2005 MSR corpus.*

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	*97.29*	±0.000998	97.03	±0.001042	*97.16*
CNG	97.02	±0.001045	96.87	±0.001069	96.95
AVS	97.24	±0.001007	96.91	±0.001063	97.07
TCB	**97.32**	±0.000993	**97.09**	±0.001033	**97.20**
TCF	97.02	±0.001044	96.70	±0.001097	96.86
AVS+TCB	97.16	±0.001020	96.91	±0.001063	97.04
AVS+TCF	97.25	±0.001005	97.00	±0.001049	97.12
LRAVS	97.20	±0.001014	97.01	±0.001046	97.10
LRAVS+TCB	97.21	±0.001012	*97.05*	±0.001040	97.13
LRAVS+TCF	*97.29*	±0.000997	96.43	±0.001139	96.86
2005 Best	96.60	±0.001110	96.20	±0.001170	96.40
2005 Baseline	91.20	±0.001733	95.50	±0.001268	93.30
Our Baseline	91.74	±0.001691	95.69	±0.001247	93.67
2005 Topline	99.20	±0.000545	99.10	±0.000578	99.10
Our Topline	99.31	±0.000510	99.10	±0.000580	99.20

Table 14. *Performance comparison of OOV on SIGHAN 2005 MSR corpus.*

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	72.22	±0.015108	60.52	±0.016487	65.85
CNG	71.37	±0.015247	62.08	±0.016365	66.40
AVS	69.88	±0.015474	61.96	±0.016375	65.68
TCB	72.96	±0.014982	**66.73**	±0.016414	66.73
TCF	**73.81**	±0.014830	58.68	±0.016608	65.38
AVS+TCB	70.41	±0.015395	62.11	±0.016362	66.00
AVS+TCF	71.12	±0.015286	62.54	±0.016325	66.56
LRAVS	70.91	±0.015319	63.02	±0.016283	*66.73*
LRAVS+TCB	71.05	±0.015297	*63.49*	±0.016239	**67.06**
LRAVS+TCF	**73.81**	±0.014830	59.28	±0.016571	65.75
2005 Best	71.70	±0.015194	N/A	N/A	N/A
2005 Baseline	0.00	±0.000000	N/A	N/A	N/A
Our Baseline	2.47	±0.005240	16.71	±0.012582	4.31
2005 Topline	99.80	±0.001507	N/A	N/A	N/A
Our Topline	99.79	±0.001552	99.37	±0.002676	99.58

Table 15. Performance comparison of accuracy on SIGHAN 2005 PKU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	93.73	±0.001512	92.70	±0.001623	93.21
CNG	**94.36**	±0.001438	**93.57**	±0.001530	**93.96**
AVS	94.21	±0.001457	93.24	±0.001566	93.72
TCB	93.97	±0.001485	92.76	±0.001616	93.36
TCF	93.94	±0.001488	92.81	±0.001611	93.37
AVS+TCB	94.33	±0.001443	93.31	±0.001559	93.81
AVS+TCF	94.25	±0.001451	93.44	±0.001544	93.85
LRAVS	*94.34*	±0.001441	*93.48*	±0.001540	*93.91*
LRAVS+TCB	94.32	±0.001443	93.44	±0.001544	93.88
LRAVS+TCF	93.91	±0.001492	92.20	±0.001672	93.05
2005 Best	94.60	±0.001400	95.30	±0.001310	95.00
2005 Baseline	83.60	±0.002292	90.40	±0.001824	86.90
Our Baseline	84.29	±0.002269	90.68	±0.001813	87.37
2005 Topline	98.80	±0.000674	98.50	±0.000752	98.70
Our Topline	98.96	±0.000634	98.62	±0.000726	98.79

Table 16. Performance comparison of OOV on SIGHAN 2005 PKU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	57.48	±0.012083	48.04	±0.012211	52.33
CNG	**65.58**	±0.011612	**57.87**	±0.012068	**61.48**
AVS	62.69	±0.011821	55.60	±0.012144	58.93
TCB	60.07	±0.011970	54.87	±0.012220	54.87
TCF	60.39	±0.011954	50.41	±0.012220	54.95
AVS+TCB	64.02	±0.011730	56.97	±0.012101	60.29
AVS+TCF	63.80	±0.011746	56.06	±0.012130	59.68
LRAVS	65.02	±0.011656	57.31	±0.012089	60.92
LRAVS+TCB	*65.42*	±0.011625	*57.60*	±0.012079	*61.26*
LRAVS+TCF	60.42	±0.011952	48.92	±0.012218	54.07
2005 Best	63.60	±0.011760	N/A	N/A	N/A
2005 Baseline	5.90	±0.005759	N/A	N/A	N/A
Our Baseline	6.86	±0.006178	6.10	±0.005850	6.46
2005 Topline	99.40	±0.001888	N/A	N/A	N/A
Our Topline	99.37	±0.001938	97.72	±0.003645	98.54

Table 17. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain A (Literature) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	92.83	±0.002754	92.37	±0.002833	92.60
CNG	*93.69*	±0.002595	91.94	±0.002906	92.81
AVS	93.47	±0.002638	92.89	±0.002744	93.18
TCB	93.12	±0.002702	92.56	±0.002801	92.84
TCF	93.18	±0.002690	92.52	±0.002808	92.85
AVS+TCB	93.68	±0.002596	92.99	±0.002726	93.33
AVS+TCF	93.67	±0.002600	93.10	±0.002705	*93.38*
LRAVS	93.55	±0.002623	93.08	±0.002709	93.31
LRAVS+TCB	93.56	±0.002620	*93.11*	±0.002703	93.33
LRAVS+TCF	**93.72**	±0.002589	**93.28**	±0.002673	**93.50**
2010 Best	94.60	±0.002390	94.50	±0.002410	94.60
2010 Baseline	86.20	±0.003648	91.70	±0.002919	88.90
Our Baseline	86.24	±0.003676	91.67	±0.002949	88.88
2010 Topline	99.00	±0.001053	98.60	±0.001243	98.80
Our Topline	99.02	±0.001052	98.57	±0.001268	98.79

Table 18. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain A (Literature) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	62.62	±0.019128	59.98	±0.01937	61.27
CNG	**65.36**	±0.018812	62.81	±0.019109	64.06
AVS	64.80	±0.018882	66.63	±0.018643	65.70
TCB	64.48	±0.018921	63.35	±0.019164	63.35
TCF	65.00	±0.018858	62.36	±0.019155	63.65
AVS+TCB	*65.04*	±0.018853	67.43	±0.018528	66.22
AVS+TCF	64.96	±0.018863	*67.60*	±0.018502	*66.26*
LRAVS	63.67	±0.019015	66.71	±0.018632	65.15
LRAVS+TCB	64.35	±0.018936	67.09	±0.018578	65.69
LRAVS+TCF	64.92	±0.018868	**68.48**	±0.018368	**66.65**
2010 Best	81.60	±0.015320	N/A	N/A	N/A
2010 Baseline	15.60	±0.014346	N/A	N/A	N/A
Our Baseline	15.69	±0.014378	30.61	±0.01822	20.74
2010 Topline	99.60	±0.002495	N/A	N/A	N/A
Our Topline	99.60	±0.002505	96.48	±0.007282	98.02

Table 19. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain B (Computer) corpus.

Configuration	*P*	C_P	*R*	C_R	*F*
6-tag	90.95	±0.003129	92.46	±0.002880	91.70
CNG	91.45	±0.003050	92.36	±0.002898	91.90
AVS	91.25	±0.003081	*92.72*	±0.002833	91.98
TCB	91.21	±0.003087	92.53	±0.002867	91.87
TCF	90.86	±0.003143	92.62	±0.002852	91.73
AVS+TCB	91.60	±0.003026	92.67	±0.002842	92.13
AVS+TCF	90.81	±0.003151	92.16	±0.002932	91.48
LRAVS	*91.71*	±0.003007	92.61	±0.002854	*92.16*
LRAVS+TCB	**91.97**	±0.002963	**92.76**	±0.002826	**92.37**
LRAVS+TCF	91.28	±0.003077	92.60	±0.002856	91.93
2010 Best	95.00	±0.002320	95.30	±0.002250	95.10
2010 Baseline	63.20	±0.005132	85.60	±0.003736	72.70
Our Baseline	63.26	±0.005258	85.68	±0.003820	72.78
2010 Topline	99.30	±0.000887	99.10	±0.001005	99.20
Our Topline	99.25	±0.000940	99.06	±0.001052	99.16

Table 20. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain B (Computer) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	70.62	±0.013380	67.66	±0.013740	69.11
CNG	70.38	±0.013412	65.17	±0.013994	67.67
AVS	69.85	±0.013479	66.16	±0.013898	67.96
TCB	71.23	±0.013297	**69.66**	±0.013684	69.66
TCF	**72.01**	±0.013187	66.02	±0.013913	68.89
AVS+TCB	70.25	±0.013429	67.22	±0.013788	68.70
AVS+TCF	69.63	±0.013507	63.73	±0.014123	66.55
LRAVS	71.25	±0.013294	68.25	±0.013673	*69.72*
LRAVS+TCB	*71.81*	±0.013216	*69.47*	±0.013528	**70.62**
LRAVS+TCF	70.92	±0.013340	66.13	±0.013902	68.44
2010 Best	82.70	±0.011111	N/A	N/A	N/A
2010 Baseline	16.30	±0.010850	N/A	N/A	N/A
Our Baseline	16.65	±0.010944	6.39	±0.007185	9.24
2010 Topline	99.00	±0.002923	N/A	N/A	N/A
Our Topline	99.00	±0.002930	98.08	±0.004028	98.54

Table 21. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain C (Medicine) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	91.27	±0.003207	91.96	±0.003089	91.61
CNG	92.84	±0.002928	92.07	±0.003069	92.46
AVS	92.40	±0.003011	92.89	±0.002919	92.64
TCB	91.55	±0.003159	92.19	±0.003048	91.87
TCF	91.62	±0.003147	92.21	±0.003045	91.91
AVS+TCB	92.73	±0.002949	92.90	±0.002917	92.82
AVS+TCF	92.82	±0.002933	93.07	±0.002885	92.94
LRAVS	*93.12*	±0.002876	*93.22*	±0.002856	*93.17*
LRAVS+TCB	**93.12**	±0.002875	**93.33**	±0.002834	**93.23**
LRAVS+TCF	93.07	±0.002884	93.20	±0.002859	93.14
2010 Best	93.60	±0.002760	94.20	±0.002630	93.90
2010 Baseline	77.40	±0.004714	88.60	±0.003582	82.60
Our Baseline	77.46	±0.004746	88.64	±0.003604	82.68
2010 Topline	99.10	±0.001064	98.90	±0.001176	99.00
Our Topline	99.18	±0.001025	98.97	±0.001146	99.08

Table 22. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain C (Medicine) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	66.70	±0.016081	61.15	±0.016630	63.80
CNG	70.90	±0.015498	70.46	±0.015567	70.68
AVS	71.02	±0.015479	69.61	±0.015692	70.31
TCB	66.41	±0.016115	60.67	±0.016667	63.41
TCF	66.44	±0.016112	60.65	±0.016668	63.41
AVS+TCB	70.10	±0.015621	69.00	±0.015780	69.54
AVS+TCF	69.66	±0.015685	69.11	±0.015765	69.38
LRAVS	**71.62**	±0.015382	**70.91**	±0.015497	**71.26**
LRAVS+TCB	71.45	±0.015410	70.39	±0.015576	70.92
LRAVS+TCF	*71.56*	±0.015392	*70.53*	±0.015556	*71.04*
2010 Best	75.00	±0.014774	N/A	N/A	N/A
2010 Baseline	12.30	±0.011206	N/A	N/A	N/A
Our Baseline	12.33	±0.011218	15.34	±0.012294	13.67
2010 Topline	98.00	±0.004777	N/A	N/A	N/A
Our Topline	98.21	±0.004519	97.21	±0.005623	97.71

Table 23. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 simplified Chinese domain D (Finance) corpus.

Configuration	P	C_P	R	C_R	F
6-tag	93.01	±0.002838	93.74	±0.002697	93.38
CNG	*94.40*	±0.002561	93.66	±0.002714	94.02
AVS	93.54	±0.002736	*94.30*	±0.002581	93.92
TCB	93.35	±0.002774	94.14	±0.002614	93.74
TCF	93.10	±0.002822	93.88	±0.002669	93.49
AVS+TCB	**94.56**	±0.002526	**94.49**	±0.002540	**94.53**
AVS+TCF	94.05	±0.002633	94.10	±0.002624	94.08
LRAVS	94.30	±0.002582	94.13	±0.002616	94.21
LRAVS+TCB	94.36	±0.002568	94.16	±0.002611	94.26
LRAVS+TCF	94.36	±0.002569	94.19	±0.002604	*94.28*
2010 Best	96.00	±0.002160	95.90	±0.002180	95.90
2010 Baseline	80.30	±0.004377	91.40	±0.003085	85.50
Our Baseline	80.26	±0.004431	91.41	±0.003119	85.48
2010 Topline	99.50	±0.000776	99.40	±0.000850	99.40
Our Topline	99.56	±0.000734	99.47	±0.000810	99.52

Table 24. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 simplified Chinese domain D (Finance) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	67.60	±0.017666	61.28	±0.018388	64.28
CNG	*73.53*	±0.016655	67.77	±0.017642	70.53
AVS	71.10	±0.017111	64.17	±0.018101	67.46
TCB	70.58	±0.017201	66.44	±0.018250	66.44
TCF	70.13	±0.017277	61.19	±0.018396	65.35
AVS+TCB	**73.80**	±0.016598	**70.79**	±0.017166	**72.26**
AVS+TCF	70.76	±0.017172	67.73	±0.017648	69.21
LRAVS	71.66	±0.017012	68.54	±0.017528	70.07
LRAVS+TCB	72.63	±0.016831	*69.82*	±0.017328	*71.20*
LRAVS+TCF	72.38	±0.016878	69.40	±0.017396	70.86
2010 Best	82.70	±0.014279	N/A	N/A	N/A
2010 Baseline	23.30	±0.015958	N/A	N/A	N/A
Our Baseline	23.32	±0.015963	14.15	±0.013157	17.61
2010 Topline	99.50	±0.002663	N/A	N/A	N/A
Our Topline	99.72	±0.001985	99.34	±0.003047	99.53

Table 25. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain A (Literature) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	93.06	±0.002672	92.31	±0.002802	92.68
CNG	93.66	±0.002562	91.16	±0.002985	92.39
AVS	93.61	±0.002572	92.78	±0.002721	93.19
TCB	93.21	±0.002646	92.33	±0.002798	92.77
TCF	93.33	±0.002623	92.58	±0.002756	92.95
AVS+TCB	93.61	±0.002572	92.85	±0.002709	93.23
AVS+TCF	93.68	±0.002559	92.98	±0.002685	93.33
LRAVS	*93.77*	±0.002542	*93.04*	±0.002676	*93.40*
LRAVS+TCB	**93.77**	±0.002541	**93.06**	±0.002673	**93.41**
LRAVS+TCF	93.65	±0.002564	92.92	±0.002697	93.28
2010 Best	94.20	±0.002450	94.20	±0.002450	94.20
2010 Baseline	78.80	±0.004286	86.30	±0.003606	82.40
Our Baseline	78.83	±0.004295	86.39	±0.003605	82.44
2010 Topline	98.80	±0.001142	98.10	±0.001432	98.50
Our Topline	98.83	±0.001130	98.11	±0.001430	98.47

Table 26. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 traditional Chinese domain A (Literature) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	75.89	±0.014654	68.68	±0.015889	72.11
CNG	74.12	±0.015004	69.46	±0.015780	71.71
AVS	75.10	±0.014816	73.34	±0.015148	74.21
TCB	**77.19**	±0.014376	69.27	±0.015807	73.01
TCF	*77.10*	±0.014395	69.82	±0.015727	73.28
AVS+TCB	75.54	±0.014727	73.46	±0.015127	74.48
AVS+TCF	75.60	±0.014715	73.92	±0.015042	74.75
LRAVS	75.42	±0.014751	*74.93*	±0.014848	*75.18*
LRAVS+TCB	75.66	±0.014703	**75.12**	±0.014810	**75.39**
LRAVS+TCF	75.27	±0.014780	74.44	±0.014944	74.85
2010 Best	78.80	±0.014003	N/A	N/A	N/A
2010 Baseline	4.10	±0.006793	N/A	N/A	N/A
Our Baseline	4.10	±0.006791	8.93	±0.009769	5.62
2010 Topline	99.80	±0.001531	N/A	N/A	N/A
Our Topline	99.82	±0.001439	99.33	±0.002804	99.57

Table 27. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain B (Computer) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	95.15	±0.002122	93.20	±0.002487	94.17
CNG	95.60	±0.002027	93.16	±0.002494	94.36
AVS	*95.67*	±0.002012	*93.83*	±0.002378	*94.74*
TCB	95.21	±0.002111	93.25	±0.002480	94.22
TCF	95.28	±0.002095	93.42	±0.002450	94.34
AVS+TCB	95.62	±0.002023	93.72	±0.002398	94.66
AVS+TCF	**95.74**	±0.001996	93.83	±0.002378	**94.77**
LRAVS	95.57	±0.002034	93.79	±0.002384	94.67
LRAVS+TCB	95.63	±0.002020	**93.85**	±0.002373	94.73
LRAVS+TCF	95.55	±0.002038	93.81	±0.002381	94.67
2010 Best	95.70	±0.001950	94.80	±0.002130	95.20
2010 Baseline	70.10	±0.004390	87.30	±0.003193	77.80
Our Baseline	70.15	±0.004522	87.33	±0.003286	77.80
2010 Topline	99.10	±0.000906	98.80	±0.001044	99.00
Our Topline	99.38	±0.000778	98.85	±0.001055	99.11

Table 28. Non-Chinese-Pattern performance comparison of OOV on SIGHAN 2010 traditional Chinese domain B (Computer) corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	58.79	±0.016769	68.17	±0.015871	63.14
CNG	*61.77*	±0.016556	70.16	±0.015589	65.70
AVS	60.59	±0.016649	72.29	±0.015248	65.93
TCB	59.09	±0.016751	68.81	±0.015784	63.58
TCF	59.34	±0.016735	69.21	±0.015727	63.89
AVS+TCB	60.89	±0.016626	72.24	±0.015257	66.08
AVS+TCF	61.35	±0.01659	72.90	±0.015143	66.63
LRAVS	61.67	±0.016564	72.84	±0.015155	*66.79*
LRAVS+TCB	**61.82**	±0.016552	**73.07**	±0.015113	**66.98**
LRAVS+TCF	61.55	±0.016574	*72.94*	±0.015135	66.76
2010 Best	66.60	±0.016069	N/A	N/A	N/A
2010 Baseline	1.00	±0.003390	N/A	N/A	N/A
Our Baseline	1.03	±0.003445	0.55	±0.002515	0.72
2010 Topline	99.60	±0.002150	N/A	N/A	N/A
Our Topline	99.34	±0.002765	99.41	±0.002609	99.37

Table 29. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain C (Medicine) corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	94.70	±0.002170	93.83	±0.002331	94.26
CNG	95.35	±0.002039	93.35	±0.002414	94.34
AVS	95.28	±0.002055	94.37	±0.002232	94.82
TCB	94.76	±0.002158	93.87	±0.002324	94.31
TCF	94.88	±0.002135	94.05	±0.002291	94.46
AVS+TCB	95.33	±0.002044	94.49	±0.002209	94.91
AVS+TCF	95.33	±0.002043	94.44	±0.002219	94.88
LRAVS	**95.52**	±0.002003	**94.60**	±0.002190	**95.06**
LRAVS+TCB	95.36	±0.002038	*94.51*	±0.002206	*94.93*
LRAVS+TCF	*95.42*	±0.002025	94.42	±0.002224	94.91
2010 Best	95.70	±0.001950	95.30	±0.002030	95.50
2010 Baseline	81.00	±0.003764	88.60	±0.003049	84.60
Our Baseline	80.98	±0.003801	88.63	±0.003075	84.64
2010 Topline	98.90	±0.001001	98.40	±0.001204	98.60
Our Topline	98.91	±0.001006	98.38	±0.001223	98.64

Table 30. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 traditional Chinese domain C (Medicine) corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	74.79	±0.015086	67.98	±0.016209	71.22
CNG	77.16	±0.014586	71.22	±0.015730	74.07
AVS	76.13	±0.014810	74.80	±0.015083	75.46
TCB	75.60	±0.014922	68.64	±0.016119	71.95
TCF	75.79	±0.014883	69.29	±0.016026	72.39
AVS+TCB	76.72	±0.014683	*75.75*	±0.014890	76.23
AVS+TCF	77.22	±0.014572	75.69	±0.014903	76.44
LRAVS	**78.65**	±0.014237	**76.37**	±0.014759	**77.49**
LRAVS+TCB	77.75	±0.014451	75.54	±0.014934	76.63
LRAVS+TCF	*78.03*	±0.014385	75.65	±0.014911	*76.82*
2010 Best	79.80	±0.013949	N/A	N/A	N/A
2010 Baseline	2.70	±0.005631	N/A	N/A	N/A
Our Baseline	2.71	±0.005639	4.34	±0.007082	3.34
2010 Topline	99.20	±0.003095	N/A	N/A	N/A
Our Topline	99.16	±0.003171	98.73	±0.003891	98.94

Table 31. Non-Chinese treatment performance comparison of accuracy on SIGHAN 2010 traditional Chinese domain D (Finance) corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	95.52	±0.001925	95.46	±0.001937	95.49
CNG	**96.13**	±0.001794	95.04	±0.002020	95.58
AVS	95.99	±0.001825	95.79	±0.001868	95.89
TCB	95.55	±0.001918	95.51	±0.001927	95.53
TCF	95.61	±0.001907	95.57	±0.001915	95.59
AVS+TCB	95.93	±0.001839	95.77	±0.001874	95.85
AVS+TCF	95.99	±0.001825	**95.88**	±0.001850	**95.93**
LRAVS	96.02	±0.001820	95.73	±0.001881	95.87
LRAVS+TCB	*96.04*	±0.001814	*95.82*	±0.001862	*95.93*
LRAVS+TCF	95.94	±0.001836	95.71	±0.001885	95.83
2010 Best	96.20	±0.001760	96.40	±0.001720	96.30
2010 Baseline	82.60	±0.003492	88.80	±0.002905	85.50
Our Baseline	82.56	±0.003531	88.77	±0.002937	85.55
2010 Topline	98.60	±0.001082	98.10	±0.001258	98.40
Our Topline	98.63	±0.001081	98.10	±0.00127	98.36

Table 32. Non-Chinese treatment performance comparison of OOV on SIGHAN 2010 traditional Chinese domain D (Finance) corpus.

Configuration	R_{OOV}	$C_{R_{oov}}$	P_{OOV}	$C_{P_{oov}}$	F_{OOV}
6-tag	80.45	±0.013488	76.61	±0.014398	78.48
CNG	**82.96**	±0.012787	78.16	±0.014053	80.49
AVS	81.33	±0.013253	81.28	±0.013267	81.30
TCB	80.99	±0.013346	77.44	±0.014216	79.17
TCF	80.92	±0.013363	77.26	±0.014255	79.05
AVS+TCB	80.99	±0.013346	81.55	±0.013193	81.27
AVS+TCF	80.99	±0.013346	81.96	±0.013077	81.47
LRAVS	*82.62*	±0.012889	*82.10*	±0.013038	**82.36**
LRAVS+TCB	82.18	±0.013016	**82.44**	±0.012942	*82.31*
LRAVS+TCF	81.86	±0.013105	82.04	±0.013054	81.95
2010 Best	81.20	±0.013288	N/A	N/A	N/A
2010 Baseline	0.60	±0.002627	N/A	N/A	N/A
Our Baseline	0.60	±0.002618	2.28	±0.005078	0.95
2010 Topline	99.70	±0.001860	N/A	N/A	N/A
Our Topline	99.69	±0.001902	98.54	±0.004076	99.11

It has been observed that using any of the unsupervised features could create short patterns for the CRF learner, which might break more English words than using the *6-tag* approach alone. AVS, TCF, and TCB, however, resolve more overlapping ambiguities of Chinese words than the *6-tag* approach and CNG. Interestingly, even for the unsupervised feature without rank or overlapping information, TCB/TCF successfully recognizes "依靠 / 单位 / 的 / 纽带 / 来 / 维持," while the *6-tag* approach sees this phrase incorrectly as "依靠 / 单位 / 的 / 纽 / 带来 / 维持." TCB/TCF also saves more factoids, such as "一二九・九 / 左右" (129.9 / around) from scattered tokens, such as "一二九 / ・ / 九 / 左右" (129 / point / 9 / around).

The above observations suggest that the quality of a string as a word-like candidate should be an important factor for the unsupervised feature injected CRF learner. Relatively speaking, CNG probably brings in too much noise. Feature combinations of LRAVS and TCF usually improve F and F_{OOV}, respectively. Improvements are significant in terms of C_R, C_P, C_{Roov}, and C_{Poov},, which confirms the hypothesis mentioned at the end of Section 1.3 that, combining information from the outer pattern of a substring (*i.e.*, LRAVS) with information from the inner pattern of a substring (*i.e.*, TCF) into a compound of unsupervised feature could help improving CWS performance of supervised labeling scheme of CRF. Nevertheless, since AVS or TCB sometimes gain better results, fine-tuning of feature engineering according to different corpora and segmentation standards is necessary.

7. Conclusion and Future Work

This work provides a unified view of CRF-based CWS integrated with unsupervised features via frequent string, and it reasons that, since LRAVS comes with inner structure and TCF comes with outer structure of overlapping string, utilizing their compound features could be more useful than applying one of them solely. The thorough experimental results show that the compound features of LRAVS and TCF usually obtain competitive performance in terms of F and F_{OOV}, respectively. Sometimes, AVS and TCB may contribute more, but generally combining the outer pattern of a substring (*i.e.*, LRAVS or AVS) with the inner pattern of a substring (*i.e.*, TCF or TCB) into a compound of unsupervised features could help improve CWS performance of a supervised labeling scheme of CRF. Recommended future investigation is unknown word extraction and named entity recognition using AVS (Li *et al.*, 2010) and TCF/TCB(Chang & Lee, 2003; Zhang *et al.*, 2010) as features for more complicated CRF (Sun & Nan, 2010).

Reference

Chang, J.-S., & Su, K.-Y. (1997). An Unsupervised Iterative Method for Chinese New Lexicon Extraction. in *Proc. Computational Linguistics and Chinese Language Processing,* 2(2), 97-148.

Chang, T.-H., & Lee, C.-H. (2003). Automatic Chinese unknown word extraction using small-corpus-based method. in *Proc. International Conference on Natural Language Processing and Knowledge Engineering*, 459-464.

Chien, L.-F. (1997). PAT-tree-based Keyword Extraction for Chinese Information Retrieval. in *Proc. 20th Annnual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-58.

Cohen, P., Adams, N., & Heeringa, B. (2007). Voting Experts: An Unsupervised Algorithm for Segmenting Sequences. *Intelligent Data Analysis,* 11(6), 607-625.

Emerson, T. (2005). The Second International Chinese Word Segmentation Bakeoff. in *Proc. 4th SIGHAN Workshop on Chinese Language Processing*.

Feng, H., Chen, K., Deng, X., & Zheng, W. (2004). Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics,* 30(1), 75-93.

Ha, L. Q., Seymour, R., Hanna, P., & Smith, F. J. (2005). Reduced N-Grams for Chinese Evaluation. *Computational Linguistics and Chinese Language Processing,* 10(1), 19-34.

Harris, Z. S. (1970). Morpheme Boundaries within Words. Paper presented at the *Structural and Transformational Linguistics*.

Huang, J. H., & Powers, D. (2003). Chinese Word Segmentation based on contextual entropy. in *Proc. 17th Asian Pacific Conference on Language, Information and Computation*, 152-158.

Jiang, T.-J., Hsu, W.-L., Kuo, C.-H., & Yang, T.-H. (2011). Enhancement of Unsupervised Feature Selection for Conditional Random Fields Learning in Chinese Word Segmentation. in *Proc. 7th IEEE International Conference on Natural Language Processing and Knowledge Engineering,* 382-389.

Jiang, T.-J., Liu, S.-H., Sung, C.-L., & Hsu, W.-L. (2010). Term Contributed Boundary Tagging by Conditional Random Fields for SIGHAN 2010 Chinese Word Segmentation Bakeoff. in *Proc. 1st CIPS-SIGHAN Joint Conf. on Chinese Language Processing*, Beijing, China.

Jin, G., & Chen, X. (2007). The Fourth International Chinese Language Processing Bakeoff : Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging. in *Proc. 6th SIGHAN Workshop on Chinese Language Processing*, 69-81.

Kit, C., & Wilks, Y. (1999). Unsupervised learning of word boundary with description length gain. in *Proc. CoNLL-99*, 1-6.

Lü, X., & Zhang, L. (2005). Statistical Substring Reduction in Linear Time. in *Proc. 1st Internal Joint Conference on Natural Language Processing*.

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields

Probabilistic Models for Segmenting and Labeling Sequence Data. in *Proc. ICML*. 282-289.

Levow, G.-A. (2006). The Third International Chinese Language Processing Bakeoff Word Segmentation and Named Entity Recognition. in *Proc. 5th SIGHAN Workshop on Chinese Language Processing,* 108-117.

Li, L., Li, Z., Ding, Z., & Huang, D. (2010). A Hybrid Model Combining CRF with Boundary Templates for Chinese Person Name Recognition. *International Journal Advanced Intelligent,* 2(1), 73-80.

Li, M., Gao, J., Huang, C., & Li, J. (2003). Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation. in *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, 17, 1-7.

Lin, Y.-J., & Yu, M.-S. (2001). Extracting Chinese Frequent Strings without a Dictionary from a Chinese Corpus and its Applications. *J. Information Science and Engineering,* 17, 805-824.

Ma, W.-Y., & Chen, K.-J. (2003). Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. in *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, 17, 168-171.

Manber, U., & Myers, G. (1993). Suffix arrays: a new method for on-line string searches. *SIAM J. Computing,* 22(5), 935-948.

O'Boyle, P. (1993). *A Study of an N-Gram Language Model for Speech Recognition.* (Ph.D.), Queen's University Belfast.

Qiao, W., Sun, M., & Menzel, W. (2008). Statistical Properties of Overlapping Ambiguities in Chinese Word Segmentation and a Strategy for Their Disambiguation. in *Proc. Text, Speech and Dialogue*, 177-186.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. in *Proc. Empirical Methods in Natural Language Processing*, 133-142.

Sproat, R., & Emerson, T. (2003). The First International Chinese Word Segmentation Bakeoff. in *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*, 17, 133-143.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. in *Proc. Spoken Language Processing*, 901-904.

Sun, M., Huang, C. N., Lu, F., & Shen, D. Y. (1997). Using Character Bigram for Ambiguity Resolution In Chinese Word Segmentation (In Chinese). *Computer Research and Development,* 34(5), 332-339.

Sun, W., & Xu, J. (2011). Enhancing Chinese Word Segmentation Using Unlabeled Data. in *Proc. Empirical Methods in Natural Language Processing,* 970-979.

Sun, X., & Nan, X. (2010). Chinese base phrases chunking based on latent semi-CRF model. in *Proc. International Conference on Natural Language Processing and Knowledge Engineering*, 1-7.

Sung, C.-L., Yen, H.-C., & Hsu, W.-L. (2008). Compute the Term Contributed Frequency. in

Proc. 8th Int. Conference Intelligent System Design and Application, 2, 325-328.

Tanaka-Ishii, K. (2005). Entropy as an Indicator of Context Boundaries: An Experiment Using a Web Search Engine. in *Proc. Internal Joint Conference on Natural Language Processing*, 93-105.

Tung, C.-H., & Lee, H.-J. (1994). Identification of Unkown Words from Corpus. *Computational Proc. Chinese and Oriental Languages*, 8, 131-145.

Wallach, H. M. (2004). *Conditional Random Fields An Introduction*. (MS-CIS-04-21).

Zhang, H., Huang, H., Zhu, C., & Shi, S. (2010). A pragmatic model for new Chinese word extraction. in *Proc. International Conference on Natural Language Processing and Knowledge Engineering*, 1-8.

Zhang, R., Kikui, G., & Sumita, E. (2006). Subword-based Tagging for Confidence-dependent Chinese Word Segmentation. in *Proc. COLING/ACL*, 961-968.

Zhao, H., Huang, C.-N., Li, M., & Lu, B.-L. (2010). A Unified Character-Based Tagging Framework for Chinese Word Segmentation. *ACM Trans. on Asian Language Information Processing,* 9(2).

Zhao, H., & Kit, C. (2007). Incorporating Global Information into Supervised Learning for Chinese Word Segmentation. in *Proc. 10th PACLIC*, 66-74.

Zhao, H., & Liu, Q. (2010). The CIPS-SIGHAN CLP2010 Chinese Word Segmentation Backoff. in *Proc. 1st CIPS-SIGHAN Joint Conf. on Chinese Language Processing*, 199-209.

Appendix

Table 33. Performance comparison of accuracy on SIGHAN 2003 AS corpus.

Configuration	P	C_P	R	C_R	F
6-tag	**97.18**	±0.003024	97.23	±0.002998	**97.21**
CNG	97.05	±0.003091	97.16	±0.003033	97.11
AVS	97.06	±0.003086	97.23	±0.002998	97.14
TCB	*97.16*	±0.003037	97.18	±0.003024	97.17
TCF	97.15	±0.003042	97.11	±0.003059	97.13
AVS+TCB	97.04	±0.003098	*97.24*	±0.002994	97.14
AVS+TCF	97.07	±0.003081	**97.30**	±0.002958	*97.19*
LRAVS	96.89	±0.003172	97.15	±0.003042	97.02
LRAVS+TCB	97.03	±0.003103	97.20	±0.003011	97.12
LRAVS+TCF	96.94	±0.003147	*97.24*	±0.002994	97.09
2003 Best	95.60	±0.003700	96.60	±0.003300	96.10
2003 Baseline	91.20	±0.005175	91.70	±0.005040	91.50
Our Baseline	91.23	±0.005168	91.74	±0.005029	91.48
2003 Topline	99.30	±0.001523	99.00	±0.001818	99.20
Our Topline	99.30	±0.001526	99.02	±0.001804	99.16

Table 34. Performance comparison of OOV on SIGHAN 2003 AS corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	*77.13*	±0.052294	75.09	±0.053848	**76.10**
CNG	73.64	±0.054857	75.10	±0.053845	74.36
AVS	70.93	±0.056540	77.22	±0.052227	73.94
TCB	76.74	±0.052603	74.44	±0.054316	*75.57*
TCF	**77.91**	±0.051658	71.02	±0.056486	74.31
AVS+TCB	70.93	±0.056540	*77.54*	±0.051960	74.09
AVS+TCF	70.93	±0.056540	**77.87**	±0.051687	74.24
LRAVS	69.77	±0.057185	76.27	±0.052971	72.87
LRAVS+TCB	69.38	±0.057391	76.50	±0.052797	72.76
LRAVS+TCF	70.16	±0.056975	76.37	±0.052894	73.13
2003 Best	36.40	±0.059910	N/A	N/A	N/A
2003 Baseline	0.00	±0.000000	N/A	N/A	N/A
Our Baseline	0.00	±0.000000	0.00	±0.000000	0.00
2003 Topline	98.80	±0.013558	N/A	N/A	N/A
Our Topline	98.84	±0.013348	97.33	±0.020079	98.08

Table 35. Performance comparison of accuracy on SIGHAN 2003 CityU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	94.77	±0.002381	94.79	±0.002377	94.78
CNG	**95.24**	±0.002278	**95.48**	±0.002222	**95.36**
AVS	95.13	±0.002302	95.20	±0.002286	95.17
TCB	94.84	±0.002367	94.87	±0.002360	94.85
TCF	94.78	±0.002380	94.77	±0.002382	94.77
AVS+TCB	*95.18*	±0.002291	95.24	±0.002278	95.21
AVS+TCF	95.08	±0.002313	95.19	±0.002288	95.14
LRAVS	95.00	±0.002332	95.21	±0.002284	95.10
LRAVS+TCB	*95.18*	±0.002292	*95.33*	±0.002256	*95.26*
LRAVS+TCF	95.00	±0.002330	95.27	±0.002271	95.14
2003 Best	93.40	±0.002700	94.70	±0.002400	94.00
2003 Baseline	83.00	±0.004018	90.80	±0.003092	86.70
Our Baseline	82.97	±0.004021	90.77	±0.003097	86.69
2003 Topline	99.10	±0.001010	98.60	±0.001257	98.90
Our Topline	99.10	±0.001009	98.62	±0.001249	98.86

Table 36. Performance comparison of OOV on SIGHAN 2003 CityU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	75.80	±0.017149	66.07	±0.018969	70.60
CNG	**77.25**	±0.016796	**73.25**	±0.017735	75.20
AVS	75.16	±0.017311	71.79	±0.018030	73.44
TCB	76.20	±0.017061	66.63	±0.018891	71.10
TCF	*76.28*	±0.017041	66.38	±0.018927	70.99
AVS+TCB	75.44	±0.017245	72.06	±0.017977	*73.71*
AVS+TCF	74.88	±0.017376	71.66	±0.018055	73.23
LRAVS	74.12	±0.017548	72.01	±0.017987	73.05
LRAVS+TCB	74.88	±0.017376	*72.92*	±0.017804	**73.89**
LRAVS+TCF	74.32	±0.017503	72.23	±0.017943	73.26
2003 Best	62.50	±0.019396	N/A	N/A	N/A
2003 Baseline	3.70	±0.007563	N/A	N/A	N/A
Our Baseline	3.69	±0.007555	5.20	±0.008896	4.32
2003 Topline	99.60	±0.002529	N/A	N/A	N/A
Our Topline	99.60	±0.002533	98.65	±0.004626	99.12

Table 37. Performance comparison of accuracy on SIGHAN 2003 PKU corpus.

Configuration	P	C_P	R	C_R	F
6-tag	92.98	±0.003897	93.67	±0.003713	93.32
CNG	94.35	±0.003521	94.70	±0.003417	94.53
AVS	**94.39**	±0.003510	94.70	±0.003417	*94.54*
TCB	93.14	±0.003856	93.69	±0.003709	93.41
TCF	93.43	±0.003780	93.58	±0.003739	93.50
AVS+TCB	**94.43**	±0.003498	**94.84**	±0.003376	94.63
AVS+TCF	94.32	±0.003529	*94.83*	±0.003377	**94.58**
LRAVS	94.18	±0.003572	94.71	±0.003415	94.44
LRAVS+TCB	94.26	±0.003548	94.81	±0.003383	94.53
LRAVS+TCF	94.04	±0.003611	94.62	±0.003441	94.33
2003 Best	94.00	±0.003600	96.20	±0.002900	95.10
2003 Baseline	82.90	±0.005743	90.90	±0.004387	86.70
Our Baseline	82.96	±0.005735	90.87	±0.004392	86.74
2003 Topline	99.60	±0.000963	99.50	±0.001076	99.50
Our Topline	99.63	±0.000930	99.45	±0.001125	99.54

Table 38. Performance comparison of OOV on SIGHAN 2003 PKU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	60.22	±0.028389	49.69	±0.029	54.45
CNG	67.70	±0.027122	63.24	±0.027966	65.39
AVS	66.36	±0.027405	64.94	±0.027676	65.64
TCB	61.14	±0.028271	51.49	±0.028988	55.90
TCF	63.58	±0.027910	54.74	±0.028870	58.83
AVS+TCB	68.54	±0.026932	**66.31**	±0.027414	**67.41**
AVS+TCF	*68.29*	±0.026990	65.22	±0.027624	*66.72*
LRAVS	67.12	±0.027249	64.56	±0.027743	65.81
LRAVS+TCB	**68.46**	±0.026952	*64.91*	±0.027681	66.64
LRAVS+TCF	66.95	±0.027284	63.02	±0.028	64.93
2003 Best	61.65	±0.025928	N/A	N/A	N/A
2003 Baseline	5.00	±0.012641	N/A	N/A	N/A
Our Baseline	4.96	±0.012596	5.12	±0.01278	5.04
2003 Topline	100.00	±0.000000	N/A	N/A	N/A
Our Topline	100.00	±0.000000	99.92	±0.001681	99.96

Table 39. Performance comparison of accuracy on SIGHAN 2003 CTB corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	87.30	±0.003334	86.83	±0.003385	87.06
CNG	**89.61**	±0.003054	**88.66**	±0.003175	**89.13**
AVS	89.38	±0.003085	88.06	±0.003246	88.71
TCB	87.46	±0.003315	86.86	±0.003382	87.16
TCF	87.18	±0.003347	86.45	±0.003426	86.81
AVS+TCB	89.31	±0.003092	88.08	±0.003244	88.69
AVS+TCF	*89.39*	±0.003082	88.17	±0.003233	*88.78*
LRAVS	89.30	±0.003094	*88.21*	±0.003228	88.75
LRAVS+TCB	89.37	±0.003086	88.09	±0.003243	88.72
LRAVS+TCF	89.31	±0.003093	88.07	±0.003244	88.68
2003 Best	87.50	±0.003300	86.60	±0.003200	88.10
2003 Baseline	66.30	±0.004731	80.00	±0.004004	72.50
Our Baseline	66.33	±0.004730	80.01	±0.004003	72.53
2003 Topline	98.80	±0.001090	98.20	±0.001331	98.50
Our Topline	98.84	±0.001072	98.19	±0.001333	98.52

Table 40. Performance comparison of OOV on SIGHAN 2003 CTB corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	69.85	±0.010805	62.24	±0.011415	65.83
CNG	**71.79**	±0.010596	**71.31**	±0.010650	**71.55**
AVS	70.59	±0.010728	69.61	±0.010830	70.09
TCB	70.23	±0.010766	62.51	±0.011398	66.14
TCF	69.49	±0.010841	61.91	±0.011434	65.48
AVS+TCB	70.73	±0.010714	70.05	±0.010785	70.39
AVS+TCF	*70.95*	±0.010690	69.80	±0.010811	70.37
LRAVS	70.35	±0.010753	69.98	±0.010793	70.16
LRAVS+TCB	70.58	±0.010730	*70.49*	±0.010739	*70.53*
LRAVS+TCF	70.24	±0.010765	70.05	±0.010785	70.15
2003 Best	70.50	±0.010738	N/A	N/A	N/A
2003 Baseline	6.20	±0.005678	N/A	N/A	N/A
Our Baseline	6.24	±0.005694	8.36	±0.006516	7.14
2003 Topline	99.00	±0.002343	N/A	N/A	N/A
Our Topline	99.02	±0.002324	97.46	±0.003703	98.23

Table 41. Performance comparison of accuracy on SIGHAN 2006 AS corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	94.57	±0.001499	95.76	±0.001333	95.16
CNG	95.13	±0.001424	96.16	±0.001271	95.64
AVS	95.25	±0.001407	96.18	±0.001267	95.71
TCB	94.74	±0.001477	95.87	±0.001316	95.30
TCF	94.80	±0.001468	95.85	±0.001319	95.32
AVS+TCB	95.32	±0.001398	96.23	±0.001260	*95.77*
AVS+TCF	*95.33*	±0.001395	96.21	±0.001263	*95.77*
LRAVS	95.24	±0.001408	*96.25*	±0.001256	95.74
LRAVS+TCB	**95.34**	±0.001394	**96.31**	±0.001247	**95.82**
LRAVS+TCF	95.12	±0.001424	95.97	±0.001300	95.55
2006 Best	95.50	±0.001371	96.10	±0.00128	95.80
2006 Baseline	87.00	±0.002224	91.50	±0.001844	89.20
Our Baseline	87.03	±0.002222	91.47	±0.001848	89.19
2006 Topline	98.70	±0.000749	98.00	±0.000926	98.30
Our Topline	98.68	±0.000754	97.98	±0.00093	98.33

Table 42. Performance comparison of OOV on SIGHAN 2006 AS corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	65.19	±0.015339	60.36	±0.015751	62.68
CNG	67.68	±0.01506	71.51	±0.014533	69.54
AVS	66.90	±0.015152	73.68	±0.01418	70.13
TCB	65.86	±0.015268	61.53	±0.015666	63.62
TCF	67.47	±0.015085	62.17	±0.015616	64.71
AVS+TCB	67.31	±0.015104	*74.18*	±0.014092	70.58
AVS+TCF	67.94	±0.015028	**74.33**	±0.014065	*70.99*
LRAVS	67.73	±0.015054	72.89	±0.014314	70.21
LRAVS+TCB	*68.25*	±0.014989	73.34	±0.014238	70.70
LRAVS+TCF	**69.62**	±0.014808	73.89	±0.014143	**71.69**
2006 Best	70.20	±0.014727	N/A	N/A	N/A
2006 Baseline	3.00	±0.005493	N/A	N/A	N/A
Our Baseline	2.98	±0.005476	5.86	±0.00756	3.95
2006 Topline	99.70	±0.001761	N/A	N/A	N/A
Our Topline	99.64	±0.001936	97.17	±0.005341	98.39

Table 43. Performance comparison of accuracy on SIGHAN 2006 CityU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	96.92	±0.000736	96.88	±0.000741	96.90
CNG	97.26	±0.000696	97.21	±0.000701	97.23
AVS	97.31	±0.000690	**97.34**	±0.000686	97.32
TCB	96.95	±0.000733	96.89	±0.000740	96.92
TCF	96.96	±0.000732	96.90	±0.000739	96.93
AVS+TCB	97.32	±0.000689	97.32	±0.000689	97.32
AVS+TCF	**97.35**	±0.000685	97.32	±0.000688	*97.33*
LRAVS	**97.35**	±0.000684	97.32	±0.000688	**97.34**
LRAVS+TCB	*97.34*	±0.000686	*97.33*	±0.000687	**97.34**
LRAVS+TCF	97.23	±0.000700	97.26	±0.000696	97.24
2006 Best	97.20	±0.000703	97.30	±0.000691	97.20
2006 Baseline	88.20	±0.002134	93.00	±0.001687	90.60
Our Baseline	88.22	±0.001374	93.06	±0.001083	90.57
2006 Topline	98.50	±0.000804	98.20	±0.000879	98.40
Our Topline	98.55	±0.00051	98.19	±0.000568	98.37

Table 44. Performance comparison of OOV on SIGHAN 2006 CityU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	78.35	±0.008738	69.60	±0.009759	73.72
CNG	79.66	±0.008540	76.97	±0.008932	78.29
AVS	79.27	±0.008600	78.08	±0.008777	78.67
TCB	78.55	±0.008708	69.97	±0.009725	74.01
TCF	78.94	±0.008651	69.94	±0.009728	74.17
AVS+TCB	79.31	±0.008595	77.93	±0.008798	78.61
AVS+TCF	79.70	±0.008533	78.30	±0.008745	78.99
LRAVS	**79.84**	±0.008512	78.32	±0.008742	*79.07*
LRAVS+TCB	*79.82*	±0.008514	*78.57*	±0.008706	**79.19**
LRAVS+TCF	79.48	±0.008568	**77.93**	±0.008798	78.70
2006 Best	78.70	±0.008686	N/A	N/A	N/A
2006 Baseline	0.90	±0.002004	N/A	N/A	N/A
Our Baseline	0.95	±0.002053	2.47	±0.003293	1.37
2006 Topline	99.30	±0.001769	N/A	N/A	N/A
Our Topline	99.31	±0.001752	95.22	±0.004526	97.22

Table 45. Performance comparison of accuracy on SIGHAN 2006 PKU corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	92.51	±0.001338	93.79	±0.001227	93.14
CNG	*93.54*	±0.001250	94.38	±0.001170	93.96
AVS	93.43	±0.001259	*94.41*	±0.001167	93.92
TCB	92.54	±0.001335	93.75	±0.001230	93.14
TCF	92.54	±0.001335	93.72	±0.001233	93.13
AVS+TCB	93.43	±0.001259	94.37	±0.001171	93.90
AVS+TCF	93.42	±0.001260	94.32	±0.001176	93.87
LRAVS	**93.59**	±0.001245	**94.44**	±0.001164	**94.01**
LRAVS+TCB	*93.54*	±0.001250	94.40	±0.001168	*93.97*
LRAVS+TCF	93.40	±0.001262	94.30	±0.001178	93.85
2006 Best	92.60	±0.001330	94.00	±0.001207	93.30
2006 Baseline	79.00	±0.002694	86.90	±0.002231	82.80
Our Baseline	79.04	±0.002069	86.87	±0.001717	82.77
2006 Topline	97.60	±0.001012	96.10	±0.00128	96.80
Our Topline	97.59	±0.000779	96.08	±0.000986	96.83

Table 46. Performance comparison of OOV on SIGHAN 2006 PKU corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	70.51	±0.007834	70.70	±0.00782	70.60
CNG	74.97	±0.007442	**78.04**	±0.007112	*76.47*
AVS	74.57	±0.007481	77.78	±0.007142	76.14
TCB	70.73	±0.007817	70.90	±0.007804	70.81
TCF	70.96	±0.007799	70.19	±0.007859	70.57
AVS+TCB	74.51	±0.007487	77.68	±0.007154	76.06
AVS+TCF	74.14	±0.007522	77.13	±0.007215	75.61
LRAVS	**75.28**	±0.007411	*77.93*	±0.007125	**76.58**
LRAVS+TCB	*75.13*	±0.007427	77.68	±0.007154	76.38
LRAVS+TCF	74.53	±0.007486	77.03	±0.007226	75.76
2006 Best	70.70	±0.007819	N/A	N/A	N/A
2006 Baseline	1.10	±0.001792	N/A	N/A	N/A
Our Baseline	1.11	±0.001803	3.42	±0.003124	1.68
2006 Topline	98.90	±0.001792	N/A	N/A	N/A
Our Topline	98.94	±0.001762	92.56	±0.004507	95.65

Table 47. Performance comparison of accuracy on SIGHAN 2006 MSR corpus.

Configuration	P	C_P	R	C_R	F
6-tag	**96.44**	±0.001169	95.71	±0.001279	*96.08*
CNG	96.19	±0.001208	95.58	±0.001298	95.88
AVS	96.30	±0.001191	95.84	±0.001260	96.07
TCB	*96.40*	±0.001177	95.74	±0.001275	96.07
TCF	96.35	±0.001183	95.69	±0.001283	96.02
AVS+TC	96.38	±0.001180	*95.87*	±0.001256	**96.12**
AVS+TCF	96.40	±0.001177	95.73	±0.001276	96.06
LRAVS	96.22	±0.001203	95.85	±0.001259	96.04
LRAVS+TCB	96.24	±0.001200	**95.88**	±0.001255	96.06
LRAVS+TC	96.16	±0.001213	95.85	±0.001259	96.01
2006 Best	96.10	±0.001222	96.40	±0.001176	96.30
2006 Baseline	90.00	±0.001984	94.90	±0.001455	92.40
Our Baseline	90.03	±0.001891	94.94	±0.001384	92.42
2006 Topline	99.30	±0.000551	99.10	±0.000625	99.20
Our Topline	99.28	±0.000534	99.08	±0.000603	99.18

Table 48. Performance comparison of OOV on SIGHAN 2006 MSR corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	**66.57**	±0.016171	55.62	±0.017031	60.60
CNG	61.60	±0.016672	58.23	±0.016906	59.87
AVS	64.60	±0.016393	*60.83*	±0.016733	*62.66*
TCB	66.86	±0.016136	55.95	±0.017018	60.92
TCF	*66.42*	±0.016189	54.67	±0.017065	59.97
AVS+TCB	64.72	±0.016380	**61.19**	±0.016705	**62.91**
AVS+TCF	62.78	±0.016571	59.86	±0.016803	61.28
LRAVS	63.92	±0.016462	59.94	±0.016797	61.87
LRAVS+TCB	62.87	±0.016563	60.40	±0.016765	61.61
LRAVS+TCF	62.96	±0.016554	59.56	±0.016824	61.21
2006 Best	61.20	±0.016704	N/A	N/A	N/A
2006 Baseline	2.20	±0.005028	N/A	N/A	N/A
Our Baseline	2.17	±0.004999	11.13	±0.010780	3.64
2006 Topline	99.90	±0.001083	N/A	N/A	N/A
Our Topline	99.85	±0.001313	99.24	±0.002975	99.55

Table 49. Performance comparison of accuracy on SIGHAN 2008 AS corpus.

Configuration	***P***	C_P	***R***	C_R	***F***
6-tag	82.36	±0.002526	83.25	±0.002475	82.80
CNG	83.00	±0.002490	83.77	±0.002444	83.38
AVS	**83.09**	±0.002484	**83.83**	±0.002440	**83.46**
TCB	82.28	±0.002531	83.20	±0.002478	82.74
TCF	82.54	±0.002516	83.37	±0.002468	82.95
AVS+TCB	82.83	±0.002499	83.62	±0.002453	83.23
AVS+TCF	82.97	±0.002492	*83.80*	±0.002442	83.38
LRAVS	82.98	±0.002491	83.78	±0.002443	83.38
LRAVS+TCB	*83.03*	±0.002488	*83.80*	±0.002442	*83.42*
LRAVS+TCF	82.86	±0.002498	83.72	±0.002447	83.29
2008 Best	94.40	±0.001527	95.01	±0.001445	94.70
2008 Baseline	82.32	±0.002534	89.78	±0.002012	85.69
Our Baseline	80.99	±0.002601	89.29	±0.002050	84.93
2008 Topline	98.80	±0.000723	98.23	±0.000876	98.52
Our Topline	98.53	±0.000796	97.84	±0.000963	98.19

Table 50. Performance comparison of OOV on SIGHAN 2008 AS corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	62.85	±0.011258	55.49	±0.011580	58.94
CNG	63.78	±0.011199	**63.07**	±0.011245	**63.42**
AVS	63.38	±0.011225	*62.50*	±0.011280	*62.94*
TCB	62.42	±0.011285	55.61	±0.011576	58.82
TCF	63.61	±0.011210	56.22	±0.011560	59.69
AVS+TCB	*62.89*	±0.011256	60.88	±0.011371	61.87
AVS+TCF	63.60	±0.011211	61.80	±0.011321	62.68
LRAVS	63.30	±0.01123	62.19	±0.011298	62.74
LRAVS+TCB	63.34	±0.011228	62.27	±0.011294	62.80
LRAVS+TCF	**62.81**	±0.011261	61.71	±0.011326	62.25
2008 Best	74.04	±0.010215	76.49	±0.009881	75.24
2008 Baseline	2.08	±0.003325	6.78	±0.005858	3.19
Our Baseline	4.03	±0.004583	8.08	±0.006348	5.38
2008 Topline	99.32	±0.001915	96.42	±0.004329	97.84
Our Topline	99.40	±0.001795	96.41	±0.004337	97.88

Table 51. Performance comparison of accuracy on SIGHAN 2008 CTB corpus.

Configuration	***P***	***C_P***	***R***	***C_R***	***F***
6-tag	95.56	±0.001682	95.51	±0.001691	95.54
CNG	95.54	±0.001686	95.53	±0.001688	95.54
AVS	95.68	±0.001660	95.71	±0.001655	95.70
TCB	95.54	±0.001687	95.54	±0.001687	95.54
TCF	95.52	±0.001689	95.54	±0.001685	95.53
AVS+TCB	95.58	±0.001680	95.61	±0.001674	95.59
AVS+TCF	**95.98**	±0.001605	**95.96**	±0.001609	**95.97**
LRAVS	95.55	±0.001684	95.56	±0.001682	95.56
LRAVS+TCB	95.53	±0.001687	95.56	±0.001683	95.55
LRAVS+TCF	*95.69*	±0.001658	*95.72*	±0.001653	*95.71*
2008 Best	95.96	±0.001386	95.83	±0.001408	95.89
2008 Baseline	84.27	±0.002563	88.64	±0.002234	86.40
Our Baseline	84.05	±0.002991	88.86	±0.002570	86.39
2008 Topline	98.25	±0.000923	97.10	±0.001181	97.67
Our Topline	98.42	±0.001018	97.55	±0.001264	97.98

Table 52. Performance comparison of OOV on SIGHAN 2008 CTB corpus.

Configuration	***R_{OOV}***	***C_{Roov}***	***P_{OOV}***	***C_{Poov}***	***F_{OOV}***
6-tag	77.63	±0.014611	70.56	±0.01598	73.92
CNG	76.28	±0.014915	74.58	±0.015266	75.42
AVS	77.69	±0.014597	*75.87*	±0.015001	76.77
TCB	77.69	±0.014597	70.71	±0.015955	74.04
TCF	77.69	±0.014597	71.03	±0.015904	74.21
AVS+TCB	77.20	±0.014710	75.14	±0.015153	76.16
AVS+TCF	**78.86**	±0.014316	77.43	±0.014657	**78.14**
LRAVS	77.11	±0.014731	75.21	±0.015139	76.15
LRAVS+TCB	77.04	±0.014745	75.19	±0.015142	76.11
LRAVS+TCF	*78.15*	±0.014488	**76.50**	±0.014865	*77.32*
2008 Best	77.30	±0.014687	77.61	±0.014615	77.45
2008 Baseline	2.83	±0.005814	7.69	±0.009341	4.14
Our Baseline	1.54	±0.004313	3.34	±0.006298	2.10
2008 Topline	99.20	±0.003123	97.07	±0.005913	98.12
Our Topline	99.54	±0.002375	97.56	±0.005409	98.54

Table 53. Performance comparison of accuracy on SIGHAN 2008 NCC corpus.

Configuration	P	C_P	R	C_R	F
6-tag	93.55	±0.001259	93.09	±0.001300	93.32
CNG	**93.84**	±0.001232	**93.90**	±0.001226	**93.87**
AVS	93.69	±0.001246	93.72	±0.001243	93.71
TCB	93.60	±0.001254	93.14	±0.001295	93.37
TCF	93.46	±0.001267	93.11	±0.001298	93.28
AVS+TCB	*93.79*	±0.001237	93.78	±0.001238	93.78
AVS+TCF	93.75	±0.001240	93.81	±0.001235	93.78
LRAVS	93.76	±0.001240	93.83	±0.001233	93.79
LRAVS+TCB	93.78	±0.001238	*93.86*	±0.001230	*93.82*
LRAVS+TCF	93.73	±0.001242	93.81	±0.001235	93.77
2008 Best	94.07	±0.001210	94.02	±0.001214	94.05
2008 Baseline	87.16	±0.001714	92.00	±0.001390	89.51
Our Baseline	87.18	±0.001713	91.99	±0.001391	89.52
2008 Topline	98.17	±0.000687	97.35	±0.000823	97.76
Our Topline	98.17	±0.000687	97.35	±0.000823	97.76

Table 54. Performance comparison of OOV on SIGHAN 2008 NCC corpus.

Configuration	R_{OOV}	C_{Roov}	P_{OOV}	C_{Poov}	F_{OOV}
6-tag	*62.32*	±0.0114	51.51	±0.011758	56.40
CNG	60.43	±0.011504	*59.39*	±0.011554	*59.90*
AVS	59.76	±0.011537	57.86	±0.011617	58.79
TCB	**63.28**	±0.011341	52.30	±0.011751	57.27
TCF	62.86	±0.011367	52.73	±0.011745	57.35
AVS+TCB	60.30	±0.011511	58.43	±0.011595	59.35
AVS+TCF	59.91	±0.01153	58.64	±0.011586	59.27
LRAVS	60.08	±0.011522	59.31	±0.011557	59.69
LRAVS+TCB	60.32	±0.01151	**59.49**	±0.011549	**59.90**
LRAVS+TCF	60.23	±0.011514	59.21	±0.011562	59.72
2008 Best	61.79	±0.011431	59.84	±0.011533	60.80
2008 Baseline	2.73	±0.003834	18.58	±0.00915	4.76
Our Baseline	2.73	±0.003831	18.58	±0.009151	4.75
2008 Topline	99.33	±0.001919	92.03	±0.006372	95.54
Our Topline	99.34	±0.001911	92.04	±0.006368	95.55

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 87-108

Strategies of Processing Japanese Names and Character Variants in Traditional Chinese Text

Chuan-Jie Lin[*], Jia-Cheng Zhan[*], Yen-Heng Chen[*], and

Chien-Wei Pao[*]

Abstract

This paper proposes an approach to identify word candidates that are not Traditional Chinese, including Japanese names (written in Japanese Kanji or Traditional Chinese characters) and word variants, when doing word segmentation on Traditional Chinese text. When handling personal names, a probability model concerning formats of names is introduced. We also propose a method to map Japanese Kanji into the corresponding Traditional Chinese characters. The same method can also be used to detect words written in character variants. After integrating generation rules for various types of special words, as well as their probability models, the F-measure of our word segmentation system rises from 94.16% to 96.06%. Another experiment shows that 83.18% of the 862 Japanese names in a set of 109 human-annotated documents can be successfully detected.

Keywords: Semantic Chinese Word Segmentation, Japanese Name Identification, Character Variants.

1. Introduction

Word segmentation is an indispensable technique in Chinese NLP. Nevertheless, the processing of Japanese names and Chinese word variants has been studied rarely. At the time when Traditional Chinese text was mostly encoded in BIG5, writers often transcribed a Japanese person's name into its equivalent Traditional Chinese characters, such as the name "滝沢秀明" (Hideaki Takizawa) in Japanese becoming "瀧澤秀明" in Traditional Chinese. After Unicode was widely adopted, we also could see names written in original Japanese Kanji in Traditional Chinese text. Another issue is how different regions may write a character

[*] Department of Computer Science and Engineering, National Taiwan Ocean University
No 2, Pei-Ning Road, Keelung, 20224 Taiwan
E-mail: cjlin@ntou.edu.tw; jjt@cyber.ntou.edu.tw; {M97570019, M97570020}@ntou.edu.tw

in a different shape. For example, the Traditional Chinese character 圖 (picture) is written as 图 in Simplified Chinese and 図 in Japanese. How these character variants impact Chinese text processing has been mentioned rarely in earlier studies; thus, it has become our interest.

Chinese word segmentation has been studied for a long time. Many recent word segmentation systems have been rule-based or probabilistic. The most common rules are longest-word-first or least-segmentation-first. The probability models are often built in Markov's unigram or bigram models, such as in Peng and Chang (1993). Word candidate sets are often vocabulary in a dictionary or a lexicon collected from a large corpus. Some systems also propose possible candidates by morphological rules (Gao *et al.*, 2003), such as NOUN+"們" (plural form of a noun) as a legal word (*e.g*. "學生們," students, and "家長們," parents). Wu and Jiang (1998) even integrated a syntactic parser in their word segmentation system.

In addition to word segmentation ambiguity, the out-of-vocabulary problem is another important issue. Unknown words include rare words (*e.g*. "躉售," for sale); technical terms (*e.g*. "三聚氰胺," Melamine, a chemical compound); newly invented terms (Chien, 1997) (*e.g*. "新流感," Swine flu); and named entities, such as personal and location names. NE recognition is an important related technique (Sun *et al.*, 2003). In recent times, machine learning approaches have been the focus of papers on Chinese segmentation, such as using SVM (Lu, 2007) or CRF (Zhao *et al.*, 2006; Shi & Wang, 2007).

There have been fewer studies focused on handling words that are not Traditional Chinese words in Traditional Chinese text. The most relevant work is discussion of the impact of the different Chinese vocabulary used in different areas on word segmentation systems. These experiments have been designed to train a system with a Traditional Chinese corpus but test on a Simplified Chinese test set or to increase the robustness of a system using a lexicon expanded by adding new terms in different areas (Lo, 2008).

The main problem in this paper is defined as follows. When a word that is not Traditional Chinese appears in a Traditional Chinese document, such as the Japanese name "滝沢秀明" (written in Japanese Kanji) or "瀧澤秀明" (written in its equivalent Traditional Chinese), word variants (*e.g*. "裡面" vs. "裏面"), and words written in Simplified Chinese, all of these words can be detected and become word segmentation candidates. This paper is constructed as follows. Section 2 introduces the basic architecture of our word segmentation system. Section 3 explains the Chinese and Japanese name processing modules. Section 4 talks about the character-variant clusters with a corresponding Traditional Chinese character. Section 5 delivers the experimental results and discussion, and Section 6 concludes this paper.

2. Word Segmentation Strategy

This paper focuses on approaches to handling words that are not Traditional Chinese during word segmentation. We first constructed a basic bigram model word segmentation system. We did not build a complicated system because its purpose is only for observing the effect of applying different handling approaches for words that are not written in Traditional Chinese on the performance of word segmentation. Word candidates were identified by searching the lexicon or applying detection rules for special-type words, such as temporal or numerical expressions. Note that identical word candidates may be proposed by different rules or the lexicon. Moreover, if no candidate of any length can be found at a particular position inside the input sentence, the system automatically adds a one-character candidate at that position. Afterward, the probabilities of all of the possible segmentations are calculated according to a bigram model. The highest probable segmentation is proposed as the result.

2.1 Special-Type Word Candidate Generation Rules

As there are many special type words, it is impossible to collect them all in a lexicon. Hence, we manually designed many detection rules to recognize such words in an input sentence. The special types handled in our system include the following: address, date, time, monetary, percentage, fraction, Internet address (IP, URL, e-mail, *etc*.), number, string written in foreign language, and Chinese and Japanese personal name. Numerical digits in the detection rules can be full-sized or Chinese numbers (一,二…壹貳…). Foreign language characters are detected according to the Unicode table; thus, any character sets, such as Korean or Arabic characters, easily can be added into our system. Consequent characters written in the same foreign language are treated as one word, as most languages use the space symbol as the word-segmentation mark.

Since the focus of this paper is not on the correctness of these special rules, only personal name detection rules will be explained in Section 3.

2.2 Bigram Probabilistic Model

After enumerating all possible segmentations, the next step is to calculate their probabilities $P(S)$. There have been many probabilistic models proposed in word segmentation research. Our system is built on Markov's bigram probabilistic model, whose definition is:

$$P(S = w_1 w_2 ... w_N) = P(w_1) \times \prod_{i=2}^{N} P(w_i \mid w_{i-1}) \qquad (1)$$

where $P(w_i)$ is the unigram probability of the word w_i and $P(w_i \mid w_{i-1})$ is the probability that w_i appears after w_{i-1}. In order to avoid the underflow problem, the equation is often calculated in its logistic form:

$$\log P(S = w_1 w_2 ... w_N) = \log P(w_1) + \sum_{i=2}^{N} \log P(w_i \mid w_{i-1}) \tag{2}$$

Data sparseness is an apparent problem, *i.e.* most word bigrams have no probability. Our solution is a unigram-back-off strategy. That is, when a bigram $<w_{i-1}, w_i>$ never occurs in a training corpus, its bigram probability $P(w_i \mid w_{i-1})$ is measured by $\alpha P(w_i)$ instead.

When determining the probability of a bigram containing special-type words, the probability is calculated by Eq. 3. Suppose that w_i belongs to a special type T; the equation is defined as:

$$P(w_i \mid w_{i-1}) P(w_{i+1} \mid w_i) = P(T \mid w_{i-1}) \times P(w_{i+1} \mid T) \times P_G(w_i \mid T) \tag{3}$$

where $P(T \mid w_k)$ and $P(w_k \mid T)$ are the special-type bigram probabilities for the type T and a word w_k and where $P_G(w_i \mid T)$ is the generation probability of w_i being in the type T. The generation probabilities are set to 1 for all special types other than the personal names, whose definitions are explained in Section 3.

As the boundaries of some special types, including address, monetary, percentage, fraction, Internet address, number, and foreign language string, are deterministic and unambiguous, their special-type bigram probabilities are all set to be 1, which means that we accept the segmentation directly.

On the other hand, characters for Chinese numbers often appear as a part of a word, such as “一切” (“一,” one; “一切,” all) and “萬一” (both characters are numbers but together mean “if it happens”). Therefore, the number-type bigram probability is trained from a corpus.

Some temporal expressions are unambiguous, such as the date expression “中華民國九十八年六月二十一日” (“June 21 of the 98th year of the R.O.C.”). Their special-type bigram probabilities are set to 1. For ambiguous temporal expressions, such as “三十年” (meaning “the 30th year” or “thirty years”), their special-type bigram probabilities are obtained by training.

Before training a bigram model, words belonging to special types first are identified by detection rules and replaced by labels representing their types so that special-type bigram probabilities can be measured at the same time.

Our special-type bigram probability model is very similar to Gao *et al.* (2003). Nevertheless, they treat all dictionary words as one class and all types of special words as a second class, while we treat different types as different classes.

2.3 Computation Reduction

When an input sentence is too long or too many possible segmentations can be found (sometimes hundreds of thousands), the computation time becomes intractable. In order to reduce the computation load, we use the beam search algorithm to prune some low probability segmentations. The main idea of the algorithm is described as follows.

Let *N* be the number of characters in an input sentence. Declare *N* priority queues (denoted as record[*i*] where *i* = 1~*N*) to record the top *k* segmentations with the highest probability scores covering the first *i* characters. For each word candidate *w* beginning with the $(i+1)^{th}$ character whose length is *b*, append the word *w* with every segmentation stored in record[*i*], compute the probability of the new segmentation, and try to insert it into the queue record[*i*+*b*]. If the new segmentation has higher probability than any segmentation stored in the queue record[*i*+*b*], the segmentation with the lowest probability in record[*i*+*b*] is discarded in order to insert this new segmentation.

At the beginning, all priority queues are empty. Start with the first character in the sentence. Recursively perform the steps described in the previous paragraph until all of the word candidates starting with the N^{th} character have been considered. In the end, the top 1 segmentation stored in record[*N*] is proposed as the result. The queue size *k* is set to be 20 in our system.

3. Chinese and Japanese Name Processing

In this section, we focus on how to find Japanese names written in Japanese Kanji that appear in a Traditional Chinese article. The method of identifying Japanese names written in corresponding Traditional Chinese characters is discussed in Section 4. As our approach to recognize Japanese personal names is similar to the one to find Chinese names, our Chinese name identification approach is introduced first.

3.1 Chinese Personal Name Identification

A Chinese personal name consists of a surname part and a first name part. A Chinese surname can be one or two syllables (one or two characters) long. In some cases, a person may have two surnames (usually both with one syllable) in his or her name for various reasons. The first name part in a Chinese name is also one or two syllables long. All name formats possibly seen in an article are listed in Table 1, where “SN” denotes “surname,” “FN” as “first name,” and “char” is “character”.

All strings matching these formats are treated as Chinese name candidates, except the format “1-char FN,” in order to prevent proposing every single character as a personal name candidate. The combination of two surnames is also restricted to two 1-syllable surnames, because one rarely sees a 2-syllable surname combined with another surname. We need to build probabilistic models for each character being in every part of a name, as well as a probabilistic model for the personal name formats.

Table 1. Chinese personal name formats **(*surnames are underlined*)**

Format	Cases	Examples	Format	Cases	Examples
SN only	1-char SN	Prof. 林	SN+FN	1-char SN+1-char FN	陳登
	2-char SN	Mr. 諸葛		1-char SN+2-char FN	王小明
FN only	1-char FN	慧		Two SNs+1-char FN	張 李娥
	2-char FN	國雄		Two SNs+2-char FN	張 陳素珠
				2-char SN+1-char FN	諸葛亮
				2-char SN+2-char FN	司馬中原

To recognize a Chinese name, first we have to prepare a complete list of Chinese surnames. We collected surnames from the Wikipedia entries "中國姓氏列表"[1] (List of Chinese Surnames) and "複姓"[2] (2-Syllable Surnames), the websites of the Department of Civil Affairs at the Ministry of Interior[3], 中華百家姓[4] (GreatChinese), and 千家姓[5] (Thousand Surnames). 2,471 surnames were collected. As for the first name part, we simply treat all of the Chinese characters as possible first name characters.

The generation probability model of a word being a Chinese name is defined as Eq. 4, where σ is the gender model (male or female), and π is a possible format matching the word w. The name format is represented as π = 'xxxx,' where 's' denotes a 1-syllable surname, 'dd' a 2-syllable surname, and 'n' a character in a first name. For example, the format "two SNs+2-char FN" is represented as π = 'ssnn' and the format "2-char SN+1-char FN" is represented as π = 'ddn'.

$$P_G(w\,|\,S_{CHname}) = \max_{\sigma,\pi} P_\sigma(w\,|\,\pi)P_G(\pi\,|\,S_{CHname}) \tag{4}$$

In Eq. 4, the ***Chinese name generation probability*** $P_\sigma(w|\pi)$ is the probability of a word w being a Chinese name whose format is π and gender is σ. The ***Chinese name format probability*** $P_G(\pi\,|\,S_{CHname})$ is the probability of the special type S_{CHname} (Chinese personal names) appearing in an article with a format π. The methods of building these probabilistic models are introduced in the following paragraphs.

When computing the Chinese name generation probability $P_\sigma(w|\pi)$, we borrowed the idea from Chen *et al*. (1998), but we assume that the choice of first names is independent of the surname, and the choice of two characters in the first name part is also independent, in order to reduce the complexity. We also assume that the surname is unrelated to the person's gender. Table 2 lists all of the definitions of the Chinese name generation probabilities for

[1] http://zh.wikipedia.org/wiki/中國姓氏列表

[2] http://zh.wikipedia.org/wiki/複姓

[3] http://www.ris.gov.tw/ch4/0940531-2.doc

[4] http://www.greatchinese.com/surname/surname.htm

[5] http://pjoke.com/showxing.php

every format, where LN_{CH} is the set of Chinese surnames and FN_{CH} is the set of characters used in a Chinese first name. A more sophisticated model may be applied but is outside the scope of this paper.

Table 2. Definitions of the Chinese name probabilities for every name format

Format π	Name Generation Probability $P_\sigma(w\|\pi)$	Format Probability
s	$P_G(c_1\|LN_{CH})$	$P_G(\pi=\text{‘s’}\|S_{CHname})$
dd	$P_G(c_1c_2\|LN_{CH})$	$P_G(\pi=\text{‘dd’}\|S_{CHname})$
sn	$P_G(c_1\|LN_{CH})\times P_\sigma(c_2\|FN_{CH})$	$P_G(\pi=\text{‘sn’}\|S_{CHname})$
nn	$P_\sigma(c_1\|FN_{CH})\times P_\sigma(c_2\|FN_{CH})$	$P_G(\pi=\text{‘nn’}\|S_{CHname})$
ddn	$P_G(c_1c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})$	$P_G(\pi=\text{‘ddn’}\|S_{CHname})$
snn	$P_G(c_1\|LN_{CH})\times P_\sigma(c_2\|FN_{CH})\times P_\sigma(c_3\|FN_{CH})$	$P_G(\pi=\text{‘snn’}\|S_{CHname})$
ssn	$P_G(c_1\|LN_{CH})\times P_G(c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})$	$P_G(\pi=\text{‘ssn’}\|S_{CHname})$
ddnn	$P_G(c_1c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})\times P_\sigma(c_4\|FN_{CH})$	$P_G(\pi=\text{‘ddnn’}\|S_{CHname})$
ssnn	$P_G(c_1\|LN_{CH})\times P_G(c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})\times P_\sigma(c_4\|FN_{CH})$	$P_G(\pi=\text{‘ssnn’}\|S_{CHname})$

The generation probability models for surnames and first name characters, $P_G(c_i|LN_{CH})$, $P_G(c_ic_{i+1}|LN_{CH})$ and $P_\sigma(c_j|FN_{CH})$, are trained from a large corpus by maximum likelihood:

1-char SN:	$P_G(c_i\|LN_{CH})$	=	count(c_i) / count(names)
2-char SN:	$P_G(c_ic_{i+1}\|LN_{CH})$	=	count(c_ic_{i+1}) / count(names)
FN char:	$P_\sigma(c_j\|FN_{CH})$	=	count(c_j) / count(FN chars) of gender σ

We adopted a list of one million personal names in Taiwan to build the probabilistic models. The list contains 476,269 male names and 503,679 female names. There are only 953 surnames and 4,000 more first name characters seen in the name list. For those unseen surnames and first name characters, we assign them a small probability (10^{-1000}, tuned by experiments) to avoid the zero probability problem.

The next step is to build the Chinese name format probability $P_G(\pi \mid \mathrm{S}_{\mathrm{CHname}})$. Since it is about the probability of a name format appearing in an article, the distribution is quite different from the ones observed in the list of one million personal names. A person is often mentioned in an article by his or her title, *e.g*. “Prof. 林” (“Prof. Lin”) or “Mr. 諸葛” (“Mr. Zhu-Ge). When referring to a person in a novel or a letter, it is quite natural to give his or her first name instead of his or her full name. Such cases cannot be captured inside the one million personal names list. Therefore, we need another corpus to train this model.

Personal names in the Academia Sinica Balanced Corpus (*Sinica Corpus* hereafter) are marked as proper nouns (POS-tagged as Nb). We extracted all of the proper nouns in the Sinica Corpus that matched any name format and assumed them to be personal names. These names occur in real documents; thus, they can satisfy our need. The precedence of format matching is defined as follows. Every personal name can only be matched to one format.

1-char word：s > n > not-Chinese-personal-name
2-char word：dd > sn > nn > not-Chinese-personal-name
3-char word：ddn > snn > ssn > not-Chinese-personal-name
4-char word：ddnn > ssnn > not-Chinese-personal-name
5-char word：not-Chinese-personal-name

Nevertheless, for the reason that some common characters are uncommon surnames, it is possible to identify a proper noun of some other type incorrectly as a personal name, such as "中興號" ("Zhong Xing Hao," a bus company name) where "中" ("Zhong") is also a surname. In order to increase the precision without sacrificing the recall, we used only frequent surnames and first name characters to do the matching. The sets of frequent characters are the ones that dominate 90% of the probabilities in the name generation model, including 64 surnames (陳,林…程), 467 male first name characters (文,明…瀛), and 293 female first name characters (美,淑…吉), together with all of the 2-syllable surnames.

There are two more formats seen in articles: SN+"姓" or SN+"氏", which call a person or a family, respectively, by the surname only. We denote them as π = 'p'. After implementing the matching procedure described above, 39,612 of the 92,314 proper nouns in the Sinica Corpus were extracted as personal names. The Chinese name format probabilities are listed in Table 3. Although there may still be false-alarm personal names in the set, we expect the scale of the corpus is large enough that it can still provide relatively accurate information. The identified personal names in the corpus also can be used to build the bigram models related to the special type S_{CHname}, Chinese personal name.

Table 3. The Chinese name format probabilities

Format Probability	Count	Prob.	Format Probability	Count	Prob.
$P_G(\pi=\text{‘s’}\|S_{CHname})$	5,431	13.71%	$P_G(\pi=\text{‘ddn’}\|S_{CHname})$	126	0.32%
$P_G(\pi=\text{‘n’}\|S_{CHname})$	815	2.06%	$P_G(\pi=\text{‘snn’}\|S_{CHname})$	19,454	49.11%
$P_G(\pi=\text{‘p’}\|S_{CHname})$	487	1.23%	$P_G(\pi=\text{‘ssn’}\|S_{CHname})$	58	0.15%
$P_G(\pi=\text{‘dd’}\|S_{CHname})$	46	0.12%	$P_G(\pi=\text{‘ddnn’}\|S_{CHname})$	24	0.06%
$P_G(\pi=\text{‘sn’}\|S_{CHname})$	2,845	7.18%	$P_G(\pi=\text{‘ssnn’}\|S_{CHname})$	61	0.15%
$P_G(\pi=\text{‘nn’}\|S_{CHname})$	10,265	25.91%	Total	39,612	

An example is given here to illustrate how the probability of a personal name is determined. The word "張德培," ("Michael Te Pei Chang") matches two name formats, $\boldsymbol{\pi}$ = {'snn', 'ssn'}, since both "張" ("Chang") and "德" ("Te") are possible surnames. Genders options are male and female, *i.e.* $\boldsymbol{\sigma}$ = {M, F}. The most probable one is a male name with the format 'snn'.

Name: 張德培		
π	σ	Probability
snn	M	$\log (P_G(張 \mid LN_{CH})\times P_M(德 \mid FN_{CH})\times P_M(培 \mid FN_{CH})\times P_G(\pi=\text{‘snn’} \mid S_{CHname}))$ $= (-1.26) + (-1.87) + (-2.74) + (-0.31) = -6.18$
snn	F	$\log (P_G(張 \mid LN_{CH})\times P_F(德 \mid FN_{CH})\times P_F(培 \mid FN_{CH})\times P_G(\pi=\text{‘snn’} \mid S_{CHname}))$ $= (-1.26) + (-2.89) + (-3.27) + (-0.31) = -7.73$
ssn	M	$\log (P_G(張 \mid LN_{CH})\times P_G(德 \mid LN_{CH})\times P_M(培 \mid FN_{CH})\times P_G(\pi=\text{‘ssn’} \mid S_{CHname}))$ $= (-1.26) + (-6.02) + (-2.74) + (-2.82) = -12.84$
ssn	F	$\log (P_G(張 \mid LN_{CH})\times P_G(德 \mid LN_{CH})\times P_F(培 \mid FN_{CH})\times P_G(\pi=\text{‘ssn’} \mid S_{CHname}))$ $= (-1.26) + (-6.02) + (-3.27) + (-2.82) = -13.37$

3.2 Japanese Personal Name Identification

When a Japanese name occurs in an article written in Chinese, there are two ways to write the name. In earlier days, when Traditional Chinese was usually encoded in BIG5, a Japanese name normally was written in its corresponding Traditional Chinese characters, such the name "滝沢秀明," Hideaki Takizawa, a Japanese performer, would be written as "瀧澤秀明" in Traditional Chinese. Nowadays, many documents are encoded in Unicode, so Japanese Kanji can be directly used in a Traditional Chinese article. Our word segmentation approach wants to identify both cases.

The format of a Japanese personal name is SN+FN, just like a Chinese name. Nevertheless, the length of a Japanese surname varies from one to three Kanji characters, as does the length of the first name part. Sometimes, a name is directly written in Katakana or Hiragana with various lengths. The number of Kanji or Kana characters in a Japanese name is strongly correlated to the number of syllables. Due to the lack of related knowledge, we only deal with the names written in all Kanji and leave the cases of names including Kana as a future work, although Kana can be detected easily by Unicode ranges.

Table 4. Japanese name formats **(*surnames are underlined*)**

Format	SN	FN	SN+FN
Example	木村 長谷川	理惠 新一	伊藤由奈 高橋留美子

As the length of Japanese names varies considerably, we only adopt three name formats, SN-only, FN-only, and SN+FN, without regarding the number of characters inside the first name part, as listed in Table 4. We know that there is no double surname in Japan.

From the experience of Chinese name processing, we know that a list of Japanese surnames and a large collection of Japanese personal names are needed in order to build name generation probability models. Also, we have to find a corpus of Chinese articles containing

Japanese names in order to build the format probability model as well as the special-type bigram probability. The probability of a Japanese personal name is defined as follows.

$$P_G(w \mid S_{JPname}) = \max_{\pi} P_G(w \mid \pi) P_G(\pi \mid S_{JPname}) \qquad (5)$$

The notations in Eq. 5 are defined as the same as in Eq. 4. One difference is that, because we do not have a large training corpus for different genders, the factor of gender in the name generation probability is omitted. Table 5 lists the definitions of each probability, where *m* and *n* are integers between 1 and 3, 'S' denotes the surname part, and 'F' denotes the first name part. Surnames and first names are also assumed to be independent, as are the characters inside a first name part.

Table 5. Definitions of the Japanese name probabilities for every format

Format	Name Generation Probability $P(w\|\pi)$	Format Probability
SN	$P_G(c_{1\ldots}c_m\|LN_{JP})$	$P_G(\pi=\text{'S'}\|S_{JPname})$
FN	$P_G(c_1\|FN_{JP})\times\ldots\times P_G(c_n\|FN_{JP})$	$P_G(\pi=\text{'F'}\|S_{JPname})$
SN+FN	$P_G(c_{1\ldots}c_m\|LN_{JP})\times P_G(c_{m+1}\|FN_{JP})\times\ldots\times P_G(c_{m+n}\|FN_{JP})$	$P_G(\pi=\text{'SF'}\|S_{JPname})$

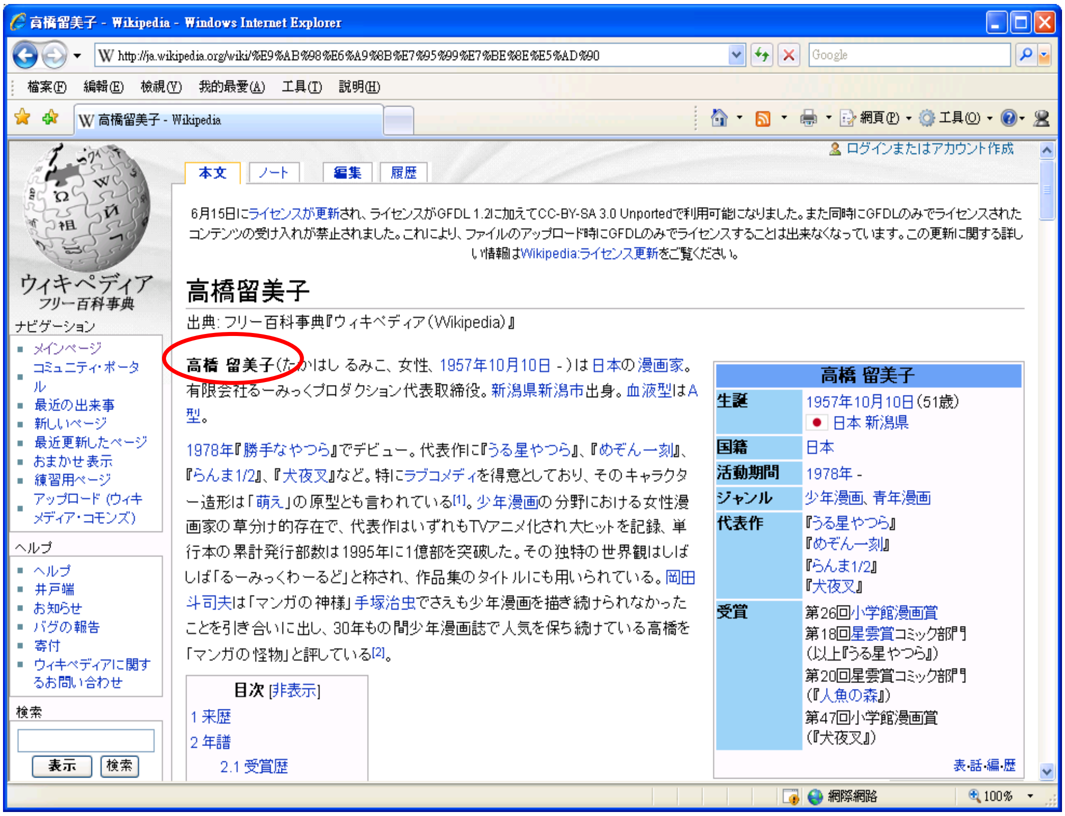
Japanese surnames were collected from a website called "日本の苗字七千傑"[6] (7,000 Surnames in Japan). This website provides 8,603 Japanese surnames along with their populations, where data came from the 117 million costumers of NTT, a Japanese Telecom company. The population data can be used to measure the distributions of the surnames. Nevertheless, according the Wikipedia entry "日文姓名,"[7] there are more than 140 thousand Japanese surnames, far more than we have collected. No complete list is available so far. Moreover, we still need another data set to train the probabilities of first name characters.

All of the Japanese Wikipedia entries that deliver biographies of persons were extracted for learning Japanese personal name distributions. In a Wikipedia page, the title of the entry will also be mentioned again in the text and marked in bold type. The surname part is often separated from the first name part by a space, as in the example of the entry "高橋留美子" ("Rumiko Takahashi"), shown in Figure 1. By detecting such kinds of strings, we can gather many Japanese names in a short time.

Nevertheless, Chinese celebrities may also become entries in the Japanese Wikipedia, such as "王建民" ("Chien-Ming Wang") or "曾國藩" ("Zeng Guofan"). We filtered out the names with a known Chinese surname and a first name part less than three characters. After processing the entire Japanese Wikipedia dumped on Jan 24, 2009 by the methods described above, 65,778 different Japanese names were extracted, including 12,907 surnames and 2,320

[6] http://www.myj7000.jp-biz.net

[7] http://zh.wikipedia.org/wiki/日文姓名


高橋留美子 - Wikipedia - Windows Internet Explorer
http://ja.wikipedia.org/wiki/%E9%AB%98%E6%A9%8B%E7%95%99%E7%BE%8E%E5%AD%90
Google
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
高橋留美子 - Wikipedia
網頁(P) 工具(O)
ログインまたはアカウント作成
本文
ノート
編集
履歴
ウィキペディア
フリー百科事典
ナビゲーション
メインページ
コミュニティ・ポータル
最近の出来事
新しいページ
最近更新したページ
おまかせ表示
練習用ページ
アップロード (ウィキメディア・コモンズ)
ヘルプ
ヘルプ
井戸端
お知らせ
バグの報告
寄付
ウィキペディアに関するお問い合わせ
検索
表示
検索
網際網路
100%

6月15日にライセンスが更新され、ライセンスがGFDL 1.2に加えてCC-BY-SA 3.0 Unportedで利用可能になりました。また同時にGFDLのみでライセンスされたコンテンツの受け入れが禁止されました。これにより、ファイルのアップロード時にGFDLのみでライセンスすることは出来なくなっています。この更新に関する詳しい情報はWikipedia:ライセンス更新をご覧ください。

高橋留美子

出典: フリー百科事典『ウィキペディア（Wikipedia）』

高橋 留美子（たかはし るみこ、女性、1957年10月10日 - ）は日本の漫画家。有限会社るーみっくプロダクション代表取締役。新潟県新潟市出身。血液型はA型。

1978年『勝手なやつら』でデビュー。代表作に『うる星やつら』、『めぞん一刻』、『らんま1/2』、『犬夜叉』など。特にラブコメディを得意としており、そのキャラクター造形は「萌え」の原型とも言われている[1]。少年漫画の分野における女性漫画家の草分け的存在で、代表作はいずれもTVアニメ化され大ヒットを記録、単行本の累計発行部数は1995年に1億部を突破した。その独特の世界観はしばしば「るーみっくわーるど」と称され、作品集のタイトルにも用いられている。岡田斗司夫は「マンガの神様」手塚治虫でさえも少年漫画を描き続けられなかったことを引き合いに出し、30年もの間少年漫画誌で人気を保ち続けている高橋を「マンガの怪物」と評している[2]。

目次 [非表示]

1 来歴
2 年譜
2.1 受賞歴

高橋 留美子	
生誕	1957年10月10日（51歳） 日本 新潟県
国籍	日本
活動期間	1978年 -
ジャンル	少年漫画、青年漫画
代表作	『うる星やつら』 『めぞん一刻』 『らんま1/2』 『犬夜叉』
受賞	第26回小学館漫画賞 第18回星雲賞コミック部門 （以上『うる星やつら』） 第20回星雲賞コミック部門 （『人魚の森』） 第47回小学館漫画賞 （『犬夜叉』）

表·話·編·歴

Figure 1. The Wikipedia entry page “高橋留美子”

first name Kanji. Table 6 lists the frequencies of these first name Kanji, where the name generation probabilities $P_G(c_j|FN_{JP})$ are listed in the third column and the accumulated probabilities are in the fourth column.

Table 6. Frequencies of the Japanese first name Kanji

FN Kanji	Freq	$P_G(c_j\|FN_{JP})$	Accm Prob.	FN Kanji	Freq	$P_G(c_j\|FN_{JP})$	Accm Prob.
子	4,821	3.60%	3.60%	亭	46	0.03%	89.99%
一	3,358	2.50%	6.10%	瑞	46	0.03%	90.03%
郎	3,237	2.41%	8.52%	…	…	…	…
美	2,230	1.66%	10.18%	褒	1	0.00%	99.99%
正	1,741	1.30%	11.48%	焔	1	0.00%	100.00%
…	…	…	…	Totally 2,320 Kanji; total freq = 134,055			

Many surnames collected from the Japanese Wikipedia did not appear in the surname list of “日本の苗字七千傑”. The two lists were merged and resulted in a list of 15,702 surnames.

The population data provided by “日本の苗字七千傑” or the frequencies in Wikipedia were used to estimate the generation probabilities of the surnames, as listed in Table 7. Note that surnames from “佐藤” to “高井良” come from “日本の苗字七千傑,” and the surnames after “斉藤” were collected from Wikipedia.

Table 7. Population of Japanese surnames

SN	Freq	Gen. Prob. $P_G(c_{1\ldots}c_m \vert LN_{JP})$	SN	Freq	Gen. Prob. $P_G(c_{1\ldots}c_m \vert LN_{JP})$
佐藤	1928000	1.65%	高井良	760	6.49×10^{-6}
鈴木	1707000	1.46%	斉藤	111	9.47×10^{-7}
高橋	1416000	1.21%	三遊亭	106	9.05×10^{-7}
田中	1336000	1.14%	⋯	…	…
渡辺	1135000	0.97%	城土	1	8.54×10^{-9}
伊藤	1080000	0.92%	駒尾	1	8.54×10^{-9}
⋯	…	…	Totally 15,702 surnames; total = 117,156,792		

The Japanese name format probability $P_G(\pi \mid S_{JPname})$ was also built by detecting Japanese names in the Sinica Corpus, but only on those proper nouns that were not determined to be Chinese names. Moreover, since the Japanese names in the Sinica Corpus are encoded in Traditional Chinese characters, the matching procedure also includes corresponding Kanji-mapping, which will be explained in Section 4.2.

When extracting Japanese names in the Sinica Corpus, only the 437 first name Kanji (子, 一…瑞), which cover 90% of the probabilities, are used, along with the whole Japanese surname set. The preference of the formats is SN+FN > SN > FN. Each name matched one format at most. After doing so, 4,849 of the 92,314 proper nouns in the Sinica Corpus were extracted as Japanese names. They were used to build the format probability model (as listed in Table 8) as well as the special-type bigram probability for the Japanese name type S_{JPname}. In our experience, however, the format FN-only often suggests too many incorrect candidates and harms the performance of word segmentation. In the end, we elected not to use it.

Table 8. Japanese name format probabilities

Format Probability	$P_G(\pi$=‘S’$\vert S_{JPname})$	$P_G(\pi$=‘F’$\vert S_{JPname})$	$P_G(\pi$=‘SF’$\vert S_{JPname})$	Total
Frequency	718	1,120	3,011	4,849
Probability	14.90%	23.24%	62.48%	

An example is given here to illustrate how the probability of a personal name is determined. The name “滝沢光” matches the Japanese name format in two ways: “滝沢” (“Takizawa”) as a surname and “光” (“Hikaru”) as a first name, or “滝” (“Taki”) the surname and “沢光”

(“Sawahikari”[8]) the first name. The highest probability suggests “滝沢” as a surname and “光” as a first name.

Name: 滝沢光	
Format	Probability
SN+FN	$\log (P_G(\text{滝沢}\vert LN_{JP})\times P_G(\text{光}\vert FN_{JP})\times P_G(\pi=\text{‘SN’}\vert S_{JPname}))$ $= (-7.35) + (-5.15) + (-0.076)$ $= -12.576$
SN+FN	$\log (P_G(\text{滝}\vert LN_{JP})\times P_G(\text{沢}\vert FN_{JP})\times P_G(\text{光}\vert FN_{JP})\times P_G(\pi=\text{‘SN’}\vert S_{JPname}))$ $= (-10.70) + (-9.40) + (-5.15) + (-0.076)$ $= -25.326$

4. Character Variant Handling

This section discusses three cases where character variants may be used: (1) a Japanese name written in its corresponding Chinese characters (*e.g*. “滝沢秀明” vs. “瀧澤秀明,” Hideaki Takizawa); (2) equivalent words in variant forms (*e.g*. “裡面” vs. “裏面,” inside); (3) Simplified Chinese terms (*e.g*. “體育館” vs. “体育馆”, the gym) appearing in a Traditional Chinese article. Although the last two cases are not often seen, especially the third case (which could not happen until Unicode was invented), we still propose approaches to handle these cases at the same time for the possibility of building a multilingual environment.

4.1 Mapping of Character Variants

A mapping table between the character variants is required for handling the three cases introduced in the previous paragraph. For Japanese names, we need a list of Japanese Kanji and their corresponding Chinese characters. For word variants, a list of the equivalent Chinese character set is necessary. The mapping between Simplified Chinese terms and the corresponding Traditional Chinese ones requires mapping between the two character sets, which is more easily acquired because there are many kinds of software providing such a mapping function.

We do not know of any well-known Japanese-Chinese Kanji mapping tables. To construct one, we adopted the character variant list[9] developed by Prof. Koichi Yasuoka and Motoko Yasuoka in the Institute for Research in Humanities, Kyoto University. There are 8,196 character variant pairs collected in the list. Following the equivalent relationship, we grouped characters in the list into many character-variant clusters. Some examples of character-variant clusters are given here.

[8] In fact, “沢光” (“Sawahikari”) is a Japanese surname and rarely used as a first name.

[9] http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/ftp/CJKtable/UniVariants.Z

丰 豊 豐 靊 靊
秇 蓺 蕓 藝
乹 乾 乾 干 漧

Note that these variants are equivalent only in some cases. Take the first cluster illustrated above as an example. The character "豊" is Japanese Kanji and "丰" is a Simplified Chinese character, and they both correspond to the Traditional Chinese character "豐". Nevertheless, "豊" (ritual vessel) and "丰" (elegance) are also legal Traditional Chinese characters that have different meanings from the one of "豐" (prosperous).

In each character-variant cluster, one Traditional Chinese character (if any) is chosen to be the *corresponding* character. If there is more than one Traditional Chinese character in a cluster, the most frequent one is chosen. The frequencies of characters are provided by the Table of Frequencies of Characters in Frequent Words[10] (常用語詞調查報告書之字頻總表) published by the Taiwan Ministry of Education in 1998. Again, considering the first cluster in the examples above, the three characters "丰," "豊," and "豐" are all Traditional Chinese characters. "豐" is the most frequent one; hence, it is chosen as the corresponding character of this cluster. By doing so, not only do the Japanese Kanji "豊" and the Simplified Chinese character "丰" have a corresponding Traditional Chinese character, but also the infrequent variants "靊" and "靊" can have a frequent corresponding character.

There are many issues in variant mapping. First, the Traditional Chinese set is larger than the BIG5 character set. Relatively infrequent Traditional Chinese characters, such as "靊," are not seen in the BIG5 set. Since we are looking for the most frequent Traditional Chinese character, this will not become a problem.

Another issue is the time when two variant characters can be regarded as equivalent. As we have mentioned, the character "豊" is equivalent to "豐" only when it is used as Japanese Kanji. Its meaning in Traditional Chinese is a ritual vessel in ancient times (*cf*. Revised Mandarin Chinese Dictionary[11], 重編國語辭典修訂本), which is completely different from the current meaning of "豐" (prosperous). This would be an interesting future topic.

4.2 Finding Corresponding Chinese Characters for Japanese Kanji

When extracting Japanese personal names inside the Sinica Corpus (as described in Section 3.2), the mapping between Japanese Kanji and Traditional Chinese characters is necessary. Characters in the tables of Japanese surnames and first name Kanji need to be transformed into Traditional Chinese first.

[10] http://www.edu.tw/files/site_content/M0001/87news/index.htm

[11] http://dict.revised.moe.edu.tw/cgi-bin/newDict/dict.sh?cond=%E0T&pieceLen=50&fld=1&cat=&ukey=1838907571&op=&imgFont=1

Each Kanji in a Japanese surname was changed into its corresponding Traditional Chinese character found by the method explained in Section 4.1. For example, the surname “滝沢” (Takizawa) was changed into “瀧澤” and “中曽根” (Nakasone) was changed into “中曾根”. The newly created surnames were merged into the original Japanese surname table, and they shared the same probabilities with the original Japanese surnames. If at least one Kanji character in a surname did not have a corresponding Traditional Chinese character (*e.g.* “畑” in the surname “古畑,” Huruhata), no new surname would be created. The first name Kanji table was expanded in a similar way, along with the assignment of the probabilities.

Merging a newly created term into the name probability table makes our system capable of identifying various methods of name writing at the same time. Our system can identify the two equivalent names in the sentence “滝沢聡就是瀧澤聰” (which means, “滝沢聡 then is 瀧澤聰”). We can see that “滝沢” and “瀧澤” can be found in the Japanese surname table, just as “聡” and “聰” are found in the Japanese first name table. Both “滝沢聡” and “瀧澤聰” are proposed as word candidates that are Japanese names and share the same probability.

Following the same idea, if we further expand the correspondent relationship to the Simplified Chinese character set, it will be possible to understand the sentence “滝沢聡和泷泽聡都是瀧澤聰” (“滝沢聡 and 泷泽聡 both are 瀧澤聰”), where “滝沢聡” is in Japanese, “滝沢聡” is in Simplified Chinese, and “瀧澤聰” is in Traditional Chinese. This part has not yet been implemented but is quite promising.

4.3 Generating Word Variants

In order to identify word variants written either in character variants or in Simplified Chinese, we expanded the dictionary vocabulary by changing the characters in a Traditional Chinese word into their character variants (including Simplified Chinese characters). For example, given a Traditional Chinese word, ABC, each character is searched in the character-variant clusters introduced in Section 4.1. Every character variant found in the character-variant clusters is used to enumerate all possible word variants. Supposing that A’, A”, B’, and C’ are variants of the characters A, B, and C, the following word variants will be enumerated: A’BC, AB’C, ABC’, A’B’C, AB’C’, A’BC’, A’B’C’, A”BC, A”B’C, A”BC’, and A”B’C’.

The newly enumerated word shares the same probability as its original form. Instead of merging the word variants and attaining a large dictionary, we assigned each group of the word variants a unique ID and indexed the bigram probability table (for word segmentation) by the group IDs.

Since the mapping between Simplified Chinese characters and Traditional Chinese characters is not one-to-one, there may be identical word variants enumerated from different words. For example, the Simplified Chinese word variants of “白面” (white-faced) vs. “白麵”

(white noodles) are both “白面,” and the Simplified Chinese word variants “改制” (rule changing) and “改製” (producing in a different model) are the same term “改制,” too. To determine the final probability of an ambiguous word variant, we experimented on three strategies where the final probability is the maximum, the minimum, or the sum of all of the probabilities of the original words. Section 5.4 reveals the results of this experiment.

5. Experiment

5.1 Experimental Data and Evaluation Metrics

The experimental data for word segmentation is the Academia Sinica Balanced Corpus, Version 3.0[12]. The Sinica corpus is designed for language analysis purposes. Words in a sentence are separated by spaces and tagged with their POSs. The documents are written in Modern Mandarin and collected from different domains and topics. There are 316 files containing 743,718 sentences.

Our evaluation was done by 5-fold cross-validation. The 316 files were divided into 5 sets. Each set was used as the test set iteratively when the other sets were used as the development set to construct the lexicon and train probability models. The number of sentences in each set is given in Table 9.

Table 9. Number of sentences in the experimental data

File ID	Test Set ID	No of Files	Sentences	Known Words	Unknown Words
000~065	ASBCset0	66	148,575	146,477	15675
066~129	ASBCset1	64	149,713	146,275	15877
130~183	ASBCset2	54	148,870	146,634	15518
184~244	ASBCset3	61	148,012	146,024	16128
245~315	ASBCset4	71	148,548	146,004	16148

The performance of word segmentation was evaluated by the following metrics, precision, recall, F-measure, and BI score:

$$precision = \frac{\text{correct words being segmented}}{\text{number of words segmented by the system}} \quad (6)$$

$$recall = \frac{\text{correct words being segmented}}{\text{number of words segmented in the test set}} \quad (7)$$

$$\text{F-}measure = \frac{2 \times recall \times precision}{recall + precision} \quad (8)$$

[12] http://godel.iis.sinica.edu.tw/CKIP/20corpus.htm

$$\text{BI-score} = \frac{\text{correct BI labels}}{\text{number of total characters in the test set}} \tag{9}$$

The BI-score labels are defined as follows. Given a sentence, each character is labeled as B (at the beginning of a word) or I (inside a word) according to the segmentation in the test set or the segmentation generated by the system. The ratio of correct BI labels also reveals the performance of a word segmentation system.

When evaluating using 5-fold cross-validation, we used micro-averaging to calculate the scores. That is, the values of the denominators and the numerators of precision, recall, and BI-score are the sums over the five experiment sets.

5.2 Word Segmentation Baseline Performance

This section shows the performance of our basic-model word segmentation system. System Sys1a uses only the known-word lexicon with bigram probability model. System Sys1b integrates the special-type word generation rules, including address, date, time, monetary, percentage, fraction, foreign string, and Internet address, as introduced in Section 2.2. The Sys2 systems further integrate the numbers, including Arabic and Chinese numbers. In order to see the impact of directly adopting the boundary of a number candidate, we experimented on two strategies for Sys2, denoted as Sys2a and Sys2b. As shown in Table 10, Sys1b performs better because of the integration of special-type word generation rules. The maximum-likelihood probability model for numbers is also a better choice.

- **Sys2a: Number generation probability is set to be 1**
- **Sys2b: Number generation probability is trained by maximum likelihood**

Table 10. Performance of the basic word segmentation integrated with special-type word generation rules

System	**R**	**P**	**F**	**BI**
Sys1a	95.66	92.72	94.16	96.96
Sys1b	95.87	93.31	94.57	97.20
Sys2a	95.97	93.57	94.76	97.30
Sys2b	**96.16**	**93.68**	**94.90**	**97.38**

5.3 Experiments on Handling Chinese and Japanese Personal Names

After integrating the Chinese personal name generation rules, the special-type probability for Chinese names is also employed. The difference between our work and Chen *et al.* (1998) is the use of Chinese name format probability and allowing personal names without surnames. Three systems were designed to observe the impact.

- **Sys3a: Using the Chinese name special-type probability, but not the format π = ‘nn’ and the format probability**
- **Sys3b: Using the Chinese name special-type probability with the format π = ‘nn’ but not the format probability**
- **Sys3c: Using the Chinese name special-type probability with the format π = ‘nn’ and the format probability**

All Sys3 systems are based on Sys2b. The evaluation results are shown in Table 11. We can see that all of these methods (using the special-type probability for Chinese name, the name format of FN-only, and the Chinese name format probability) improve the performance. This confirms the success of name formats in personal name recognition and word segmentation.

Table 11. Performance after integrating Chinese name processing

System	**R**	**P**	**F**	**BI**
Sys3a	96.39	94.97	95.68	97.90
Sys3b	96.42	95.49	95.95	98.05
Sys3c	**96.57**	**95.53**	**96.04**	**98.10**

Two systems were designed to observe the effectiveness of the Japanese name special-type probability and the format probability. As the test set is encoded in BIG5, the Japanese name processing is performed under the BIG5 Traditional Chinese character set. Both Sys4 systems are based on Sys3c.

- **Sys4a: Using the Japanese name special-type probability without the format probability**
- **Sys4b: Using both the Japanese name special-type probability and the format probability**

Table 12. Performance after integrating Japanese name processing

System	**R**	**P**	**F**	**BI**
Sys3c	**96.57**	95.53	96.04	98.10
Sys4a	96.54	95.54	96.04	98.10
Sys4b	96.56	**95.56**	**96.06**	98.10

Table 12 illustrates the performance after integrating Japanese name processing. We found that using only the Japanese name special-type probability resulted in a decline of the word segmentation performance, while using both probability models outperformed Sys3c, but not significantly. The reason may be the small number of Japanese names appearing in the Sinica Corpus, as we know that only 4,849 words in the 743,718 sentences were considered to be Japanese names (*cf.* Section 3.2). The improvement of Japanese name processing did not affect the performance of word segmentation significantly.

In order to observe the real performance of Japanese name processing, we designed another experiment. A collection of 109 news articles was prepared, and the Japanese names in it were manually annotated. 862 occurrences of 216 distinct Japanese names were found.

Two kinds of observations were performed. The first one was to verify the ratio of Japanese names being correctly segmented before and after the integration of Japanese name processing. The results are shown in Table 13, which were obtained by applying Sys3c and Sys4b on the 109 documents. This confirms that integrating Japanese name processing greatly improves the success rate.

Table 13. Ratio of Japanese names successfully being segmented

System	**Number of Successfully Segmented Japanese Names**	**Ratio**
Sys3c	154	17.87%
Sys4b	717	83.18%
Total	862	

The second observation is to measure the precision and recall of Japanese name recognition. That is, the ratio of correct ones among the Japanese name candidates proposed by the system (precision) and the ratio of correctly proposed ones among the Japanese names in the test set (recall). The results are listed in Table 14, where both recall and precision are about 75%, which is fair correctness but still needing improvement. This also shows that Japanese name processing is not an easy problem.

Table 14. Precision and recall of Japanese name recognition

System	P	R
Sys4b	74.31% (648/872)	75.17% (648/862)

Some examples of correct and incorrect word segmentation results before and after integrating the Japanese name processing are given here.

Successful examples:

Sys3c	Sys4b	Sys3c	Sys4b
小 林恭二	小林恭二	大 前 研一	大前研一
石原慎 太郎	石原慎太郎	藥師 丸 博子	藥師丸博子

Incorrect examples:

Sys3c	Sys4b	Sys3c	Sys4b
麻布 和 木材	麻布和 木材	瓦斯井 原有	瓦斯 井原有
國小 林佩萱 老師	國 小林 佩萱 老師	廣島 亞運 時	廣島亞運時

5.4 Word Variant Experiments

This section discusses the performance of handling word variants. Unfortunately, we cannot find a suitable test set that contains annotations of character variants. The documents in the Sinica Corpus are encoded in BIG5, a subset of Traditional Chinese characters. There are only a few character variants appearing in the Sinica Corpus.

Two experimental datasets were constructed for the evaluation. The first dataset was a copy of the Sinica Corpus with every character transformed into its Simplified Chinese form (the mapping is unambiguous and can be done by a lot of software). This dataset can be used to verify the ability of Simplified Chinese word handling of a word segmentation system. It can also be used to decide the probabilistic model for homographic variants from different words. The second one was a real corpus written in Simplified Chinese.

As mentioned in Section 4.3, the character mapping from Simplified Chinese to Traditional Chinese is many-to-one. It is possible that a Simplified Chinese word is related to two or more different Traditional Chinese words. Three systems were designed to determine the unigram or bigram probability for such homographic word variants: Sys5a chose the maximum probability among the corresponding Traditional Chinese words, Sys5b chose the minimum, and Sys5c used the sum of the probabilities. Note that Chinese and Japanese name processing also suffers from this problem if the names are written in Simplified Chinese characters. To focus on word variant handling, the experiments were performed without personal name processing. All Sys5 systems were developed based on Sys2b, a system that has not integrated the name processing module. The evaluation results are listed in Table 15. We can see that the method of probability determination does not affect the performance as much, which also shows that the system is capable of dealing with Simplified words in Traditional Chinese text. We chose Sys5a, the one with the maximum values, as our final system.

- **Sys5a: Using the maximal probability of the corresponding source words**
- **Sys5b: Using the minimal probability of the corresponding source words**
- **Sys5c: Using the sum of the probabilities of the corresponding source words**

Table 15. Probability model determination for homographic variants

System	**R**	**P**	**F**	**BI**
Sys5a	96.11	93.53	94.80	97.33
Sys5b	95.95	93.16	94.54	97.21
Sys5c	96.11	93.53	94.80	97.33

The second experiment was done on GHAN 1st Peking University Test Set, a Simplified Chinese word segmentation benchmark. The test set contained 380 sentences. We did not use its development set and lexicon to train our system. Instead, we used Sys5a and the lexicon constructed from the Sinica Corpus. The experimental results show that the performance is worse, where precision is 86.56%, recall is 81.47%, and F-measure is 83.94%. This is because the documents in the Peking University Test Set came from Mainland China, where the vocabulary is quite different from the one in Taiwan. The lower performance is not surprising. The main purpose of this experiment is to show that our system can do word segmentation on documents written in Simplified Chinese with a certain correctness level.

6. Conclusion

In this paper, we propose methods to find word candidates that are Japanese personal names (written in either Japanese Kanji or their equivalent Traditional Chinese characters) or word variants when doing word segmentation. Documents are encoded in UTF-8 so that characters in different languages can appear in the same document. Our word segmentation is based on a bigram probabilistic model, and it integrates the generation rules and probability models for different kinds of special types of words.

When handling Chinese and Japanese personal names, we propose the idea of the name format probability model and discuss how the model can be built. We also propose a method to find corresponding Traditional Chinese characters for Japanese Kanji so that a Japanese name can be detected whenever it is written in a different language. The experimental results show that the name format probability model does improve the performance, and the mappings between Japanese Kanji and Traditional Chinese characters do help to detect Japanese names more successfully.

The size of the Japanese surname list in our system, which contains only 15,702 surnames, is far less than the amount of 140 thousand mentioned in Wikipedia. Nevertheless, once a larger Japanese surname list can be found, it can be easily integrated into our system as long as we assign a small probability to those unseen surnames for smoothing. Furthermore, our knowledge in Japanese name processing is still not sufficient. As a future work, a syllable probabilistic model regarding the pronunciation of a name will be studied. The most important of all is to find a large collection of Japanese names for training.

Using the character variant clusters, Chinese words written in any character variants can be successfully detected as word candidates. Although the set of newly enumerated word variants is large, the computational complexity remains the same if denoting word variants by their group ID and using hash tables to do searching.

Reference

Chen, H.H., Ding, Y.W., Tsai S.C., & Bian, G.W. (1998). Description of the NTU System Used for MET2. In *Proceedings of 7th Message Understanding Conference* (*MUC-7*). Available: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.

Chien, L.F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of SIGIR97*, 27-31.

Gao, J., Li, M., & Huang, C.N. (2003). Improved Source-Channel Models for Chinese Word Segmentation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (*ACL 2003*), 272-279.

羅永聖 (Lo) (2008). 結合多類型字典與條件隨機域之中文斷詞與詞性標記系統研究, Master Thesis, National Taiwan University.

Lu, X. (2007). Combining machine learning with linguistic heuristics for Chinese word segmentation. In *Proceedings of the FLAIRS Conference*, 241-246.

彭載衍 (Peng) and 張俊盛 (Chang) (1993). 中文辭彙歧義之研究－斷詞與詞性標示. In 第六屆中華民國計算語言學研討會論文集 (*ROCLING-6*), 173-194.

Shi, Y. & Wang, M. (2007). A dual-layer CRFs based joint decoding method for cascaded segmentation and labeling tasks. In *Proceedings of International Joint Conference on Artificial Intelligence* (*IJCAI '07*), 2007, 1707-1712.

Sun, J., Zhou, M., & Gao, J.F. (2003). A Class-based Language Model Approach to Chinese Named Entity Identification. In *International Journal of Computational Linguistics and Chinese Language Processing*, 8(2), 1-28.

Wu, A. & Jiang, Z. (1998). Word segmentation in sentence analysis. In *Proceedings of the 1998 International Conference on Chinese Information Processing*, 169-180.

Zhao, H., Huang, C.N., & Li, M. (2006). An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 162-165.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 109-128

Evaluation of TTS Systems in Intelligibility and Comprehension Tasks: a Case Study of HTS-2008 and Multisyn Synthesizers[1]

Yu-Yun Chang[*]

Abstract

This paper explores the relationship between intelligibility and comprehensibility in speech synthesizers, and it designs an appropriate comprehension task for evaluating the speech synthesizers' comprehensibility. Previous studies have predicted that a speech synthesizer with higher intelligibility will have higher performance in comprehension. Also, since the two most popular speech synthesis methods are HMM-based and unit selection, this study tries to compare whether the HTS-2008 (HMM-based) or Multisyn (unit selection) speech synthesizer has better performance in application. Natural speech is applied in the experiment as a control group to the speech synthesizers. The results in the intelligibility test show that natural speech is better than HTS-2008, which, in turn, is much better than the Multisyn system. In the comprehension task, however, all three of the speech systems display minimal differences in the speech comprehension process. This is because the two speech synthesizers have reached the threshold of having enough intelligibility to provide high speech comprehension quality. Therefore, although there is equal comprehensible speech quality between the HTS-2008 and Multisyn systems, the HTS-2008 speech synthesizer is recommended due to its higher intelligibility.

Keywords: Speech Synthesizers, Intelligibility Evaluation, Comprehension Evaluation, HTS-2008, Multisyn.

[1] Acknowledgement: Thanks to the people from University of Edinburgh - for the suggestions and comments offered by Simon King and Catherine Mayo, the perception lab reserving and statistics help supported by Rob Clark, and the HTS-2008 speech synthesizer provided by Junichi Yamagishi.

[*] Graduate Institute of Linguistics, National Taiwan University, 3F, Le-Xue Building, No. 1, Sec. 4, Roosevelt Rd., Taipei Taiwan, 106
E-mail: june06029@gmail.com

1. Introduction

Recently, text-to-speech (TTS) system synthesizers have been evaluated from different aspects, such as intelligibility, naturalness, and preference of the synthetic speech, as noted by Stevens, Lees, Vonwiller, and Burnham (2005). Since the final purpose of applying synthetic speech is to make it usable to applications, carrying out experiments measuring the synthesizers' performance with human listeners is worthwhile.

In previous studies, while mentioning the evaluation of speech synthesizers, most researchers only focused on intelligibility evaluation due to the experiment being easy and quick to carry out. Nevertheless, it is necessary to involve perception factors in synthetic speech evaluation, rather than merely evaluating the intelligibility, in order to better assess speech synthesizers, as indicated by Pisoni, Nusbaum, and Greene (1985). Sydeserff, Caley, Isard, Jack, and Monaghan (1992) also evaluated the aspect of the listener's perception on a comprehension task to learn how well synthetic speech could be understood by the listeners. Moreover, Pisoni *et al*. (1985) demonstrated that intelligibility had a strong impact on comprehension, and specified that intelligibility was one of the important factors affecting listening comprehension. Thus, it is worth observing the linkage between intelligibility and comprehension in speech synthesizers.

Although several studies have evaluated the intelligibility of speech synthesizers successfully, very few researchers have examined its effect on comprehension. This may be because the comprehension measuring experiment is difficult to construct, as it involves cognitive processes that are difficult to capture and take into account. Recent studies have taken post-perceptual comprehension tests instead to investigate listeners' comprehension, but many have failed to distinguish differences between TTS systems. An appropriate strategy for evaluating comprehension still has not been found. Therefore, this research is intended to design an adequate comprehension test for speech synthesis evaluation and to discover the effect of intelligibility on comprehension.

In this study, the word "intelligibility" means the degree of accuracy with which each word is produced in a sentence and the word "comprehension" means the degree of received messages being understood. This study assumes that intelligibility has a strong influence on comprehension, which indicates that speech synthesizers with higher intelligibility can be expected to obtain higher comprehension. In addition, this paper also compares the latest version of speech synthesizers used in the Blizzard Challenge (Black & Tokuda, 2005), which are the unit selection (Clark, Richmond, & King, 2007) based Multisyn synthesizer (Clark, *et al*., 2007) and the hidden Markov models (HMMs) (Zen *et al*., 2007) based HTS-2008 synthesizer (Yamagishi *et al*., 2008). Since these two speech synthesizers are built by adapting the most popular methods used in producing TTS systems, it will be interesting to find out whether the HMM-based or unit selection approach can generate better synthetic speech in terms of both intelligibility and comprehension.

2. Literature Review

2.1 HMM-based and Unit Selection Speech Synthesizers

In recent years, HMMs have been used to generate synthesized speech (Yoshimura, Tokuda, Masuko, Kobayashi, & Kitamura, 1999). The basic procedures of implementing HMM-based speech synthesizers to produce synthetic speech can be grouped into two parts: a training part and a synthesis part (Heiga & Tomoki, 2005). There are two main advantages of using HMMs to generate speech synthesizers. One is that the produced synthesized speech can be smoothed and made to sound natural. The other is that, since the synthetic speech is created from HMM models with parameters (Heiga & Tomoki, 2005), the characteristics of the voice can be modified easily with adequate parameter transformations. The latest version of the HTS (HMM-based Speech Synthesis System) used in the Blizzard Challenge is the HTS-2008. HTS-2008 used the speaker adaptive approach, rather than the speaker-dependent method, to generate HMM-based synthesizers. The training database used to create the average voice model for HTS-2008 was a 41-hour speech collection. In addition, to reduce the expensive computing time, the forward-backward algorithm was introduced in HTS-2008 (Yamagishi *et al.*, 2008).

As for the unit selection speech synthesizers, basically, a natural speech database will be recorded by a single speaker and the units are extracted directly from the speech inventory and concatenated together to generate new utterances. A number of different unit sizes can be used to construct various types of unit selection speech synthesizers, such as phones, half phones, diphones, and variable-sized units (Clark, Richmond, & King, 2004). In the recent Festival speech synthesis system, the Multisyn unit selection algorithm was introduced (Clark, *et al.*, 2007) with the diphone sized units, which could carry better acoustic features and higher-level linguistic information than the phone sized units used in CHATR (Hunt & Black, 1996) and clunits (Black & Taylor, 1997). It can produce open-domain speech voices in high speech quality and does not need to be based on the context domain speech to produce better quality. In other words, higher quality synthesized speech can be created using the Multisyn unit selection algorithm even if the synthesized utterance is not one of the sentences in the collected databases.

Since the Multisyn speech synthesis approach has the advantage of generating natural synthesized voices by extracting the diphone sized units straight from the speech signal with less expensive signal processing, an investigation of its distinction from the HTS-2008 HMM-based speech synthesizer would be interesting and useful.

2.2 Evaluation of Intelligibility

When evaluating the intelligibility of a speech synthesizer, semantically unpredictable sentences (SUS) are used frequently. SUS sentences have been widely used in dictation tasks and are recommended in evaluating intelligibility of speech synthesizers (Pols, van Santen, Abe, Kahn, & Keller, 1998). SUS sentences are sentences that are semantically unpredictable, but are still constructed grammatically syntactically. SUS sentences are used to prevent the process of assessing intelligibility from being influenced by linguistic cues. If semantically predictable sentences are used, listeners will learn the semantic and syntactic cues from the context, which will influence their performance in the intelligibility task (Benoît, Grice, & Hazan, 1996). They claimed that using SUS sentences in the intelligibility task could disrupt the predictable context. This conclusion was also supported by Miller and Isard (1963), reporting that using SUS sentences could prevent the learning effect.

2.3 Evaluation of Comprehension

The performance of various speech synthesizers can also be evaluated through comprehension tasks. Several researchers have indicated that comprehension evaluation is a valid way to assess intelligibility (Hustad, 2008; Yorkston, Strand, & Kennedy, 1996). This is because, in the intelligibility task, listeners will emphasize recognizing individual words, rather than focusing on the meaning of sentences. Nevertheless, the deeper information that lies within intelligibility cannot be examined by merely identifying each word.

There are four types of questions that have been used in speech synthesizer comprehension evaluation: surface structure questions, high proposition questions, low proposition questions, and inference questions. These questions were designed based on different levels of memory used during comprehension (Luce, 1981; Pisoni, Nusbaum, Luce, & Schwab, 1983; Salasoo, 1982). Surface structure questions required participants to recall specific words that occurred in the speech content. High proposition questions examined whether listeners could get a general idea from the speech content, whereas low proposition questions asked for more detailed information about the speech content than high proposition questions. Finally, the inference questions measured whether the listeners could draw a conclusion from the speech. Since surface structure questions did not involve much comprehension ability, which did not meet the purpose of the present experiment, this type of question was not included in the present study.

2.4 Some Influential Factors in Intelligibility and Comprehension

2.4.1 Short-term Memory

Short-term memory is the biggest cognitive factor influencing the comprehension task. This is because short-term memory is used to store fractions of information temporarily until full information can be completely comprehended. Therefore, the technique is essential during the comprehension task, and the load of short-term memory needs to be considered as well. As demonstrated from the concurrent task experiment by Ralston, Pisoni, and Mullennix (1989), short-term memory has limited capacity. Goldstein (1995) identified two different levels of short-term memory, which are the nominal level and supra-nominal level. He further said that nominal level short-term memory was involved in intelligibility tasks, focusing on qualitative evaluation, whereas supra-nominal level short-term memory was used in comprehension tasks, which required the information to be identified, processed, and understood. Therefore, as specified by previous researchers, it would be important to take short-term memory into account in this study.

2.4.2 Listeners' Preferences

Another factor that may influence task performance is the listeners' preferences. Nusbaum *et al*. (1984) judged listeners' preferences from listeners' feedback on one natural speech and two speech synthesizers, MITalk and Votrax. The measurement was to assess adjectives from the feedback. The researchers found that, although people preferred to listen to natural speech rather than the two speech synthesizers, they liked the MITalk system more than the Votrax system. Also, they investigated the intelligibility in the MITalk system and evaluated it as higher than the Votrax system. As indicated in the paper, this result showed that a relationship existed between the subjects' preferences and intelligibility of different speech synthesizers. Besides, Nusbaum, Francis, and Henly (1995) contended that listeners' preferences depended greatly on the quality of speech intelligibility. Moreover, Terken and Lemeer (1988) and Paris, Thomas, Gilson, and Kincaid (2000) found that, as the intelligibility got better, the degree of preference would also increase.

Therefore, in this paper, HTS-2008 and Multisyn systems would be taken as the representatives of HMM-based and unit selection speech synthesizers during the evaluation. Also, by modifying the evaluation approaches used in the previous studies and considering the cognitive factors, I try to design an appropriate comprehension test, which has not been found yet, rather than designing an intelligibility test. In addition, through the newly modified comprehension test, I hope that a stronger relationship of "higher intelligibility will gain better comprehension" could be revealed.

3. Methodology

3.1 Subjects

Twenty-five native English speakers participated in the experiment, with 6 males and 19 females.[2] Table 1 shows the subjects' level of education.

Table 1. Participants' level of education status

Degree of Education	Undergraduate	Master	PhD
Number of Subjects	5	11	9

All of the participants were students studying at University of Edinburgh at the time of the survey. There were 5 undergraduates, 11 master's students, and 9 PhD students involved in this experiment. The subjects' average age was 25.44 years old, with a standard deviation (SD) of 3.465 years.

Table 2. Participants' English accents

English Accent	British	American	Scottish	Irish	Welsh	Indian
Number of Subjects	13	6	3	1	1	1

Table 2 presents the survey results of the participants' English accents. The accent survey reported 13 people with a British accent, 6 with an American accent, 3 with a Scottish accent, 1 with an Irish accent, 1 with a Welsh accent, and 1 with an Indian accent. Additionally, only three participants indicated that they were speech experts. No one reported having a hearing disorder.

3.2 Materials

3.2.1 SUS Sentences for Intelligibility Evaluation

Thirty SUS sentences were used as the material in the intelligibility task. These SUS sentences were adopted from the 2008 Blizzard Challenge (Karaiskos, King, Clark, & Mayo, 2008). The structure of these sentences is "The (Determiner) + (Adjective) + (Noun) $_{\text{plural}}$ + (Verb) $_{\text{past tense}}$ + the (Determiner) + (Adjective) + (Noun) $_{\text{singular}}$". Although this was the only structure used in the experiment, the English words chosen to construct SUS sentences are all low-frequency words, in order to prevent the listeners from predicting meanings easily. For example, one of the sentences used in the experiment is "The amicable chests became the unprepared cockroach". As the example shows, the intelligibility task tends to make it difficult for

[2] Although the numbers of male and female participants were not balanced, the gender did not display any significance in statistical analysis. Therefore, the gender difference is not considered in this paper.

listeners to predict the unheard information. In addition, listening to each sentence more than once was allowed, but subjects were requested to keep this to as few times as possible.

3.2.2 News Articles for Comprehension Evaluation

Six news articles from BBC News online that were considered to contain few story line cues were used in the comprehension task. As in the study of Lai, Wood, and Considine (2000), in order to reduce the news articles' textual familiarity to the listeners, all of the topics chosen were research reports, which were likely to be less familiar to most of the listeners. The answers to the questions were designed with the assumption that there was no global and general knowledge to the articles. In other words, participants could not learn the answers to questions without listening. The average article was about 238.8 words (SD = 21.1 words).

Each news article was attached to ten questions. Five of the questions were designed as multiple-choice questions, while the other five questions were open-ended questions. Only the questions that required inferential skills would be arranged as multiple-choice questions with four choices. On the other hand, factual questions with low-level proposition information were assigned to open-ended questions. Figures 1 and 2 present examples of the questions involved in the main experiment.

Inferential Question

Question: What would be the best topic for the news?

A. The poor quality of recent education.
B. The cpmpetition between colleges.
C. Colleges face the financial crisis.
D. Education revolution.

Figure 1. An example of inferential question in the main experiment

Factual Question

Question: How long would the growth of stubble usually appears?

____

Figure 2. An example of factual question in the main experiment

3.2.3 Synthesized Speech and Natural Speech Recording

HTS-2008 and Multisyn speech synthesizers were included in this experiment. Both speech synthesizers were constructed by collecting the voice from a single male speaker with a British accent, "Roger". Also, the male speaker's natural speech was taken as a control group, to compare with the experimental materials (30 SUS sentences and 6 news articles) produced by the two synthesizers.

The recording was held in a sound lab of University of Edinburgh. The lab was equipped with a professional recording room and a control room. The voice was recorded through a Sennheiser MKH 800 microphone, with the volume set at 60 dB. The recorded wav files were all single channel, with a frequency of 16 kHz. The recording duration was approximately one hour.

The male speaker was a well-trained professional reader and had cooperated with the Centre for Speech Technology Research (CSTR) for a long while, participating in speech data recording. Therefore, steady and good quality natural speech was guaranteed.

3.2.4 Questionnaires

A questionnaire was assigned at the end of the experiment, asking for participants' basic information, whether they were speech experts, and the average number of times each sentence in the intelligibility task was played. Some empty blanks were left for participants to write down their comments and suggestions about the experiment.

3.3 Procedure

There were two tasks in the experiment. The first part was an intelligibility task (listening to 30 SUS sentences), and the other part was the comprehension task (listening to 6 BBC News reports and answering questions). The experiment took place at the Perception Lab in the Informatics Forum building. The lab consisted of individual rooms. Each room was equipped with a SAMSUNG 2043 screen monitor and a set of Beyerdynamic DT 770 PRO headphones. Every participant was arranged into one of the single rooms. The experiment was carried out by applying an online webpage. All of the voices would come from the headphones throughout the experiment, and the volume had been set to an adequate loudness for the listeners. No participants complained about the sound volume.

3.3.1 Producing Wav Files

For the intelligibility task and comprehension task, all wav files of SUS sentences and news passages were produced by natural speech and the two synthesizers, HTS-2008 and Multisyn. In order to generate higher-quality synthesized speech for news passages, all of the sentences in each article were synthesized individually before being concatenated together with a silence interval of about 500 milliseconds in between.

There were some cases needing careful consideration when producing synthesized speech, where the TTS systems could not identify the pronunciation as predicted in natural speech. For example, if the input text was “500MB,” the synthesizers would not be able to pronounce it as “five hundred megabytes”. Instead, the pronunciation turned out to be “five zero zero M B”. Since the purpose of this comprehension test was to measure whether the synthesized passages were comprehensible to listeners, every word in the experiment should be made understandable to listeners.

3.3.2 Pilot Tests for Comprehension Task

Since the material used in the intelligibility test was the same as in the Blizzard Challenge, pilot tests for evaluating the intelligibility test were unnecessary. Nevertheless, pilot tests were needed for the comprehension test in this study. The pilot tests for the comprehension test were done three times, measuring the length of the articles, the difficulty of the text and questions, and the familiarity of the text. Two native English speakers were invited to do the pilot test and help evaluate the design of the comprehension task.

3.3.3 Main Experiment

To make the wav files produced from HTS-2008, Multisyn, and natural speech equally distributed through the experiment, the wav files were equally arranged into 6 different groups via Latin Squares. Each group included 30 SUS sentences in the intelligibility test and 6 news articles in the comprehension test. Then, each listener would be assigned to one of the six groups. Also, in order to prevent the participants from having pressure taking the exams, an announcement was made beforehand indicating that they were testing the systems, not being tested.

The intelligibility task was taken before the comprehension task. It was arranged this way due to more effort being required in the comprehension task than in the intelligibility test, where participants needed to answer questions rather than simply type the words they heard. Therefore, it would be better not to depress the listeners’ patience and willingness in the first task. The listeners were informed in advance that the sentences in the intelligibility task might not be meaningful to them and were requested to try to listen as few times as possible. For the comprehension task, listeners were only allowed to listen to each news article once before answering questions without taking notes. Also, two extra subjective questions followed each news article, asking about the participants’ confidence in completing the questions and their feelings about the speech quality, scaled from 1 (very low) to 5 (extremely high). Finally, a questionnaire was given after completing the two tasks.

The intelligibility task of this experiment took around 15 to 20 minutes, while the comprehension test was about 25 to 30 minutes. Delogu *et al.* (1998) pointed out that many researchers had found that participants would fail to maintain their attention after 20 to 35 minutes of doing the task. Due to this finding, participants were asked to take a 5-minute break between the two tasks.

4. Results[3]

4.1 Intelligibility Task

Most of the participants specified that they only listened to each sentence once, and typed what they heard. For assessing SUS sentences, the measurement was based on calculating word error rates (WER) occurring in every sentence. Typographical errors and homophones were allowed.

Table 3. Significant differences in intelligibility of the three speech systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems.[4]

	Natural	HTS-2008	Multisyn
Natural		■	■
HTS-2008	■		■
Multisyn	■	■	

In Pairwise Comparisons, as presented in Table 3, there are significant differences found between natural speech and HTS-2008 ($p = 0.005$), natural speech and Multisyn ($p < 0.001$), and HTS-2008 and Multisyn systems ($p < 0.001$). To further verify the main effects in Pairwise Comparisons, the results in the Tests of Within-Subjects Contrasts show that there are significant effects when natural speech is compared to HTS-2008, $F(1, 249) = 10.135$[5], $p = 0.002$; and when HTS-2008 is compared to the Multisyn system, $F(1, 249) = 26.685$, $p < 0.001$. Therefore, it can be concluded that natural speech has significantly lower WER (M = 4.2%, SD = 10%) than HTS-2008 (M = 6.7%, SD = 11.4%) and HTS-2008 is even better than the

[3] Since a detailed table of the scored collected from intelligibility and comprehension tests might be too much to confuse the results description in this section, I simply provide tables with further analysed statistical results here.

[4] There are a total of 4 figures in this paper describing the statistical significant differences between speech synthesizers based on experimental results. Combined with the results presented in the figures, the statistical mean value (M) and standard deviation value (SD) are also given to further investigate their performance.

[5] In this section, you will find that a lot of statistical values are provided. In the presented form, F(a, b) = c, F is the symbol of degree of freedom (df); a is the df value in the whole tested data set; b is the df value of the deviation between the data set; and c is the output value of df. When the distance between a and b values gets larger, the greater the c value represents a stronger significant difference existed within the data set, usually followed with a p value as a reference.

Multisyn system (M = 14.3%, SD = 21.6%).

4.2 Comprehension Task

4.2.1 The Results from News Articles

A 3-point scale (0, 1, 2) was applied in the experiment to score answers in the open-ended questions. If the responses to the comprehension questions were judged to be incorrect, 0 points were earned; if part of the answers were correct or the answers were too general and nonspecific, yet not wrong, 1 point would be given; and 2 points were given to the responses with fully correct and specific answers. A total of 10 points for 5 open-ended questions per news article was possible. The examples of assessing the responses from open-ended questions are provided in Table 4.

Table 4. Examples of assessing the responses from open-ended questions

Open-ended Question	Correct Answer	Listener Response	Score
What are the two new news channels that have been launched by Russia?	English and Arabic	English, Arabic	2
		English and Polish	1
		Arabic	1
		Don't know	0

The 3-point scoring system was adopted from Hustad (2008). The reason for not taking a 2-point binomial scoring scale was because, in real life comprehension, it is not always an all correct or wrong situation, as described by Hustad & Beukelman (2002). Nevertheless, since the multiple-choice questions only had one correct answer, the binomial scoring system was introduced to assess the responses. If the participants chose the correct choice, then 2 points would be earned; if they chose the wrong answer, 0 points would be awarded. There would be a sum of 10 points for 5 multiple-choice questions per news article. Therefore, the total score in each article was 20 points.

There is no significance found in the three speech systems and none in the interaction between systems and the question types. Nevertheless, there is a significant effect occurring in the question types, $F(1, 24) = 29.004$, $p < 0.001$. Therefore, the performance in open-ended questions was considerably worse (mean of error rate = 39.1%) than multiple-choice questions (mean of error rate = 28%). Furthermore, there is no significance found in the interactions between the systems and multiple-choice questions. Nevertheless, there is a main effect observed in the interaction between systems and open-ended questions, $F(1.569, 37.649) = 7.348$, $p = 0.004$. Due to this fact, it can be interpreted that the results from open-ended questions shows the differences of the three systems.

Table 5. Significant differences in open-ended questions of the three systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems

	Natural	HTS-2008	Multisyn
Natural			
HTS-2008			■
Multisyn		■	

As presented in Table 5, in the open-ended questions, a significant effect is revealed only when the comparison is between HTS-2008 and the Multisyn system, $F(1, 24) = 25.939$, $p < 0.001$. Also, HTS-2008 performs much better (mean of error rate = 29.2%) than the Multisyn system (mean of error rate = 49.8%) in answering the open-ended questions correctly.

4.2.2 A 5-point Scale for Subjective Judgments

Two individual subjective questions were given at the end of each news articles: the confidence in making right responses to the questions (Confidence) and the feeling about the displayed speech quality (Quality). Both the Confidence and Quality tests used a 5-point scale (from 1 to 5) in assessing the subjective questions. Higher points represented listeners with higher satisfaction, as shown below in Table 6.

Table 6. The 5-point scale measurement for the Confidence and Quality subjective tests

1 = Very low.
2 = Low.
3 = Average
4 = High.
5 = Extremely high.

Accordingly, there are main effects found in the systems, $F(1.45, 34.806) = 25.365$, $p < 0.001$, and in the interaction between systems and the subjective tests, $F(2, 48) = 58.808$, $p < 0.001$. Nevertheless, there is no significant main effect observed in the subjective tests.

Table 7. Significant differences in the overall subjective test performance of the three systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems

	Natural	HTS-2008	Multisyn
Natural		■	■
HTS-2008	■		
Multisyn	■		

In Table 7, highly significant effects occurred when HTS-2008 was compared to natural speech, $F(1, 24) = 24.758$, $p < 0.001$; and when the Multisyn system was compared to natural speech, $F(1, 24) = 37.536$, $p < 0.001$. While Quality compares to Confidence, two main effects are discovered in the interactions when HTS-2008 is compared to natural speech, $F(1, 24) = 89.161$, $p < 0.001$, and when Multisyn is compared with natural speech, $F(1, 24) = 73.059$, $p < 0.001$. Therefore, it can be concluded that HTS-2008 is evaluated lower (M = 52.4%) than natural speech (M = 71.6%) in the subjective tests and lower points are given to Multisyn (M = 52.2%) than to natural speech. Therefore, it is known that natural speech has better results from the subjective tests than the HTS-2008 and Multisyn systems.

The Confidence test does not show any significant effect on the systems. This result indicates that listeners have equal confidence in natural speech, HTS-2008, and the Multisyn system in answering the questions of each news article. As for the results from the Quality test, there is a significance discovered in the systems, $F(1.462, 35.085) = 61.249$, $p < 0.001$.

Table 8. Significant differences in Quality test of the three systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems

	Natural	HTS-2008	Multisyn
Natural		■	■
HTS-2008	■		
Multisyn	■		

In the Quality test, natural speech has an extremely high score in speech quality identification (M = 82.8%), compared to the HTS-2008 (M = 48.8%) and Multisyn (M = 49.6%) systems. The results in Table 8 show no significance when HTS-2008 is compared to the Multisyn system. As a result, in the subjective judgment of speech quality, natural speech is scored significantly higher than HTS-2008 and Multisyn. On the other hand, the HTS-2008 and Multisyn systems are rated with nearly the same synthetic speech quality by listeners.

The results also demonstrate that, although all of the news articles were generated by concatenating the individual sentences together, natural speech still has better speech prosody than the other two speech synthesizers. This is because the recorder of natural speech knows the context and will be able to articulate the sentences with adequate prosody contours while recording. Nevertheless, the news articles produced by HTS-2008 and Multisyn systems were simply synthesized into individual sentences, without considering the context prosody factor. As stated by Sanderman and Collier (1997), listeners preferred the speech systems with higher prosody quality. Therefore, the listeners graded natural speech with the highest score, compared to HTS-2008 and Multisyn.

5. Discussion

5.1 The Discussion in the Experiment Results

5.1.1 The Relationships between Intelligibility and Comprehension

In the intelligibility task, the results prove there are significant differences between the three systems. In the intelligibility performance, natural speech is better than HTS-2008, while HTS-2008 has greater performance than the Multisyn system. According to the initial assumption in this paper, assuming systems with higher achievement in the intelligibility task would also preserve better accomplishment in the comprehension task, we can estimate that the three systems might have the same rankings in the comprehension task as presented in the intelligibility task.

Nevertheless, in the overall comprehension task performance, no significant effects are noticed within the three systems, which signifies that natural speech, HTS-2008, and Multisyn all have a similar understandability quality for listeners. The outcomes in the comprehension task are against the results in the intelligibility task and violate the assumptions of this paper. Although it seems that the comprehension task in this study has also failed to distinguish various speech systems, this is mainly because the three systems have reached the threshold of producing comprehensible speech quality. This can be demonstrated from the results in the Confidence test.

In the Confidence test, there was no significant difference observed in the three systems, which meant that listeners have equivalent confidence in completing the comprehension task produced by the systems. This implies that the three systems have given identical comprehension quality to the listeners. In addition, the techniques required for evaluating intelligibility and comprehension are different.

In the comprehension task, the main intention is to understand and comprehend the global meanings offered in each news article, whereas the intelligibility task is not evaluated by focusing on the meanings of the words but on paying attention to every single word that can be heard. During the process of comprehension, even if some of the words are not clear to the listeners, the comprehension process will not be interrupted. Listeners can still acquire general meanings from the context of the articles. Benoît *et al*. (1996) found that, with sufficient linguistic cues, it would be easy for listeners to derive learning effects and process the effects while comprehending. Thus, with sufficient cues provided from the three systems, no significant differences could be found within the three systems in the comprehension task. In other words, although natural speech, HTS-2008, and Multisyn are significantly different from each other in intelligibility, they all obtain enough intelligibility quality for listeners to learn the linguistic cues and comprehend the texts. In addition, the WER of 14.3% in the Multisyn system can be taken as an intelligibility threshold reference for achieving high comprehensibility in speech synthesizers.

5.1.2 The Influence of Different Question Types used in the Comprehension Task

In the comprehension task, different question types used in the experiment will bring a significant effect to the systems' measurement. In this experiment, only the open-ended questions have a significant effect on the systems. This may be affected by the design purpose of each type of question.

For the multiple-choice questions, they are assigned to be inferential questions, which need to be processed and comprehended before answering. Thus, this procedure is very much the same as in the real comprehension process and shows that natural speech, HTS-2008, and Multisyn have the same comprehensibility. Nevertheless, the open-ended questions are designed to be factual questions, which makes the process of answering the questions similar to the way of completing the intelligibility task. Both the open-ended questions and intelligibility task involve listening to the speech first and focusing on the key words they can capture or understand.

The only difference between them is that the load of memory will be larger in open-ended questions than in the intelligibility task. As seen in the results of open-ended questions, the consequences diverge a little from the results in the intelligibility task. In the open-ended questions, the performance in natural speech is identical to HTS-2008, but is better than the Multisyn system. The intelligibility task, however, shows that natural speech is better than HTS-2008 and Multisyn. In addition, even the overall subjective tests and quality test show that natural speech has better achievement than HTS-2008. This may be due to there not being enough participants included in the experiment (only 25 participants in this study).

Therefore, it is assumed that, if the number of participants increases, the significant effect between natural speech and HTS-2008 in open-ended questions might occur. Apart from the intelligibility and comprehension task, the overall subjective tests and quality test are both consistent with the results specifying that the performances in HTS-2008 and the Multisyn system are the same. In general, the entire experiment in the present study has found that natural speech has greater impact and performance than HTS-2008 and Multisyn.

5.2 Listeners' Feedback and some Suggestions for Future Studies

5.2.1 Listeners' Feedback

Most of the participants found the intelligibility task interesting. Since the materials were all semantically unpredictable sentences, there could be many unexpectedly funny sentences. Still, some of the participants specified that there were a few words they had seldom heard or seen in their life, which might lead to some misspelling or making up the spelling. This problem was solved in this study by allowing typographical errors and homophones while calculating the WER in the intelligibility task. They also indicated that, in sentences with poor speech quality, it would be difficult for them to recognize the words as real words.

Most of the participants reported that the second part of the experiment (comprehension task) was harder than the first part (intelligibility task). They stated that the display duration of news articles was a bit long for them to remember all of the information. Besides, the listeners stated that, if the article were presented with low speech quality, it would be harder for them to concentrate and follow up. In addition, they tended to focus more on the topics they were interested in and answered these questions correctly more often. Some participants suggested that there should be an option of "do not know the answer" added to the multiple-choice questions to prevent them from guessing the answers.

Although there were comments coming from the participants, they still responded that the whole experiment was interesting, and they had a lot of fun during the process.

5.2.2 Suggestions and Modifications for Future Works

According to the feedback received from the participants, some things can be modified in the comprehension design to make the task better. First, since most of the participants replied that the durations of news articles were a little bit too long, a pilot test for measuring the participants' feelings of duration needs to be applied before carrying out the main experiment. In addition, with long news articles as experimental materials, there may be too much redundant information embedded, which may interfere with the comprehension testing. Furthermore, since each news article had different topics, there is no guarantee that the degree of text complexity and familiarity would remain the same between articles. The word "text complexity" used here means the degree of comprehension effort that needs to be devoted to listening to the article.

Due to the limitation of time, there were not enough listeners participating in each pilot test. In order to remove the individual problems and increase the objectivity of the results of the test, it will be better to have at least 10 people included in the pilot test.

6. Conclusion

From the results in the intelligibility task, we find that the performance in natural speech is better than HTS-2008, and HTS-2008 is proven better than the Multisyn system. Nevertheless, the results in the comprehension task show that the natural speech, HTS-2008, and Multisyn systems display equal quality for listeners to comprehend. The explanation has been given in Section 5.1.1, discussing the issue that all three systems obtain enough intelligibility quality to be used in comprehending the news passages. Although the outcomes in the intelligibility task show that there are significant differences in the three systems, their intelligibility has reached the comprehension threshold to produce understandable high quality speech. In spite of the objective results in the comprehension task, in the overall subjective tests and the Quality test, both of them show that listeners consider natural speech to be the best system of all, compared to the two speech synthesizers (HTS-2008 and Multisyn). Besides, the listeners feel that there is no difference between HTS-2008 and the Multisyn system.

For the design of the comprehension task, there is still one thing that needs to be mentioned. That is the comprehension task designed in this experiment could not directly evaluate the comprehension process, as stated by Pisoni *et al*. (1985). Since the questions are derived after listening, this kind of measurement is a post-perceptual comprehension. Therefore, the comprehension strategies involved in this study are all evaluating the products of comprehension, rather than the process itself.

In general, from the results of this experiment, the HTS-2008 speech synthesizer is preferable and more usable than the Multisyn system. Although the two systems have the same performance in comprehension, HTS-2008 is significantly better than the Multisyn system in intelligibility.

References

Benoît, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication,* 18, 381-392.

Black, A. W., & Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of the Eurospeech 1997.*

Black, A. W., & Tokuda, K. (2005). The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common dataset. In *Proceedings of Interspeech 2005,* Lisbon, Portugal.

Clark, R. A. J., Richmond, K., & King, S. (2004). Festival 2 - build your own general purpose unit selection speech synthesiser. In *Proceedings of 5th ISCA Speech Synthesis Workshop,* Pittsburgh, USA.

Clark, R. A. J., Richmond, K., & King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication,* 49, 317-330.

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication,* 24, 153-168.

Goldstein, M. (1995). Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication,* 16(3), 225-244.

Heiga, Z., & Tomoki, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proceedings of Interspeech 2005,* Lisbon, Portugal.

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the ICASSP 1996,* Atlanta, USA.

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research,* 51, 562-573.

Hustad, K. C., & Beukelman, D. R. (2002). Listener coomprehension of severely dysarthric speech: Effects of linguistic cues and stimulus cohesion. *Journal of Speech, Language, and Hearing Research,* 45, 545-558.

Karaiskos, V., King, S., Clark, R. A. J., & Mayo, C. (2008). The Blizzard Challenge 2008. In *Proceedings of the Blizzard Challenge 2008 workshop*, Brisbane, Australia.

Lai, J., Wood, D., & Considine, M. (2000). *The effect of task conditions on the comprehensibility of synthetic speech*. Paper presented at the CHI Letters.

Luce, P. A. (1981). *Comprehension of fluent synthetic speech produced by rule* (Research on Speech Perception Progress Report No. 7). Bloomington, IN 47405: Indiana University.

Miller, G. A., & Isard, S. D. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior,* 2, 217-228.

Nusbaum, H., Francis, A., & Henly, A. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology,* 1, 7-19.

Nusbaum, H. C., Schwab, E. C., & Pisoni, D. B. (1984). *Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability* (Research on Speech Perception Progress Report No. 10). Bloomington, IN47405: Speech Research Laboratory, Indiana University.

Paris, C. R., Thomas, M. H., Gilson, R. D., & Kincaid, J. P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors,* 42, 421-431.

Pisoni, D. B., Nusbaum, H. C., & Greene, B. G. (1985). Perception of synthetic speech generated by rule. In *Proceedings of the IEEE.*

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Schwab, E. C. (1983, Apr 1983). Perceptual evaluation of synthetic speech: Some considerations of the user/System interface. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83*.

Pols, L. C. W., van Santen, J. P. H., Abe, M., Kahn, D., & Keller, E. (1998). The use of large text corpora for evaluation text-to-speech systems. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Ralston, J. V., Pisoni, D. B., & Mullennix, J. W. (1989). *Comprehension of synthetic speech produced by rule* (Research on Speech Perception Progress Report No. 15). Bloomington, IN47405: Speech Research Laboratory, Indiana University.

Salasoo, A. (1982). *Cognitive Processes and comprehension measures in silent and oral reading* (Research on Speech Perception Progress Report No. 8). Bloomingtion, IN 47405: Speech Research Laboratory, Indiana University.

Sanderman, A. A., & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech,* 40, 391-409.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language,* 19, 129-146.

Sydeserff, H. A., Caley, R. J., Isard, S. D., Jack, M. A., & Monaghan, A. I. C. (1992). Evaluation of speech synthesis techniques in a comprehension task. *Speech Communication,* 11, 189-194.

Terken, J., & Lemeer, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics,* 16, 453-457.

Yamagishi J., *et al.* (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge 2008,* Brisbane, Australia.

Yorkston, K., Strand, E., & Kennedy, M. (1996). Comoprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology,* 5(1), 55-66.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultanious modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of Eurospeech 1999*.

Yu, S.-Z., & Kobayashi, T. (2003). An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Signal Processing Letters,* 10, 11-14.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., et al. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of ISCA SSW6,* Bonn, Germany.

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member：	NT$ 1,000 (US$50.-)
group membership：	NT$20,000 (US$1,000.-)

life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502　Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw　Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： ____________________

Name： ____ Date of Birth： __________________

Country of Residence： ________________________ Province/State： ________________________

Passport No.： __________________________________ Sex: ____________________________

Education(highest degree obtained)： ____

Work Experience： ____

____

Present Occupation： ____

Address： ____

____

Email Add： ____

Tel. No： __________________________________ Fax No： ________________________________

Membership Category：□ Regular Member □ Life Member

Date： _____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member	：	US$ 50.- （NT$ 1,000）
Life Member	：	US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號		（由本會填寫）	
姓名		性別		出生日期	年 月 日
				身分證號碼	
現職		學歷			
通訊地址	□□□				
戶籍地址	□□□				
電話		E-Mail			
申請人:　　　　　（簽章） 中華民國　　年　　月　　日					

審查結果:

1. 年費：

   終身會員： 10,000.-
   個人會員： 1,000.-
   學生會員： 500.-（限國內學生）
   團體會員： 20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

PAYMENT FORM

Name :____(Please print) Date: ___________________

Please debit my credit card as follows: US$__________________________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:_________________________

Card No.: __________- __________ - __________ - __________ Exp. Date:________________

3-digit code: ______________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE : ____________________________________

Tel.: ___________________________________ E-mail: ___________________________________

Add: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (CLCLP)

Quantity Wanted: ______

US$ _____ ❑ Publications:___________________________________

US$ _____ ❑ Text Corpora: _________________________________

US$ _____ ❑ Speech Corpora:________________________________

US$ _____ ❑ Others: _______________________________________

US$ __________ ❑Life Member Fee ❑ New Member ❑Renew

US$ __________ = Total

Fax : 886-2-2788-1638 or Mail this form to :

ACLCLP
% Institute of Information Science, Academia Sinica
R502, 128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會

信用卡付款單

姓名: ________________________________(請以正楷書寫)　日期:： ______________

卡別：❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD　發卡銀行：________________

卡號:_________-_________-_________- ________　有效日期：_________________

卡片後三碼：_____________________（卡片背面簽名欄上數字後三碼）

持卡人簽名：　__________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：_______________________　E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________❑ 中文計算語言學期刊(IJCLCLP)

NT$__________❑ 中研院詞庫小組技術報告

NT$__________❑ 中文（新聞）語料庫

NT$__________❑ 平衡語料庫

NT$__________❑ 中文詞庫八萬目

NT$__________❑ 中文句結構樹資料庫

NT$__________❑ 平衡語料庫詞集及詞頻統計

NT$__________❑ 中英雙語詞網

NT$__________❑ 中英雙語知識庫

NT$__________❑ 語音資料庫______________________________

NT$__________❑ 會員年費　❑續會　❑新會員　❑終身會員

NT$__________❑ 其他: __________________________________

NT$_______________=　合計

填妥後請傳真至 02-27881638 或郵寄至:
115台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

			AIR	AIR		
		Surface	(US&EURP)	(ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本)　ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02　V-N 複合名詞討論篇 & 92-03　V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05　中文詞類分析	10	15	13	_____	_____
7.	no.93-06　現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02　古漢語字頻表	11	16	14	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02　論語詞頻表	9	14	12	_____	_____
16.	no.98-01　詞頻詞典	18	30	26	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03　自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01　現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT：□ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or "中華民國計算語言學學會"
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會
相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01　「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期） 年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：______________________　收據抬頭：__________________________

地　　址：____

電　　話：______________________　E-mail: __________________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.
1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.
2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.
3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.
4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).
5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript
6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.
7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

> Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

> Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).
Please visit the following websites for details.
(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)
(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

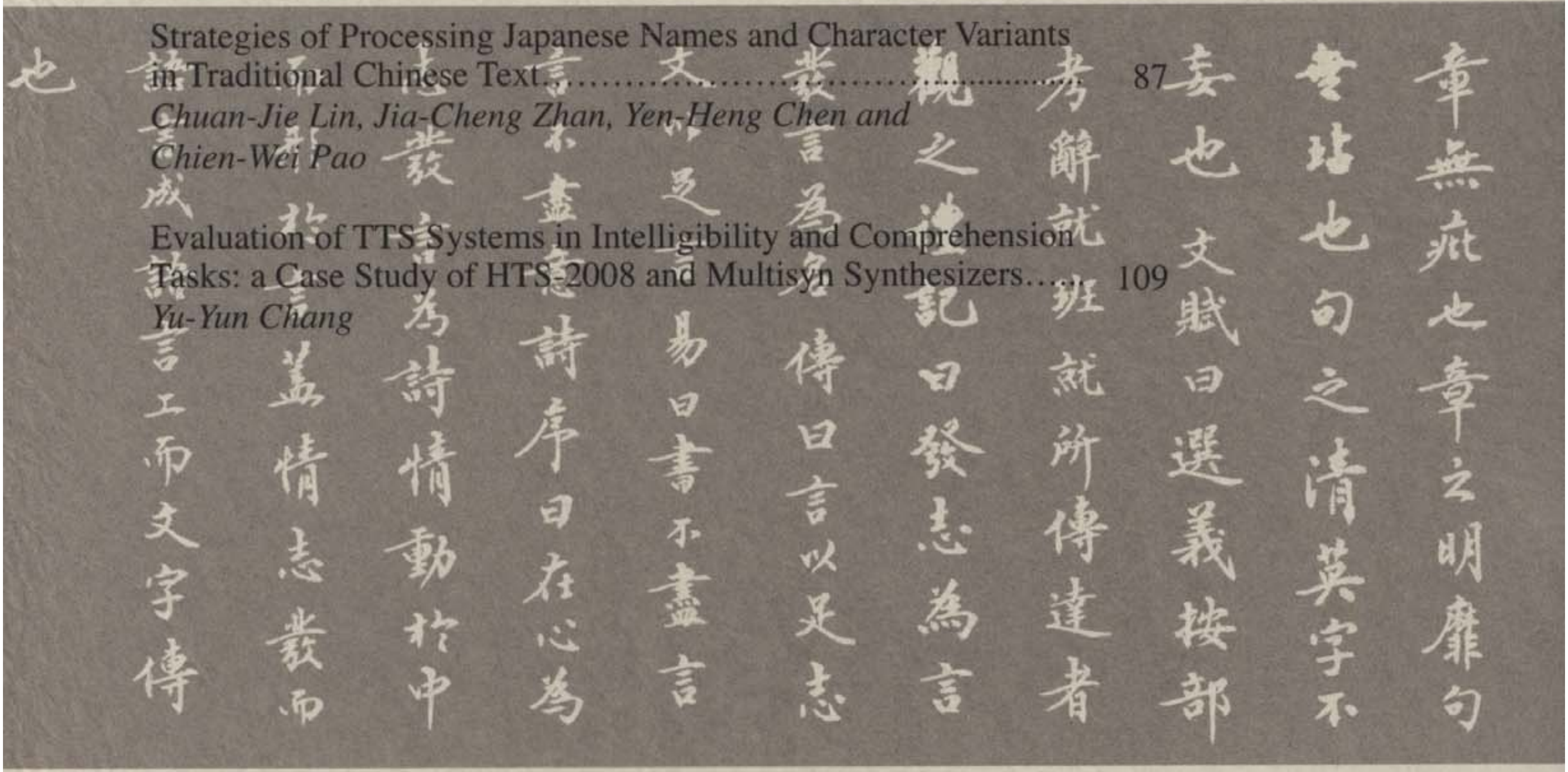
Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.17 no.3 September 2012

Contents

Papers



ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing