Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 1-18 


# Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription

**Berlin Chen*, Jen-Wei Kuo* and Wen-Hung Tsai***

**Abstract**

This article investigates the use of several lightly supervised and data-driven approaches to Mandarin broadcast news transcription. With the special structural properties of the Chinese language taken into consideration, a fast acoustic look-ahead technique for estimating the unexplored part of a speech utterance is integrated into lexical tree search to improve search efficiency. This technique is used in conjunction with the conventional language model look-ahead technique. Then, a verification-based method for automatic acoustic training data acquisition is proposed to make use of large amounts of untranscribed speech data. Finally, two alternative strategies for language model adaptation are studied with the goal of achieving accurate language model estimation. With the above approaches, the overall system was found in experiments to yield an 11.88% character error rate when applied to Mandarin broadcast news collected in Taiwan.

**Keywords:** acoustic look-ahead, lightly supervised acoustic model training, language model adaptation, Mandarin broadcast news

## 1. Introduction

With the continuing growth of the amount of multimedia information accessible over the Internet, large volumes of real-world speech information, such as that in broadcast radio and television programs, digital libraries, and so on, are now being accumulated and made available to the public. Substantial efforts and very encouraging results for broadcast news transcription, retrieval, and summarization have been reported [Woodland 2002; Gauvain *et al.* 2002; Beyerlein *et al.* 2002; Chen *et al.* 2002; Chang *et al.* 2002; Meng *et al.* 2004; Furui *et al.* 2004]. However, in order to obtain better recognition performance, most of the transcription systems require not only large amounts of manually transcribed speech materials for acoustic training in the data preparation phase, but also much time and memory in the recognition

* Graduate Institute of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan, Republic of China
E-mail: {berlin, rogerkuo, louis}@csie.ntnu.edu.tw

phase. Moreover, because the subject domains and lexical regularities of the linguistic contents of news articles are very diverse and often change with time, it is extremely difficult to build well-estimated language models for speech recognition. Hence, in the recent past, several attempts have been made to investigate the possibility of achieving automatic acquisition of speech or language training data for system refinement or for rapid prototyping of a new recognition system to new domains, and very encouraging results have been obtained [Kemp and Waibel 1999; Wessel and Ney 2001; Macherey and Ney 2002; Bacchiani 2003]. On the other hand, quite a few studies have also explored ways to improve recognition efficiency, and many good approaches have been proposed [Schuster 2000; Aubert 2002; Evermann and Woodland 2003]. In this paper, several lightly supervised and data-driven approaches to Mandarin broadcast news transcription are presented. First, considering the special structural properties of the Chinese language, a fast acoustic look-ahead technique that employs syllable-level heuristics is integrated into lexical tree search to improve search efficiency. It is used in conjunction with the conventional language model look-ahead technique [Ortmanns and Ney 2000]. Then, a verification-based method for automatic acoustic training data acquisition is proposed to make use of large speech corpora. Finally, two alternative strategies for language model adaptation are studied with the goal of achieving accurate language model estimation.

The remainder of this paper is organized as follows. In section 2, we review the major constituents of our broadcast news system and introduce the experimental speech and language data used in this research. The acoustic look-ahead technique using syllable-level heuristics is presented in section 3, while the lightly supervised acoustic model training and language model adaptation approaches are described in sections 4 and 5, respectively. Then, the results of a series of speech recognition experiments are discussed in section 6. Finally, conclusions are drawn in section 7.

## 2. The NTNU Broadcast News System

The major constituent parts of the broadcast news system developed at National Taiwan Normal University (NTNU) as well as the speech and language data used in this paper will be described in this section [Chen *et al.* 2004]. Figure 1 depicts the overall framework of the broadcast news system.

### 2.1 Front-End Processing

Front-end processing is conducted with two feature extraction approaches: the conventional MFCC-based (Mel-frequency Cepstral Coefficients) [Davis and Mermelstein 1980] and the data-driven LDA-based (Linear Discriminant Analysis) [Duda and Hart 1973] approaches. In the case of the MFCC-based approach, 13-dimensional cepstral coefficients derived from 18

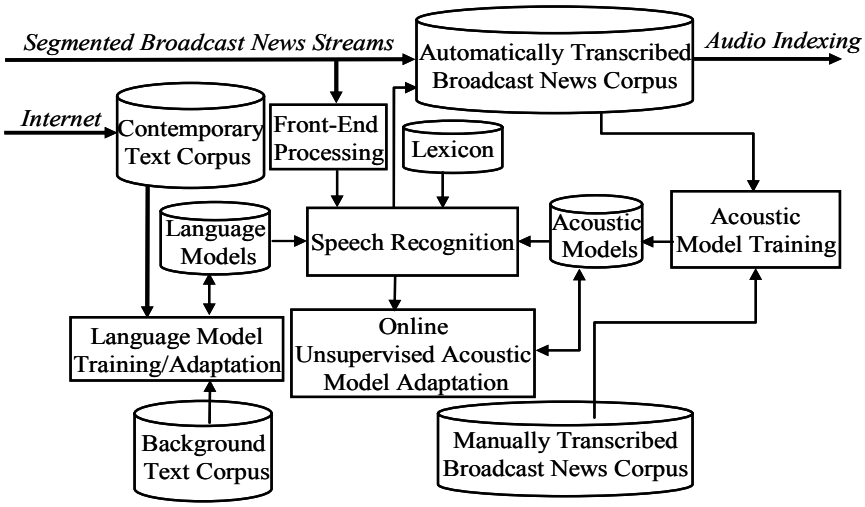



***Figure 1. The overall framework of the NTNU broadcast news system.***

filter bank outputs are incorporated along with their first- and second-order time derivatives. As for the LDA-based approach, the states of each HMM (Hidden Markov Model) are taken as the units for class assignment. Either the outputs of filter banks or the cepstral coefficients are chosen as the basic vectors. The basic vectors from every nine successive speech frames are spliced together to form supervectors for constructing the LDA transformation matrix, which is then used to project the supervectors to a lower feature space. The dimension of the resultant vectors is set to 39, which is just the same as that used in the MFCC-based approach. Finally, in both the MFCC- and LDA-based feature extraction approaches, utterance-based cepstral mean subtraction and variance normalization are applied.

## 2.2 Speech Corpus and Acoustic Modeling

The speech data set consists of about 112 hours of FM radio broadcast news, which was collected from several radio stations located in Taipei during 1998-2002 using a wizard FM radio connected to a PC and digitized at a sampling rate of 16 kHz with 16-bit resolution [Chen *et al.* 2002]. All the speech materials were manually segmented into separate stories, each of which is a news abstract spoken by one anchor speaker. Some of these stories contain background noise and music. For 7.7 hours of speech data, we have corresponding orthographic transcripts. About 4.0 hours of this data collected from 1998 to 1999 was used to bootstrap the acoustic training, and the other 3.7 hours of data collected in September 2002 was used for testing. The remaining 104.3 hours of untranscribed speech data was reserved for lightly supervised acoustic model training, which will be described in more detail in section 4.

The acoustic models chosen for speech recognition were 112 right-context-dependent INITIAL's and 38 context-independent FINAL's. They were selected based on consideration of the phonetic structure of Mandarin syllables [Chen *et al.* 2002]. Here, INITIAL means the initial consonant of a syllable and FINAL is the vowel (or diphthong) part but also includes an optional medial or nasal ending. Each INITIAL is represented by an HMM with 3 states, while each FINAL is represented with 4 states. The Gaussian mixture number per state ranges from 2 to 128, depending on the quantity of training data. In all the experiments, gender-independent models were used.

## 2.3 Lexicon, Text Corpus and Language Modeling

In the Chinese language, each character (at least 7,000 characters are commonly used) is pronounced as a monosyllable and is a morpheme with its own meaning. New words are very easily generated by combining a few characters but nevertheless are tokenized into several single-character words or words with fewer characters when the text corpus is processed for language model training. This definitely makes the out-of-vocabulary problem especially serious in the case of Mandarin broadcast news transcription. In order to alleviate the degradation of speech recognition accuracy caused by the out-of-vocabulary problem, compound words must be carefully selected and added to the lexicon according to their statistical properties in the corpus. Hence, we explored the use of the geometrical average of the forward and backward bigrams of any word pair $(w_i,w_j)$ occurring in the corpus for compound word selection [Saon and Padmanabhan 2001; Wang *et al.* 2002]:

$$FB(w_i, w_j) = \sqrt{P_f(w_j \mid w_i) P_b(w_i \mid w_j)}, \tag{1}$$

where

$$P_f(w_j \mid w_i) = \frac{P(w_{t+1} = w_j, w_t = w_i)}{P(w_t = w_i)} \qquad \text{and} \tag{2}$$

$$P_b(w_i \mid w_j) = \frac{P(w_{t+1} = w_j, w_t = w_i)}{P(w_{t+1} = w_j)}. \tag{3}$$

We started with a lexicon composed of 67K words and iteratively used the above measures with varying thresholds to find all possible word pairs which could be merged together. Eventually, a set of about 5K compound words was added to the lexicon to form a new lexicon of 72K words. The *n*-gram language modeling approach was adopted in the study; thus, the background language models consisted of word-based trigram and bigram models, which were estimated using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2000 and 2001 (the Chinese Gigaword Corpus released by LDC [LDC 2003]). On the other hand, a corpus consisting of 50 million Chinese characters in newswire texts collected from the Internet from August to October 2002 [Chang *et al.* 2003]

was used as a contemporary corpus for language model adaptation. The language models were trained with Kneser-Ney backoff smoothing [Kneser and Ney 1995] using the SRI Language Modeling Toolkit (SRILM) [Stolcke 2000].

## 2.4 Speech Recognition

Our baseline recognizer was implemented with left-to-right frame-synchronous tree search as well as lexical prefix tree organization of the lexicon [Aubert 2002; Beyerlein *et al.* 2002; Woodland 2002]. Each tree arc (or phonetic arc) in the lexical tree corresponded to the HMM for an INITIAL or FINAL in Mandarin Chinese, and each tree leaf denoted a word boundary for words sharing the same pronunciation. At each speech frame, the so-called word-conditioned method was used to group the path hypotheses that shared the same history of predecessor words (or more precisely, the same search history of $n$-1 predecessor words for $n$-gram language modeling) into identical copies of the lexical tree, and they were then expanded and recombined according to the tree structure until a possible next word ending was reached. At word boundaries, the path hypotheses among the tree copies that had equivalent search histories (the same last $n$-1 words) were recombined and then propagated into the existing tree copies or used to start new ones if none existed. Note that these tree copies were built according to a conceptual view. During the search process, only one lexical tree structure was built for reference purposes, and all path hypotheses were stored in a list structure instead. These path hypotheses were accessed by means of four-dimensional coordinates, each of which represented the history of $n$-1 predecessor words, the tree arc in the lexical tree, the HMM state, and the speech frame, respectively. At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their corresponding language model look-ahead scores, was used to select the most promising path hypotheses. Language model look-ahead was adopted because the search structure was implemented with a lexical prefix tree and the current word identity of a particular path hypothesis could not be determined until it reached a tree leaf. In addition, language model look-ahead has the merit of early application of language model constraints, which can help guide the search process. In this research, unigram language model look-ahead was adopted. The unigram language model look-ahead score for a tree arc was defined as the maximum unigram probability over all the words that could be reached via this specific arc, which could be easily calculated and stored beforehand. Therefore, for a path hypothesis ending at speech frame $t$, which had a search history $h$ and stayed at tree arc $k$ and HMM state $q$, its corresponding decoding score, $D(t, h, arc_k, s_q)$, could be modified via the following equation:

$$\log \hat{D}(t, h, arc_k, s_q) = m_1 \cdot \log D(t, h, arc_k, s_q) + m_2 \cdot \log L_{LM}(arc_k), \tag{4}$$

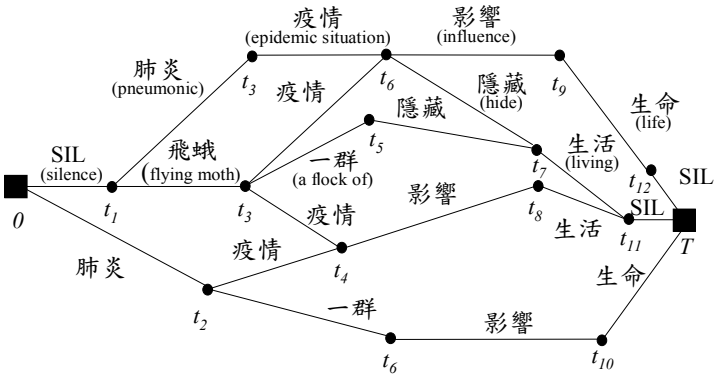



***Figure 2. An illustration of a word graph, in which each arc, together with its corresponding start and end speech frames, represents a candidate word hypothesis.***

where $L_{LM}(arc_k)$ is the unigram language model look-ahead score for tree arc $k$ (notice that the HMM states within the same tree arc share the same language model look-ahead score), and $\hat{D}(t,h,arc_k,s_q)$ is the modified decoding score. $m_1$ and $m_2$ are the weighting parameters, which were set to 1 and 8, respectively, in this research. During beam pruning, we first computed the modified decoding score of the best path hypothesis at each speech frame $t$:

$$\log \hat{D}_{\max}(t) = \max_{h,k,q} \log \hat{D}(t,h,arc_k,s_q). \tag{5}$$

Then, an unpromising path hypothesis was pruned if the logarithm of its modified decoding score, $\log\hat{D}(t,h,arc_k,s_q)$, was lower than a predefined threshold:

$$\log \hat{D}(t,h,arc_k,s_q) < \log \hat{D}_{\max}(t) - \log f_{Thr}, \tag{6}$$

where $f_{Thr}$ is an empirically set pruning factor. Moreover, if the word hypotheses ending at each speech frame had scores that were higher than the predefined threshold, their associated decoding information, such as the word start and end speech frames, the identities of current and predecessor words, and the acoustic score, were kept in order to build a word graph for further language model rescoring [Ortmanns *et al*. 1997]. Once the word graph had been built, as illustrated in Figure 2, forward-backward search with a more sophisticated language model was conducted to generate the most likely word sequence. In this study, the bigram language model was used in the tree search procedure, while the trigram language model was used in the word graph rescoring procedure.

## 3. Acoustic Look-Ahead Using Syllable-level Heuristics

In a baseline recognizer, language model look-ahead and beam pruning techniques can be incorporated together to help retain the most promising path hypotheses for further expansion.

However, the crucial problem with such an approach is that it does not consider the potential likelihood of the unexplored portion of a speech utterance when beam pruning is applied. Thus, many unpromising path hypotheses and ambiguities will unavoidably be included during the search process. Therefore, the search efficiency may be degraded, since a large number of path hypotheses will have to be examined at each speech frame. On the other hand, the Chinese language is well known for its monosyllabic structure, in which each Chinese word is composed of one or more syllables (or characters); thus, syllables are the very important constituent units of Chinese words [Lee 1997; Chen *et al.* 2002; Meng *et al.* 2004]. In addition, Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio data if the tonal information is further ignored. This implies that syllable recognition will be much faster than word recognition. Thus, in this study, we utilized syllable-level heuristics to enhance search efficiency. A compact syllable lattice based on the structural information of words in the lexicon was automatically built and used to estimate the likelihood of the unexplored portion of a speech utterance. Each HMM state in the syllable lattice could be easily related to its corresponding HMM states in the lexical tree, and the relation between them was a one-to-many mapping. In the first pass, the syllable lattice was calculated in a right-to-left time-synchronous manner, and at each speech frame, the acoustic scores for the HMM states in the lattice were stored and taken as the likelihood estimation for acoustic look-ahead. In the second pass, frame-synchronous tree search was performed by incorporating the language model look-ahead scores together with the acoustic look-ahead scores for beam pruning:

$$\log \widetilde{D}\left(t,h,arc_k,s_q\right) = m_1' \cdot \log D\left(t,h,arc_k,s_q\right) + m_2' \cdot \log L_{LM}\left(arc_k\right) + m_3' \cdot \log L_{AC}\left(t,arc_k,s_q\right), \quad (7)$$

where $L_{AC}\left(t,arc_k,s_q\right)$ is the acoustic look-ahead score, and $m_1'$, $m_2'$ and $m_3'$ are the weighting parameters, which were set to 1, 8 and 1, respectively, in this research. Though speech recognition was carried out in a two-pass mode, the time spent on calculating acoustic look-ahead scores was almost negligible. The word graph rescoring procedure also could be applied after the second-pass search.

## 4. Lightly Supervised Acoustic Model Training

The purpose of acoustic modeling is to provide a method to calculate the likelihood of a speech utterance occurring given a word sequence. In principle, a word sequence can be decomposed into a sequence of phone-like (subword, or INITIAL or FINAL in Mandarin Chinese) units, each of which is represented by an HMM, and the corresponding model parameters can be efficiently estimated from a corpus of orthographically transcribed training utterances using the Expectation-Maximum (EM) algorithm [Dempster *et al.* 1977]. Accordingly, in order to obtain acceptable performance in speech recognition, large amounts of manually transcribed speech data are inevitably required, especially when porting the

system to new application domains. However, generating manually transcribed data is an expensive process in terms of both manpower and time. Based on this observation, we investigated here the lightly supervised acoustic model training approach for Mandarin broadcast news recognition. Unlike the previous approaches [Lamel *et al.* 2002; Nguyen and Xiang 2004], which aligned closed-captions with automatic transcripts and kept only portions that agreed for acoustic training, in this study, we developed a verification-based method for automatic acoustic training data acquisition. The prototype system, initially trained with only 4 hours of manually transcribed speech corpus, was used to recognize the remaining more than one hundred hours of unannotated speech corpus, as described previously in section 2.2. For each candidate word segment generated by the forward-backward search in the word graph rescoring procedure, its associated word-level posterior probability as well as subword-level acoustic verification score, or more specifically, sub-syllable-level verification score, were incorporated together. The word-level posterior probability of a specific word segment $w$ in the word graph with the start and end speech frames $t_s$ and $t_e$, respectively, can be defined as [Wessel *et al.* 2001]

$$P_{Post}(w_{t_s}^{t_e} \mid X_1^T) = \frac{\sum_{W_1^{t_s-1}} \sum_{W_{t_e+1}^T} p(W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^T, X_1^T)}{\sum_{W_1^T} p(W_1^T, X_1^T)}, \tag{8}$$

where $X_1^T$ is the speech utterance $X$ which starts at speech frame $1$ and ends at speech frame $T$, $W_1^{t_s-1}$ denotes the word sequence $W$ which starts at speech frame $1$ and ends at speech frame $t_s-1$, and $p(W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^T, X_1^T)$ denotes the joint probability of word sequence $W_1^{t_s-1} \cdot w_{t_s}^{t_e} \cdot W_{t_e+1}^T$ and speech utterance $X_1^T$. On the other hand, the subword-level acoustic verification score of word segment $w$, which starts at speech frame $t_s$ and ends at speech frame $t_e$, can be expressed as [Chen *et al.* 1998]

$$Score_{AV}(w_{t_s}^{t_e}) = \frac{1}{N_w} \sum_{i=1}^{N_w} \frac{2}{1+\exp[-\tau \cdot LLR(Sub(i)) + \eta]}, \tag{9}$$

where $N_w$ is the number of subword (INITIAL or FINAL) units involved in the word segment $w$; $\frac{2}{1+\exp[-\tau \cdot LLR(Sub(i)) + \eta]}$ is a sigmoid function which provides the acoustic verification score for the subword unit $Sub(i)$; $\tau$ and $\eta$ are used to control the slope and shift of the sigmoid function, respectively; and $LLR(Sub(i))$ is the log likelihood ratio for $Sub(i)$. In this research, $\tau$ and $\eta$ were set to 0.5 and zero, respectively. The value of $LLR(Sub(i))$ can be calculated using the following equation:

$$LLR(Sub(i)) = \log \frac{p(X_{t_1}^{t_2} \mid Sub(i))}{\max_{Sub^*} p(X_{t_1}^{t_2} \mid Sub^*)}, \tag{10}$$

where $t_1$ and $t_2$ are, respectively, the start and end speech frames of subword unit $Sub(i)$, $P\left(X_{t_1}^{t_2} \mid Sub(i)\right)$ is the likelihood that the speech segment $X_{t_1}^{t_2}$ will generated by $Sub(i)$, and $\max_{Sub^*} p(X_{t_1}^{t_2} \mid Sub^*)$ is the likelihood that $X_{t_1}^{t_2}$ will be generated by the corresponding top 1 subword unit, which acts here as the competing subword unit. From Equations (9) and (10), it is clear that the subword-level acoustic verification score for $Sub(i)$ becomes 1 if $Sub(i)$ is just the top 1 candidate and decreases to zero as $P\left(X_{t_1}^{t_2} \mid Sub(i)\right)$ becomes much smaller than $\max_{Sub^*} p(X_{t_1}^{t_2} \mid Sub^*)$. The word-level posterior probability and subword-level acoustic verification score were set within the range of 0 to 1 and can be weighted to form the word confidence measure:

$$CM(w_{t_s}^{t_e}) = c_1 \cdot P_{Post}(w_{t_s}^{t_e} \mid X_1^T) + c_2 \cdot Score_{AV}(w_{t_s}^{t_e}), \tag{11}$$

where $c_1$ and $c_2$ are weighting parameters, whose values were set here to be equal, that is, $c_1$=$c_2$=0.5. Thus, we can use the word confidence measure to locate the most probably correct words. As the word confidence thresholds were varied, different amounts of automatically transcribed data were accordingly selected and used in combination with the original 4-hour manually transcribed corpus to retrain different sets of acoustic models. The LDA transformation matrix employed in the feature extraction process needed to be reestimated, and the acoustic features were recalculated as well, according to the speech data selected for training.

## 5. Language Model Adaptation

Statistical language modeling, which aims to capture regularities in human natural language and quantify the acceptance of a given word sequence, has been a focus of active research in speech and language processing over the past two decades. The *n*-gram modeling (especially the bigram and trigram modeling) approach, which determines the probability of a word given the previous *n*-1 word history, has been widely used [Rosenfeld 2000; Goodman 2001; Bellegarda 2004]. The *n*-gram probabilities are usually computed based on either the maximum likelihood (ML) principle or the maximum entropy (ME) principle [Berger *et al.* 1996]. However, to tackle the inevitable data sparseness problems that occur when estimating the *n*-gram probabilities from a specific text corpus, a variety of smoothing or interpolation techniques have been proposed in the past several years [Chen and Goodman 1999; Chen and Rosenfeld 2000]. In addition, statistical language modeling was also introduced to information retrieval (IR) problems in the late 1990s, and research at a number of sites has confirmed that such a modeling paradigm does provide a theoretically attractive and potentially very effective probabilistic framework for building IR systems [Croft and Lafferty 2003; Liu and Croft 2005; Zhai and Lafferty 2004]. However, for complicated speech recognition tasks, such as

broadcast news transcription, it is still extremely difficult to build well-estimated language models because the subject domains and lexical characteristics of the linguistic contents of news articles are very diverse and often change with time. Various approaches have been applied to adapt language models by making use of either the contemporary corpus [Federico and Bertoldi 2001] or the recognition hypotheses cached so far [Jelinek *et al.* 1991]. Two of the most widely-used approaches to language model adaptation are count merging and model interpolation, which can be viewed as maximum *a posteriori* (MAP) language model adaptation with different parameterizations of the prior distribution and can be easily integrated into the *n*-gram language modeling framework to capture the local regularities of word usage in the new task domain. The adaptation formulae (e.g., for trigram modeling) for count merging and model interpolation can be, respectively, written as

$$\hat{P}_{Adapt-1}(w_i|w_{i-2}w_{i-1}) = \frac{\alpha \cdot C_{d,Cont}(w_{i-2}w_{i-1}w_i) + \beta \cdot C_{d,Back}(w_{i-2}w_{i-1}w_i)}{\alpha \cdot C_{Cont}(w_{i-2}w_{i-1}) + \beta \cdot C_{Back}(w_{i-2}w_{i-1})}, \tag{12}$$

and

$$\hat{P}_{Adapt-2}(w_i|w_{i-2}w_{i-1}) = \gamma \cdot P_{Cont}(w_i|w_{i-2}w_{i-1}) + (1-\gamma) \cdot P_{Back}(w_i|w_{i-2}w_{i-1}). \tag{13}$$

For the count merging formula in Equation (12), $C_{d,Cont}(w_{i-2}w_{i-1}w_i)$ and $C_{d,Back}(w_{i-2}w_{i-1}w_i)$ are, respectively, the discounted trigram counts [Chen and Goodman 1999] accumulated from the contemporary and background text corpora; $C_{Cont}(w_{i-2}w_{i-1})$ and $C_{Back}(w_{i-2}w_{i-1})$ are, respectively, the bigram counts accumulated from the contemporary and background text corpora; and $\alpha$ and $\beta$ are tunable weighting parameters. For the model interpolation formula in Equation (13), $P_{Cont}(w_i|w_{i-2}w_{i-1})$ and $P_{Back}(w_i|w_{i-2}w_{i-1})$ are the trigram probabilities, respectively, estimated from the contemporary and background text corpora, and $\gamma$ is a tunable weighting parameter. A more detailed derivation of Equations (12)-(13) also can be found in [Bacchiani and Roark 2003]. In this study, we investigated the use of the above two language model adaptation approaches for Mandarin broadcast news transcription. As mentioned earlier, a corpus of contemporary Internet newswire texts collected from August to October 2002 was used for additional prediction for the linguistic events of the testing broadcast news stories collected in September 2002.

## 6. Experimental Results

In this section, we will present a series of experiments performed to assess recognition performance as a function of the feature extraction approaches, the decoding methods, and the acoustic learning and language adaptation approaches.

***Table 1. The baseline character error rates (%) achieved using different feature extraction approaches.***

| | Character Error Rate (%) | |
|---|---|---|
| | TS | WG |
| MFCC | 26.34 | 22.55 |
| LDA-1 | 23.10 | 19.90 |
| LDA-2 | 23.13 | 19.97 |
| LDA-2+Acoustic Look-ahead | 23.24 | 20.12 |

## 6.1 The Baseline Results

The baseline broadcast news system was alternatively configured using the conventional MFCC-based and data-driven LDA-based feature extraction approaches. The results are shown in rows 3 to 5 of Table 1, where the third (MFCC) row lists the results obtained using the MFCC-based approach, and the fourth (LDA-1) and fifth (LDA-2) rows list, respectively, the results obtained when different sets of basic vectors were adopted during the construction of the LDA transformation matrix. In LDA-1, the cepstral coefficients are taken as the basic vector, while in LDA-2, the outputs of filter banks as the basic vector. As can be seen in Table 1, the character error rates obtained, respectively, using the two variant LDA-based approaches, after either tree search (TS) or word-graph rescoring (WG), were significantly better than those obtained using the standard MFCC-based approach. Moreover, LDA-2, which uses the filter bank outputs directly as the basic vector, was even more efficient than the MFCC-based approach due to the fact that the discrete cosine transform as well as the first- and second-order time derivative operations could be excluded from front-end processing. The LDA-2 features were, thus, chosen as the default acoustic features for the experiments described below.

## 6.2 Experiments on Acoustic Look-Ahead Using Syllable-Level Heuristics

The recognition performance and efficiency, after the acoustic look-ahead technique was integrated into the system, were evaluated. These results were obtained by using the same beam pruning threshold as that previously reported in section 6.1 and were run on an ordinary 2.6 GHz Pentium IV PC. The search efficiency results are shown in columns 2 to 6 of Table 2, which list, respectively, the real time factors for feature extraction and HMM state emission probability calculation (FE), acoustic look-ahead ($L_{AC}$), tree search (TS), word-graph rescoring (WG), and the overall recognition time (Total), while the recognition accuracy results are shown in the last row of Table 1. The numbers in the parentheses in the last row of Table 2 are the relative speedups achieved compared to the results shown in the second row.

***Table 2. Recognition efficiency achieved as acoustic model look-ahead was further applied. The recognition efficiency is expressed in terms of the real time factor.***

| | FE | $L_{AC}$ | TS | WG | Total |
|---|---|---|---|---|---|
| Without Acoustic Look-ahead | 0.323 | 0.000 | 1.264 | 0.196 | 1.783 |
| With Acoustic Look-ahead | 0.323 | 0.004 | 0.738 (41.61%) | 0.149 (23.98%) | 1.214 (31.91%) |

Comparing the results shown in the last two rows of Table 1, it can be found that the recognition accuracy was slightly degraded (e.g., the character error rate increased from 19.97% to 20.12% after word-graph rescoring) when acoustic look-ahead was used. However, according to the results shown in Table 2, the recognition efficiency for tree search improved significantly (a relative improvement of 41.61% was obtained) while the time spent on acoustic look-ahead (0.004 real time factor) was almost negligible. In summary, the acoustic look-ahead method proposed here achieves an overall speedup of more than 31% and enables the whole system to run almost in real time.

***Table 3. The character error rates (%) achieved with different amounts of automatically transcribed speech training data.***

| | Character Error Rate (%) | |
|---|---|---|
| | WG | +MLLR |
| Original 4 Hours | 20.12 | 18.77 |
| +5 Hours (Thr=0.9) | 16.60 | 15.84 |
| +21 Hours (Thr=0.8) | 15.34 | 14.71 |
| +33 Hours (Thr=0.7) | 15.78 | 15.02 |
| +48 Hours (Thr=0.6) | 15.62 | 14.93 |
| +54 Hours (Thr=0.5) | 15.60 | 14.92 |
| +60 Hours (Thr=0.4) | 15.49 | 14.84 |

## 6.3 Experiments on Lightly Supervised Acoustic Model Training

Table 3 summarizes the performance of lightly supervised acoustic model training. Column 2 (WG) shows the recognition results achieved using several sets of acoustic models, which were trained by selectively combining different amounts of automatically transcribed speech data with the original 4-hour manually transcribed speech data. Column 1 indicates the actual sizes of the automatically transcribed speech data selected, and the numbers in parentheses are the corresponding word confidence thresholds used. In addition, the third column presents the

results obtained when online unsupervised MLLR (Maximum Likelihood Linear Regression) speaker adaptation was further included [Gales and Woodland 1996]. It can been found from Table 3, that with careful selection of automatically transcribed speech data, the character error rate could be effectively reduced from 20.12% to 15.34% (a relative improvement of 23.76% was obtained) when a total of 21 hours of automatically transcribed data were selected for acoustic training, in combination with the original 4-hour manually transcribed data. Use of the word confidence measure aided selection of the best subset of automatically transcribed data for acoustic training. Meanwhile, use of the online unsupervised MLLR speaker adaptation technique also resulted in additional performance gains under all experimental conditions.

***Table 4. The character error rates (%) and perplexities achieved as the language models are adapted with contemporary text corpus using either the count merging and model interpolation strategies.***

| | | Character Error Rate (%) | | Perplexity |
|---|---|---|---|---|
| | | WG | +MLLR | |
| No LM Adaptation | | 15.34 | 14.71 | 670.23 |
| Count Merging | $\alpha$ =1, $\beta$ =1 | 13.22 | 12.60 | 437.87 |
| | $\alpha$ =3, $\beta$ =1 | 12.89 | 12.17 | 367.18 |
| | $\alpha$ =5, $\beta$ =1 | 12.95 | 12.22 | 397.80 |
| | $\alpha$ =7, $\beta$ =1 | 13.06 | 12.36 | 425.22 |
| | $\alpha$ =9, $\beta$ =1 | 13.15 | 12.46 | 450.99 |
| Model Interpolation | $\gamma$ =0.1 | 13.19 | 12.48 | 517.12 |
| | $\gamma$ =0.3 | 12.63 | 11.99 | 411.62 |
| | $\gamma$ =0.5 | 12.47 | 11.88 | 373.92 |
| | $\gamma$ =0.7 | 12.49 | 11.91 | 359.26 |
| | $\gamma$ =0.9 | 12.68 | 12.06 | 363.34 |

## 6.4 Experiments on Language Model Adaptation

The language adaptation results obtained using the contemporary text corpus are shown in Table 4. The second row shows the character error rates and perplexity for the system without language model adaptation. It can be seen that the character error rates are the best ones shown in Table 3, and that the initially achieved perplexity value was 670.23. This high perplexity value was probably obtained because the local word regularity properties of the tested broadcast news stories were not modeled very well by the background language models. The

rest of the rows show, respectively, the results obtained for the systems when either the count merging adaptation strategy or the model interpolation adaptation strategy was adopted. In this study, for count merging, the value of weighting parameter $\beta$ was fixed at 1, and the value of weighting parameter $\alpha$ was varied from 1 to 9 with a step size of 2; meanwhile, for model interpolation, the value of weighting parameter $\gamma$ was varied from 0.1 to 0.9 with a step size of 0.2. Comparatively speaking, the best results for model interpolation ( $\gamma = 0.5$ or $\gamma = 0.7$ ) were slightly better than those for count merging ( $\alpha = 3, \beta = 1$ ),in terms of either the character error rate or perplexity reductions. The character rate decreased significantly from 14.71% to 11.88% ( $\gamma = 0.5$ and +MLLR), and the perplexity value also can be reduced from 670.23 to 359.26 ( $\gamma = 0.7$ ), which is just about a half of the original perplexity value. The above results reveal that the local word regularity (or contextual) information that can be obtained from the contemporary corpus is vital for the task of Mandarin broadcast news recognition, whereas the subject domains or topical information embedded in the contemporary corpus may be worth taking into account and exploring further when performing language model adaptation.

## 7. Conclusions

This paper has presented the initial results of a long-term research project on automatic recognition, indexing and summarization of Mandarin speech information. Several improved approaches to Mandarin broadcast news speech recognition have been presented. With the special structural properties of the Chinese language taken into consideration, a fast acoustic look-ahead technique using syllable-level heuristics has been proposed, and an overall speedup of more than 31% has been achieved in experiments. A verification-based method for automatic acoustic data acquisition has also been proposed to make use of large amount of untranscribed speech data, and very encouraging recognition results have been obtained. Two alternative strategies for language model adaptation have also been shown to be helpful in reducing both the character error rate and perplexity. The broadcast news system finally yielded an 11.88% character error rate when applied to a Mandarin broadcast news test set.

## Acknowledgements

The authors are thankful to the National Science Council, R.O.C., for financial supports of this work (grant no. 91-2218-E-003-002 and 92-2213-E-003-008). They also thank the NTU Speech Processing Lab for providing the necessary speech and language data.

## References

Aubert, X. L., “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language*, vol. 16, 2002, pp. 89-114.

Bacchiani, M. and B. Roark, "Unsupervised Language Model Adaptation," *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2003, pp. 224-227.

Bellegarda, J. R., "Statistical Language Model Adaptation: Review and Perspectives," *Speech Communication*, vol. 42, 2004, pp. 93-108.

Berger, A., S. D. Pietra and V. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, 1996, pp. 39-71.

Beyerlein, P., X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz and A. Sixtus, "Large Vocabulary Continuous Speech Recognition of Broadcast News – The Philips/RWTH Approach," *Speech Communication*, vol. 37, 2002, pp. 109-131.

Chang, E., F. Seide, H. Meng, Z. Chen, Y. Shi and Y.C. Li, "A System for Spoken Query Information Retrieval on Mobile Devices," *IEEE Trans. on Speech and Audio Processing,* vol. 10, no. 5, 2002, pp. 531-541.

Chang, P.C., S.P. Liao and L.S. Lee, "Improved Chinese Broadcast News Transcription by Language Modeling with Temporally Consistent Training Corpora and Iterative Phrase Extraction," *Proc. European Conference on Speech Communication and Technology,* 2003, pp. 421-424.

Chen, L., J.-L. Gauvain, L. Lamel and G. Adda, "Unsupervised Language Model Adaptation for Broadcast News," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing,* vol. I, 2003, pp. 220-223.

Chen, B., H.M. Wang, L.F. Chien and L.S. Lee, "A*-Admissible Key-Phrase Spotting with Sub-Syllable Level Utterance Verification," P*roc. International Conference on Spoken Language Processing,* 1998, CD-ROM.

Chen, B., H.M. Wang and L.S. Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese," *IEEE Trans. on Speech and Audio Processing,* vol. 10, no. 5, 2002, pp. 303-314.

Chen, B., J.W. Kuo and W.H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing,* vol. I, 2004, pp. 777-780.

Chen, S.F. and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," *Computer Speech and Language,* vol. 13, 1999, pp. 359-394.

Chen, S.F. and R. Rosenfeld, "A Survey of Smoothing Techniques for ME Models," *IEEE Trans. on Speech and Audio Processing,* vol. 8, no. 1, 2000, pp. 37-50.

Croft, W.B., (editor) and J. Lafferty (editor), Language Modeling for Information Retrieval, Kluwer-Academic Publishers, 2003.

Davis, S.B. and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustic, Speech, and Signal Processing,* vol. 28, no. 4, 1980, pp. 357-366.

Dempster, A.P., N. M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of Royal Statistical Society B*, 1977, vol. 39, no. 1, pp. 1-38.

Duda, R.O. and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973.

Gales, M.J.F. and P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language,*" vol. 10, 1996, pp. 249-264.

Gauvain, J.-L., L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication,* vol. 37, 2002, pp. 89-108.

Goodman, J., "A Bit of Progress in Language Modeling," *Computer Speech and Language,* vol. 15, 2001, pp. 403-434.

Evermann, G. and P.C. Woodland, "Design of Fast LVCSR Systems," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding,* 2003, pp. 7-12.

Federico, M. and N. Bertoldi, "Broadcast News LM adaptation Using Cotemporary Texts," *Proc. European Conference on Speech Communication and Technology,* vol. 1, 2001, pp. 239-342.

Furui, S., T. Kikuchi, Y. Shinnaka and C. Hori, "Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech," *IEEE Trans. on Speech and Audio Processing,* vol. 12, no. 4, 2004, pp. 401-408.

Jelinek, F., B. Merialdo, S. Roukos and M. Strauss, "A Dynamic Language Model for Speech Recognition," *Proc. Speech and Natural Language DARPA Workshop,* 1991, pp. 293-295.

Kemp, T. and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Proc. European Conference on Speech Communication and Technology,* vol. 6, 1999, pp. 2725-2728.

Kneser, R. and H. Ney, "Improved Backing-off for M-gram Language Modeling," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing,* vol. I, 1995, pp. 181-184.

Lamel, L., J.L. Gauvain and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, vol. 16, no.1, 2002, pp. 115-229.

LDC 2003, Chinese Gigaword Corpus: http://www.ldc.upenn.edu.

Lee, L.S., "Voice Dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, vol. 14 no. 4, 1997, pp. 63-101.

Liu, X. and W.B. Croft, "Statistical Language Modeling for Information Retrieval," to appear in *Annual Review of Information Science and Technology,* vol. 39, 2005.

Macherey, W. and H. Ney, "Towards Automatic Corpus Preparation for a German Broadcast News Transcription System," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing,* vol. I, 2002, pp. 733-736.

Meng, H., B. Chen, S. Khudanpur, G. A. Levow, W. K. Lo, D. Oard, P. Schone, K. Tang, H.M. Wang and J. Wang, "Mandarin English Information (MEI): Investigating Translingual Speech Retrieval," *Computer Speech and Language,* vol. 18, no. 2, 2004, pp. 163-179.

Nguyen, L. and B. Xiang, "Light Supervision in Acoustic Model Training," *Proc. IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2004, pp. 185-188.

Ortmanns, S., H. Ney and X. Aubert, "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language,* vol. 11, 1997, pp. 43-72.

Ortmanns, S. and H. Ney, "Look-ahead Techniques for Fast Beam Search," *Computer Speech and Language,* vol. 14, 2000, pp. 15-32.

Rosenfeld, R., "Two Decades of Statistical Language Modeling: Where Do We Go from Here," *Proc. IEEE*, vol. 88, no. 8, 2000, pp. 1270-1278.

Saon, G. and M. Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition," *IEEE Trans. on Speech And Audio Processing,* vol. 9, no. 4, 2001, pp. 327-332.

Schuster, M., "Memory-efficient LVCSR Search Using a One-Pass Stack Decoder," *Computer Speech and Language,* vol. 14, 2000, pp. 47-77.

Stolcke, A., SRI language Modeling Toolkit, version 1.3.3, 2000. http://www.speech.sri.com/projects/srilm/.

Wang, C.J., B. Chen and L.S. Lee, "Improved Chinese Spoken Document Retrieval with Hybrid Modeling and Data-driven Indexing Features," *Proc. International Conference on Spoken Language Processing,* 2002, pp. 1985-1988.

Wessel, F. and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding,* 2001, pp. 307-310.

Wessel, F., R. Schluter, K. Macherey and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing,* vol. 9, no 3, 2001, pp. 288-298.

Woodland, P.C., "The Development of the HTK Broadcast News Transcription System: An Overview," *Speech Communication,* vol. 37, 2002, pp. 47-67.

Zhai, C.X. and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," *ACM Trans. on Information Systems,* vol. 22, no. 2, 2004, pp. 179-214.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 19-34 


# Reduced *N*-Grams for Chinese Evaluation

**Le Quan Ha[*], R. Seymour[+], P. Hanna[*] and F. J. Smith[*]**

**Abstract**

Theoretically, an improvement in a language model occurs as the size of the *n*-grams increases from 3 to 5 or higher. As the *n*-gram size increases, the number of parameters and calculations, and the storage requirement increase very rapidly if we attempt to store all possible combinations of *n*-grams. To avoid these problems, the reduced *n*-grams' approach previously developed by O' Boyle and Smith [1993] can be applied. A reduced *n*-gram language model, called a reduced model, can efficiently store an entire corpus's phrase-history length within feasible storage limits. Another advantage of reduced *n*-grams is that they usually are semantically complete. In our experiments, the reduced *n*-gram creation method or the O' Boyle-Smith reduced *n*-gram algorithm was applied to a large Chinese corpus. The Chinese reduced *n*-gram Zipf curves are presented here and compared with previously obtained conventional Chinese *n*-grams. The Chinese reduced model reduced perplexity by 8.74% and the language model size by a factor of 11.49. This paper is the first attempt to model Chinese reduced *n*-grams, and may provide important insights for Chinese linguistic research.

**Keywords:** Reduced *n*-grams, reduced *n*-gram algorithm / identification, reduced model, Chinese reduced *n*-grams, Chinese reduced model

## 1. Introduction to the Reduced *N*-Gram Approach

P O' Boyle and F J Smith [1992, 1993] proposed a statistical method to improve language models based on the removal of overlapping phrases.

The distortion of phrase frequencies were first observed in the Vodis Corpus when the bigram "RAIL ENQUIRIES" and its super-phrase "BRITISH RAIL ENQUIRIES" were examined and reported by O' Boyle. Both occur 73 times, which is a large number for such a small corpus. "ENQUIRIES" follows "RAIL" with a very high probability when it is preceded by "BRITISH." However, when "RAIL" is preceded by words other than "BRITISH," "ENQUIRIES" does not occur, but words like "TICKET" or "JOURNEY" may. Thus, the

[*] Computer Science School, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK.
Email: {q.le, p.hanna, fj.smith}@qub.ac.uk
[+] Email: rowan@rowan.ws

bigram "RAIL ENQUIRIES" gives a misleading probability that "RAIL" is followed by "ENQUIRIES" irrespective of what precedes it. At the time of their research, Smith and O' Boyle reduced the frequencies of "RAIL ENQUIRIES" by using the frequency of the larger trigram, which gave a probability of zero for "ENQUIRIES" following "RAIL" if it was not preceded by "BRITISH." This problem happens not only with word-token corpora but also corpora in which all the compounds are tagged as a unit since overlapping *n*-grams still appear.

Therefore, a phrase can occur in a corpus as a reduced *n*-gram in some places and as part of a larger reduced *n*-gram in other places. In a reduced model, the occurrence of an *n*-gram is not counted when it is a part of a larger reduced *n*-gram. One algorithm to detect/identify/extract reduced *n*-grams from a corpus is the so-called reduced *n*-gram algorithm. In 1992, P O' Boyle and F J Smith were able to store the entire content of the Brown corpus of American English [Francis and Kucera 1964] (of one million word tokens, whose longest phrase-length is 22), which was a considerable improvement at the time. There was no additional way for O' Boyle to evaluate the reduced *n*-grams, so his work was incomplete. We have developed and present here our perplexity method, and we discuss its usefulness for reducing *n*-gram perplexity.

## 2. Similar Approaches and Capability

Recent progress in variable *n*-gram language modeling has provided an efficient representation of *n*-gram models and made the training of higher order *n*-grams possible. Compared to variable *n*-grams, class-based language models are more often used to reduce the size of a language model, but this typically leads to recognition performance degradation. Classes can alternatively be used to smooth a language model or provide back-off estimates, which have led to small performance gains but also an increase in language model size.

For the LOB corpus, the varigram model obtained 11.3% higher perplexity in comparison with the word-trigram model [Niesler and Woodland 1996], but it also obtained a 22-fold complexity decrease.

Reinhard Kneser [1996] built up variable-context length language models based on North American Business News (NAB - 240 million words of newspaper data) and the German Verbmobil (300,000 words with a vocabulary of 5,000 types). His results show that the variable-length model outperforms conventional models of the same size, and if a moderate loss in performance is acceptable, that the size of a language model can be reduced drastically by using his pruning algorithm. Kneser's results improve with longer contexts and the same number of parameters. For example, reducing the size of the standard NAB trigram model by a factor of 3 results in a loss of only 7% in perplexity and 3% in the word error rate.

The improvement obtained by Kneser's method depends on the length of the fixed context and on the amount of available training data. In the case of the NAB corpus, the improvement was 10% in perplexity.

M. Siu and M. Ostendorf [2000] developed Kneser's basic ideas further and applied the variable 4-gram, thus improving the perplexity and word error rate results compared to a fixed trigram model. The obtained word error reductions of 0.1 and 0.5% (absolute) in development and evaluation test sets, respectively, were not statistically significant. However, the number of parameters was reduced by 60%. By using the variable 4-gram, they were able to model a longer history while reducing the size of the model by more than 50%, compared to a regular trigram model, and at the same time improve both the test-set perplexity and recognition performance. They also reduced the size of the model by an additional 8%.

Another related work was that of Hu, Turin, Brown [1997].

## 2.1 The first algorithm [R Kneser 1996]

Variable-length models are determined by the set $S$ of word sequences. If $T$ is the set of all word sequences in the training data with a maximal length of $M$, then variable-length models can be created by finding a suitable subset $S$ of the set $T$ of all the $M$-gram sequences in the training data with a given maximal context length $M$. The distance measure between model $P_S$ and model $P_T$ is as follows:

$$D_2(P_T \parallel P_S) := \sum_{k=0}^{M-1} \sum_{(h_k, w) \in T \setminus S} d_1(h_k, w), \tag{1}$$

where the terms of the sum are defined by the average Kullback Leiber distance

$$d_1(h_k, w) := P_T(h_k, w) \log \frac{P_T(w \mid h_k)}{\gamma_T(h_k) P_T(w \mid h_{k-1})}, \tag{2}$$

where $h_k$ is a phrase history of word $w$ and $\gamma$ is the normalisation factor.

In the implementation, they store the word sequences of $S$ in a tree structure. Each node of the tree corresponds to a word sequence, and each arc is labeled with a word identity. For each node $W = (h_k, w) \in S$, $Succ(W)$ is the set of all longer word sequences starting with the same words as $W$. If a node $W$ is removed, then all $Succ(W)$ will be removed.

Therefore, the average contribution to the sum $d_2$ is

$$d_2(W) = \frac{d_1(w) + \sum_{V \in Succ(W)} d_1(V)}{1 + |Succ(W)|}. \tag{3}$$

The pruning algorithm is as follows:

*Start with* $S = T$

*While* $(|S| > K)$

*For all nodes in S calculate* $d_2$

*Remove node with lowest* $d_2$

## 2.2 The second algorithm [T R Niesler and P C Woodland 1996]

1. *Initialisation:* $L = -1$
2. $L = L + 1$
3. *Grow:* Add level #$L$ to level #$(L-1)$ by adding all the $(L+1)$-Grams occurring in the training set for which the $L$-Grams already exist in the tree.
4. *Prune:* For every (newly created) leaf in level #$L$, apply a quality criterion and discard the leaf if it fails.
5. *Termination:* If there is a nonzero number of leaves remaining in level #$L$, goto step 2.

The quality criterion checks for improvement in the leaving-one-out cross-validation training set likelihood achieved by the addition of each leaf.

## 2.3 Combination of variable *n*-grams and other language model types

Using the first algorithm, M Siu and M Ostendorf [2000] combined their variable *n*-gram method with the skipping distance method and class-based method in a study on the Switchboard corpus, consisting of 2 million words. In 1996, using the second algorithm, T R Niesler and P C Woodland developed the variable *n*-gram based category in a study on LOB, consisting of 1 million English words. In order to obtain an overview of variable *n*-grams, we combine all of these authors' results in Table 1.

# 3. O' Boyle and Smith's Reduced *N*-Gram Algorithm and Application Scope

The main goal of this algorithm is to produce three main files from the training text:

- The file that contains all the complete *n*-grams appearing at least *m* times is called the PHR file ($m \geq 2$).
- The file that contains all the *n*-grams appearing as sub-phrases, following the removal of the first word from any other complete *n*-gram in the PHR file, is called the SUB file.

***Table 1. Comparison of combinations of variable n-grams and other Language Models***

| COMBINATION OF LANGUAGE MODEL TYPES | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Basic *n*-gram** | **Variable *n*-grams** | **Category** | **Skipping distance** | **Classes** | **#params** | **Perpl-exity** | **Size** | **Source** |
| Trigram √ | | | | | 987k | 474 | 1M | LOB |
| | | Bigram √ | | | - | 603.2 | | |
| | | Trigram √ | | | - | 544.1 | | |
| | √ | √ | | | - | 534.1 | | |
| Trigram √ | | | | | 743k | 81.5 | 2M | Switch board Corpus |
| | Trigram √ | | | | 379k | 78.1 | | |
| | Trigram √ | | √ | | 363k | 78.0 | | |
| | Trigram √ | | √ | √ | 338k | 77.7 | | |
| | 4-gram √ | | | | 580k | 108 | | |
| | 4-gram √ | | √ | | 577k | 108 | | |
| | 4-gram √ | | √ | √ | 536k | 107 | | |
| | 5-gram √ | | | | 383k | 77.5 | | |
| | 5-gram √ | | √ | | 381k | 77.4 | | |
| | 5-gram √ | | √ | √ | 359k | 77.2 | | |

- The file that contains any overlapping *n*-grams that occur at least *m* times in the SUB file is called the LOS file.

Therefore, the final result is the FIN file of all reduced *n*-grams, where

$$\textbf{FIN} := \textbf{PHR} + \textbf{LOS} - \textbf{SUB}. \tag{4}$$

Before O' Boyle and Smith's work, Craig used a loop algorithm that was equivalent to FIN := PHR – SUB. This yields negative frequencies for resulting *n*-grams with overlapping, hence the need for the LOS file.

There are 2 additional files:

1. To create the PHR file, a SOR file is needed that contains all the complete *n*-grams regardless of *m* (the SOR file is the PHR file in the special case where $m = 1$). To create the PHR file, words are removed from the right-hand side of each SOR phrase in the SOR file until the resultant phrase appears at least *m* times (if the phrase already occurs more than *m* times, no words will be removed).

2. To create the LOS file, O' Boyle and Smith applied a POS file: for any SUB phrase, if one word can be added back on the right-hand side (previously removed when the PHR file was created from the SOR file), then one POS phrase will exist as the added phrase. Thus, if any POS phrase appears at least *m* times, its original SUB phrase will be an overlapping *n*-gram in the LOS file.

The application scope of O' Boyle and Smith 's reduced *n*-gram algorithm is limited to small corpora, such as the Brown corpus (American English) of 1 million words [1992], in which the longest phrase has 22 words. Now their algorithm, re-checked by us, still works for medium size and large corpora with training sizes of 100 million word tokens.

## 4. Reduced *N*-Grams and Zipf's Law

By re-applying O'Boyle and Smith's algorithm, we obtained reduced *n*-grams from the Chinese TREC corpus of the Linguistic Data Consortium[1], catalog no. LDC2000T52. TREC was collected from full articles in the People's Daily Newspaper from 01/1991 to 12/1993 and from Xinhua News Agency articles from 04/1994 to 09/1995. Originally, TREC had 19,546,872 syllable tokens but only 6,300 syllable types. Ha, Sicilia-Garcia, Ming and Smith [2002] proposed an extension of Zipf 's law and applied it to the TREC syllable corpus. Then in 2003, they produced a compound word version of TREC with 50,000 types, this version was employed in our study for reduced *n*-gram creation.

We will next present the Zipf curves for Chinese reduced *n*-grams, starting with syllables.

### 4.1 Chinese syllables

The TREC syllable reduced *n*-grams were created in 28 hours on a Pentium II with 512 MB of RAM and 2 GB of free hard-drive space.

The most common TREC syllable reduced unigrams, bigrams, trigrams, 4-grams and 5-grams are shown in Table 3. It can be seen that much noise existed in the unigram frequency observations when only one syllable "年 YEAR" re-appeared in the top ten syllable unigrams [Ha, Sicilia-Garcia, Ming and Smith 2002], listed in Table 2.

[1] http://www.ldc.upenn.edu/

***Table 2. The 10-highest frequency unigrams in the conventional Chinese TREC syllable corpus [Ha, Sicilia-Garcia, Ming and Smith 2002]***

| Rank | Unigrams | | |
|---|---|---|---|
| | **Freq** | **Token** | **Meaning** |
| 1 | 620,619 | 的 | Of |
| 2 | 308,826 | 国 | State |
| 3 | 219,543 | 一 | One |
| 4 | 209,497 | 中 | Centre / Middle |
| 5 | 176,905 | 在 | In / At |
| 6 | 159,861 | 和 | And |
| 7 | 143,359 | 人 | Human |
| 8 | 139,713 | 了 | Perfective Marker |
| 9 | 133,696 | 会 | Get Together / Meeting / Association |
| 10 | 128,805 | 年 | Year |

The Zipf [1949] curves are plotted for TREC syllable reduced unigrams and *n*-grams in Figure 1. It can be seen that none of the syllable unigram, bigram, trigram, 4-gram and 5-gram curves are straight. The unigram curve has an average slope of –1, while the bigram, trigram, 4-gram and 5-gram curves have slopes of around –0.5. At the beginning, they are very turbulent, crossing each other due to much observed noise at high frequencies.

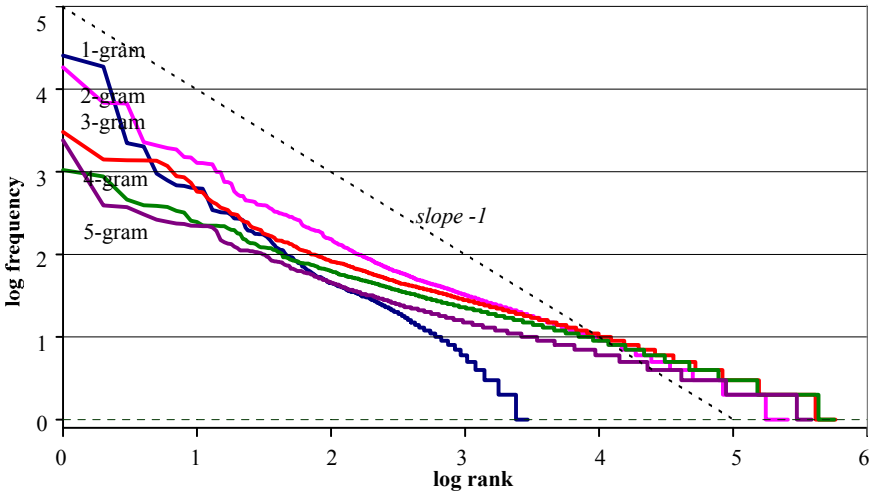



***Figure 1. TREC syllable reduced n-gram Zipf curves***

***Table 3. Most common TREC syllable reduced n-grams***

| Rank | Unigrams | | | Bigrams | | | Trigrams | | | 4-grams | | | 5-grams | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Tokens | Meaning | Freq | Tokens | Meaning | Freq | Tokens | Meaning | Freq | Tokens | Meaning | Freq | Tokens | Meaning |
| 1 | 25,434 | 月 | Month | 18,344 | 日期 | Date | 3,034 | 小标题 | Subtitle | 1,047 | 与此同时 | At the same time | 2,402 | 新华社北京 | Xinhua News Agency Beijing |
| 2 | 18,633 | 完 | Complete/ Finish | 6,828 | 日电 | News on | 1,406 | 据介绍 | As is reported | 876 | 本报北京 | The Beijing Agency of the Paper | 390 | 新华社东京 | Xinhua News Agency Tokyo |
| 3 | 2,211 | 年 | Year | 6,698 | 他说 | He said | 1,377 | 亿美元 | Hundred million dollars | 459 | 据新华社 | According to Xinhua News | 371 | 到目前为止 | Up to date |
| 4 | 2,019 | 秒 | Second | 2,270 | 目前 | At the present | 1,366 | 近年来 | In recent years | 394 | 今年以来 | Since this year | 303 | 新华社巴黎 | Xinhua News Agency Paris |
| 5 | 949 | 个 | Individual | 2,080 | 因此 | Therefore | 1,360 | 据了解 | According (to) investigation | 385 | 近几年来 | In recent years | 263 | 新华社伦敦 | Xinhua News Agency London |
| 6 | 786 | 到 | Arrive/ Reach | 1,929 | 但是 | But | 1,168 | 据统计 | According (to) statistics | 374 | 容加金旁 | Contain plus gold beside | 247 | 新华社广岛 | Xinhua News Agency Guang Dao |
| 7 | 683 | 倍 | Double | 1,854 | 此外 | In addition | 881 | 他指出 | He indicates | 337 | 钱其琛说 | Qian Qi Chen said | 237 | 今年上半年 | In the first half of the year |
| 8 | 671 | 米 | Meter/ Rice | 1,506 | 据悉 | As is known | 859 | 据报道 | It is reported | 321 | 江泽民说 | Jiang Zemin said | 233 | 新华社天津 | Xinhua News Agency Tian Jin |
| 9 | 641 | 二 | Two | 1,482 | 同时 | At the same time | 735 | 他认为 | He thinks | 255 | 另一方面 | On the other hand | 223 | 他表示相信 | He said he believed |
| 10 | 630 | 元 | Dollar | 1,278 | 今年 | This year | 573 | 李鹏说 | Li Peng said | 246 | 今天上午 | This morning | 221 | 新华社波恩 | Xinhua News Agency Bonn |

## 4.2 Chinese Compound Words

The TREC compound word reduced *n*-grams obtained using O' Boyle and Smith 's algorithm were created in 20 hours (we executed the algorithm non-stop for less than one day on a Pentium II with 512 MB of RAM) with a storage requirement of only 1 GB.

The most common TREC word reduced unigrams, bigrams, trigrams, 4-grams and 5-grams are shown in Table 5. One can observe noises in the unigram frequency observations when words with more than 1 syllable appeared in the top ten ("日期 Date," "目前 Currently," "小标题 Subtitle," "因此 Therefore," and "同时 Simultaneously"), but the 2-syllable word "中国 China" disappeared, as shown in Table 4, which lists the most common traditional TREC word unigrams [Ha *et al*. 2003].

Our observations of reduced *n*-grams show that they increase the semantic completeness of longer *n*-grams with large *n* in comparison with conventional Chinese word *n*-grams [Ha *et al*. 2003].

***Table 4. The 10-highest frequency unigrams in the conventional Chinese TREC word corpus [Ha et al. 2003]***

| Rank | Unigrams | | |
|---|---|---|---|
| | **Freq** | **Token** | **Meaning** |
| 1 | 609,395 | 的 | Of |
| 2 | 154,827 | 在 | In / At |
| 3 | 144,524 | 和 | And |
| 4 | 126,134 | 了 | Perfective Marker |
| 5 | 99,747 | 是 | Be |
| 6 | 86,928 | 一 | One |
| 7 | 77,037 | 中国 | China |
| 8 | 69,253 | 中 | Centre / Middle |
| 9 | 60,230 | 日 | Sun |
| 10 | 57,045 | 为 | For |

***Table 5. Most common TREC compound word reduced n-grams***

| Rank | Unigrams | | | Bigrams | | | Trigrams | | | 4-grams | | | 5-grams | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Tokens | Meaning | Freq | Tokens | Meaning | Freq | Tokens | Meaning | Freq | Tokens | Meaning | Freq | Tokens | Meaning |
| 1 | 26,831 | 月 | Month | 7,253 | 日 电 | Daily News | 2,849 | 星期 星期二 版次 | Week Tuesday order | 642 | 现在 广播 完 了 | Modern broadcast finishes | 405 | 版名 政治 法律 社会 标题 | Political social law page title |
| 2 | 19,325 | 完 | Complete/ Finish | 6,849 | 他 说 | He says | 2,703 | 星期 星期一 版次 | Week Monday order | 374 | 容 加 金 旁 | Contain plus gold beside[2] | 366 | 一 九 九 四 年 | Year nineteen ninety four |
| 3 | 18,406 | 日期 | Date | 2,404 | 新华社 北京 | Xinhua news agency Beijing | 2,686 | 星期 星期六 版次 | Week Saturday order | 273 | 人民币 市场 汇 价 | Renminbi market exchange price | 340 | 版名 教育 科技 文化 标题 | Educational cultural & technology page title |
| 4 | 8,423 | 年 | Year | 1,530 | 万 元 | Ten-thousand yen | 2,629 | 星期 星期四 版次 | Week Thursday order | 263 | 据 不 完全 统计 | According (to) incomplete statistics | 271 | 外 币 名称 中间 价 | Foreign money Ming-Chien transferring price |
| 5 | 4,188 | 日 | Sun | 1,516 | 亿 美元 | Ten-million US Dollars | 2,585 | 星期 星期五 版次 | Week Friday order | 239 | 中国 经济 简 讯 | China economic brief news | 270 | 一 九 九 三 年 | Year nineteen ninety three |
| 6 | 3,998 | 目前 | Currently | 1,450 | 亿 元 | Ten-million yen | 2,491 | 星期 星期日 版次 | Week Sunday order | 221 | 十 万 日 元 | One hundred thousand Japanese yen | 185 | 个 国家 和 地区 的 | Individual international and region of |
| 7 | 3,292 | 分 | Unit | 1,407 | 据 介绍 | According (to) introduction | 2,408 | 星期 星期三 版次 | Week Wednesday order | 220 | 一百 欧洲 货币 单位 | A hundred European currency units | 160 | 日 电 据 外电 报道 | Daily news according to foreign newspaper report |
| 8 | 3,034 | 小标题 | Subtitle | 1,360 | 据 了解 | According (to) understanding | 876 | 本 报 北京 | Our newspaper Beijing | 212 | 上 接 第一 版 | Continue from No.1 edition | 139 | 一 九 九 二 年 | Year nineteen ninety two |
| 9 | 2,348 | 因此 | Therefore | 1,019 | 日 至 | Date to | 788 | 版名 要闻 标题 | Abstract page title | 134 | 中国 文化 简 讯 | China cultural brief news | 96 | 版名 要 闻 正 文 | Brief news major content page |
| 10 | 2,299 | 同时 | Simultaneously | 891 | 他 指出 | He indicates | 572 | 版名 经济 标题 | Economic page title | 127 | 日 电 综 述 | Daily News generally speaking | 91 | 一 九 九 一 年 | Year nineteen ninety one |

[2]It is remarkable that many of the *n*-grams with larger *n* values contain content markup information.

Zipf curves for TREC word reduced unigrams and *n*-grams are plotted in Figure 2. It can be observed that the unigram curve not straight, but rather exhibits a two-slope behaviour, beginning with a slope of –0.67 and then falling-off with a slope of approximately -2 at the end. All the bigram, trigram, 4-gram, and 5-gram curves have slopes in the range [-0.6, -0.5] and have become more parallel and straighter. Noise is visible among the TREC word reduced bigrams, trigrams, 4-grams and 5-grams where they turbulently cross each other at the beginning.

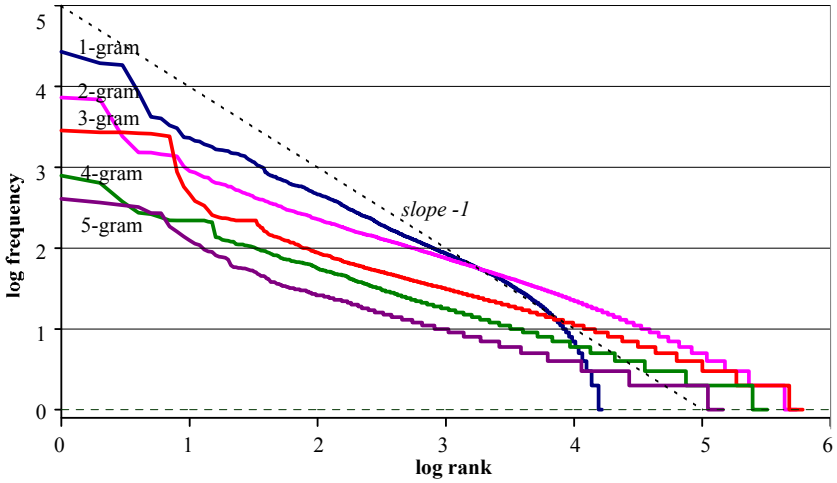



***Figure 2. TREC word reduced n-gram Zipf curves***

Usually, Zipf 's rank-frequency law is contradicted by empirical data, and the syllable and compound-word reduced *n*-grams from Chinese shown in Figure 1 and Figure 2 also contradict it. In fact, various more sophisticated models for frequency distributions have been proposed by Baayen [2001] and Evert [2004].

## 5. Perplexity for Chinese Reduced *N*-Grams

The reduced *n*-gram approach was also checked by means of Chinese compound-word perplexity calculations based on the Weighted Average Model of O' Boyle and Smith [1993, 1994, 1995, 1997], which was further developed by Sicilia-Garcia, Ming, Smith and Hanna [1999, 2000, 2001]. We rewrote this famous model in formulae (5) and (6):

$$wgt\left(w_j^i\right) = \log\left(f\left(w_j^{i-1}\right)\right) \times 2^{i-j+1}, \tag{5}$$

$$P_{WA}\left(w_i \mid w_{i-N+1}^{i-1}\right) = \frac{wgt\left(w_i\right) \times P\left(w_i\right) + \sum_{l=1}^{N-1} wgt\left(w_{i-l}^i\right) \times P\left(w_i \mid w_{i-l}^{i-1}\right)}{\sum_{l=0}^{N-1} wgt\left(w_{i-l}^i\right)}. \tag{6}$$

Next, we will analyse the main difficulties arising from perplexity calculations for our reduced model: the statistical model problem, unseen word problem and unknown word problem.

## 5.1 Statistical model problem

In a reduced model, the following rules apply:

- If $f\left(w_{i-l}^i\right) > 0$, but $f\left(w_{i-l}^{i-1}\right) = 0$, then the maximum likelihood $P\left(w_i \middle| w_{i-l}^{i-1}\right) = \frac{f\left(w_{i-l}^i\right)}{f\left(w_{i-l}^{i-1}\right)}$ and the weight $wgt\left(w_{i-l}^i\right) = \log\left(f\left(w_{i-l}^{i-1}\right)\right) \times 2^{l+1}$ will be undefined.
- Once the weight calculation has been performed, if $P\left(w_i \middle| w_{i-l-L_0}^{i-1}\right) > 0$ and the previous $L_0$ phrases $P\left(w_i \middle| w_{i-l-L_0+1}^{i-1}\right)$, $P\left(w_i \middle| w_{i-l-L_0+2}^{i-1}\right)$, …, $P\left(w_i \middle| w_{i-l}^{i-1}\right)$ are all 0, then we should include the $L_0$ phrases' weights of zero probability $wgt\left(w_{i-l-L_0+1}^i\right)$, $wgt\left(w_{i-l-L_0+2}^i\right)$, …, $wgt\left(w_{i-l}^i\right)$ into the sum of weights in the denominator of formula (6).

## 5.2 Unseen word problem

If $P_{WA}\left(w_i \mid w_{i-N+1}^{i-1}\right) = 0$ but $w_i$ occurs in other reduced *n*-grams, then how can we calculate the probability ?

## 5.3 Unknown word problem

If $P_{WA}\left(w_i \mid w_{i-N+1}^{i-1}\right) = 0$ and $w_i$ does not occur in any other reduced *n*-grams, then $w_i$ will be totally unknown and we will not be able to apply the Turing-Good probability. This is because an unusual phenomenon will occur with the hapax legomena $n_1$ and dis legomena $n_2$ when $n_1 < n_2$, and because the Turing-Good probabilities will become too high if we use $T_{\text{reduced}}$, as shown in Table 6.

***Table 6. Unusual Turing-Good observations with respect to reduced models***

| | $n_1$ | $N_2$ | $T_{\text{reduced}}$ | Turing-Good reduced probability |
|---|---|---|---|---|
| TREC reduced words | 1,620 | 2,171 | 17,515 | 0.0001530258 |

The Turing-Good probability for the conventional TREC word corpus is 7.082435E-08. For the reduced model shown in Table 6, the Turing-Good value is 2,161 times higher, which is unusual.

## 5.4 Solutions for reduced perplexities

- If $f\left(w_{i-l}^{i}\right) > 0$ but $f\left(w_{i-l}^{i-1}\right) = 0$ and the reduced training size is $R$, then the degraded weight $wgt\left(w_{i-l}^{i}\right) = \ln(R)$ and the maximum likelihood $P\left(w_i \middle| w_{i-l}^{i-1}\right) = \dfrac{f\left(w_{i-l}^{i}\right)}{R}$ are defined in the case of an isolated unigram.
- If $P\left(w_i \middle| w_{i-l-L_0}^{i-1}\right) > 0$ but all the previous $L_0$ phrases $P\left(w_i \middle| w_{i-l-L_0+1}^{i-1}\right)$, $P\left(w_i \middle| w_{i-l-L_0+2}^{i-1}\right)$, …, $P\left(w_i \middle| w_{i-l}^{i-1}\right)$ are 0, then we will include all the weights of zero probability, i.e., $wgt\left(w_{i-l-L_0+1}^{i}\right)$, $wgt\left(w_{i-l-L_0+2}^{i}\right)$, …, $wgt\left(w_{i-l}^{i}\right)$, into the sum of the weight denominator in formula (6). This should reduce the weighted average probability in comparison with the probabilities in other cases where all the previous $L_0$ phrases exist.
- Unseen word problem: If $P_{WA}\left(w_i \middle| w_{i-N+1}^{i-1}\right) = 0$ but $w_i$ occurs in other reduced $n$-grams, then we degrade $w_i$ when words have been eliminated from the reduced model because they appear less than $m$ times. Therefore, $w_i$ will have an estimated probability of $\dfrac{m-1}{T}$, where $T$ is the overall conventional training size.
- Unknown word problem: If $P_{WA}\left(w_i \middle| w_{i-N+1}^{i-1}\right) = 0$ and $w_i$ does not occur in any other reduced $n$-grams, then $w_i$ will be assigned the Turing-Good probability.

## 5.5 Results and Discussion

The perplexities for the Chinese TREC compound-word corpus were calculated. We obtained very poor and confusing perplexity results when we investigated short contexts of reduced $n$-grams, but coped well with long contexts since the purpose and the strength of reduced

models are their ability to store phrase histories that are as long as possible as well as an entire large corpus in a compact database. The test text file had 27,485 words in 3,093 sentences and 927 paragraphs (along with 7 unknown words of 6 types and 109 unseen words of 48 unseen types) for the TREC conventional *n*-gram model and also for the conventional and reduced models.

Our Chinese TREC word reduced model was stored in 50 MB of memory, and the perplexity investigation started with phrase-lengths of 10 words and more and increased until all the phrases had been analysed. The perplexity results obtained using the TREC word reduced model are shown in Table 7.

***Table 7. Reduced perplexities for Chinese TREC words obtained using the weighted average model***

| | **Traditional *n*-grams** | | **Reduced *n*-grams** | | **The cost of reduced *n*-grams on baseline trigrams** | **Factor of reduced model size** |
|---|---|---|---|---|---|---|
| **Phrase Length** | **Unigram** | 1,515.03 | **10-grams** | 128.35 | **-19.25%** | |
| | **Bigram** | 293.96 | **11-grams** | 131.95 | **-16.99%** | |
| | **Trigram** | **158.95** | **12-grams** | 134.77 | **-15.21%** | |
| | **4-gram** | 140.81 | **13-grams** | 136.98 | **-13.82%** | |
| | **5-gram** | 137.61 | **14-grams** | 138.68 | **-12.75%** | |
| | **6-gram** | 137.31 | **15-grams** | 140.01 | **-11.91%** | |
| | **7-gram** | 137.25 | **Complete contexts** | **145.06** | **-8.74%** | **11.49** |

Surprisingly for TREC word reduced *n*-grams, we achieved an 8.74% perplexity reduction, and the model size was reduced by a factor of 11.49. Thus, in our study on Chinese TREC words, we achieved improvement in both perplexity and model size.

## 6. Conclusions

The conventional *n*-gram language model is limited in terms of its ability to represent extended phrase histories because of the exponential growth in the number of parameters. To overcome this limitation, we have re-investigated the approach of O' Boyle and Smith [1992, 1993] and created a Chinese reduced *n*-gram model. The main advantage of Chinese reduced *n*-grams is that they have quite complete semantic meanings thanks to their creation process, starting from execution of whole sentence contexts.

Chinese reduced word and character Zipf curves and perplexity calculations along with the model size for TREC, a large Chinese corpus, have been presented. The reduced Chinese

syllable unigram Zipf curve has a slope of –1, which satisfies Zipf's law, and the reduced TREC word unigram Zipf curve shows a two-slope behaviour, similar to the curves reported by Ferrer and Solé [2002]. The difficulties with reduced model perplexity calculations due to statistical, unseen and unknown problems have been solved using the Weighted Average Model, a back-off probability model developed by O' Boyle and Smith [1993, 1994, 1995, 1997]. By extending TREC word reduced *n*-grams, we achieved an 8.74% perplexity reduction, and we were able to reduce the model size by a factor of 11.49. This remarkable improvement in the Chinese TREC reduced *n*-gram distribution may be smaller that that possible with the English language, in which the meaning of a word is clearer. This confirms Siu and Ostendorf 's [2000] conclusions concerning the potential application of their variable *n*-grams to Chinese (and Japanese) and other languages besides English.

## Acknowledgements

The authors would like to thank Maria Husin and Dr Sicilia-Garcia for valuable support, the reviewers for their valuable comments and David Ludwig for his revision.

## References

Baayen, H., "Word Frequency Distributions," Kluwer Academic Publishers, 2001.

Evert, S., "A Simple LNRE Model for Random Character Sequences," *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, 2004, pp. 411-422.

Ferrer I Cancho, R. and R. V. Solé, "Two Regimes in the Frequency of Words and the Origin of Complex Lexicons," *Journal of Quantitative Linguistics*, 8(3) 2002, pp. 165-173.

Francis, W. N. and H. Kucera, "Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers," Department of Linguistics, Brown University, Providence, Rhode Island, 1964.

Ha, L. Q., E. I. Sicilia-Garcia, J. Ming and F. J. Smith, "Extension of Zipf 's Law to Word and Character *N*-Grams for English and Chinese," *Journal of Computational Linguistics and Chinese Language Processing*, 8(1) 2003, pp. 77-102.

Ha, L. Q., E. I. Sicilia-Garcia, J. Ming and F. J. Smith, "Extension of Zipf 's Law to Words and Phrases," *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, 2002, pp. 315-320.

Hu, J., W. Turin and M. K. Brown, "Language Modeling using Stochastic Automata with Variable Length Contexts," *Computer Speech and Language*, vol. 11, 1997, pp. 1-16.

Kneser, R., "Statistical Language Modeling Using a Variable Context Length," *ICSLP*, vol. 1, 1996, pp. 494-497.

Niesler, T. R., "Category-based statistical language models," St. John's College, University of Cambridge, 1997.

Niesler, T. R. and P. C. Woodland, “A Variable-Length Category-Based *N*-Gram Language Model,” *IEEE ICASSP*, vol. 1, 1996, pp. 164-167.

O' Boyle, P., J. McMahon and F. J. Smith, "Combining a Multi-Level Class Hierarchy with Weighted-Average Function-Based Smoothing,” *IEEE Automatic Speech Recognition Workshop*, Snowbird, Utah, 1995.

O' Boyle, P. and M. Owens, "A Comparison of human performance with corpus-based language model performance on a task with limited context information,” *CSNLP*, Dublin City University, 1994.

O' Boyle, P., M. Owens and F. J. Smith, "A weighted average *N*-Gram model of natural language,” *Computer Speech and Language*, vol. 8, 1994, pp. 337-349.

O’ Boyle, P. L., J. Ming, M. Owens and F. J. Smith, "Adaptive Parameter Training in an Interpolated *N*-Gram language model,” *QUALICO*, Helsinki, Finland, 1997.

O’ Boyle, P. L., “A study of an *N*-Gram Language Model for Speech Recognition,” PhD thesis, Queen’s University Belfast, 1993.

Sicilia-Garcia, E. I., “A Study in Dynamic Language Modelling,” PhD thesis, Queen’s University Belfast, 2001.

Sicilia-Garcia, E. I., J. Ming and F. J. Smith, “A Dynamic Language Model based on Individual Word Domains,” *Proceedings of the 18th International Conference on Computational Linguistics* COLING 2000, Saarbrucken, Germany, vol. 2, 2000, pp. 789-794.

Sicilia-Garcia, E. I., J. Ming and F. J. Smith, “Triggering Individual Word Domains in *N*-Gram Language Model,” *Proceedings of the European Conference on Speech Communication and Technology* (EuropeSpeech), vol. 1, 2001, pp. 701-704.

Sicilia-Garcia, E. I., F. J. Smith and P. Hanna, “A Dynamic Language Model based on Individual Word Models,” *Pre-proceedings of the 10th Irish Conference on Artificial Intelligence and Cognitive Science* AICS 99, Cork Ireland, 1999, pp. 222-229.

Siu, M. and M. Ostendorf, “Integrating a Context-Dependent Phrase Grammar in the Variable *N*-Gram framework,” *IEEE ICASSP*, vol. 3, 2000, pp. 1643-1646.

Siu, M. and M. Ostendorf, “Variable *N*-Grams and Extensions for Conversational Speech Language Modelling,” *IEEE Transactions on Speech and Audio Processing,* 8(1) 2000, pp. 63-75.

Smith, F. J. and P. O’ Boyle, “The *N*-Gram Language Model,” *The Cognitive Science of Natural Language Processing Workshop*, Dublin City University, 1992, pp. 51-58.

Zipf, G. K., “Human Behaviour and the Principle of Least Effort,” Reading, MA: Addison-Wesley Publishing Co., 1949.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 35-52 


# Automated Alignment and Extraction of a Bilingual Ontology for Cross-Language Domain-Specific Applications

**Jui-Feng Yeh*, Chung-Hsien Wu*, Ming-Jun Chen* and Liang-Chih Yu***

**Abstract**

This paper presents a novel approach to ontology alignment and domain ontology extraction from two existing knowledge bases: WordNet and HowNet. These two knowledge bases are automatically aligned to construct a bilingual ontology based on the co-occurrence of words in a bilingual parallel corpus. The bilingual ontology achieves greater structural and semantic information coverage from these two complementary knowledge bases. For domain-specific applications, a domain-specific ontology is further extracted from the bilingual ontology using the island-driven algorithm and domain-specific corpus. Finally, domain-dependent terminology and axioms between domain terminology defined in a medical encyclopedia are integrated into the domain-specific ontology. In addition, a metric based on a similarity measure for ontology evaluation is also proposed. For evaluation purposes, experiments were conducted comparing an automatically constructed ontology with a benchmark ontology constructed by ontology engineers or experts. The experimental results show that the constructed bilingual domain-specific ontology mostly coincided with the benchmark ontology. As for application of this approach to the medical domain, the experimental results show that the proposed approach outperformed the synonym expansion approach to web search.

**Keywords:** Ontology, island driven algorithm, cross language application, WordNet, HowNet

* Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
E-mail: {jfyeh, chwu, mjchen, lcyu}@csie.ncku.edu.tw

## 1. INTRODUCTION

In the past few decades, a considerable number of studies been invested focused on developing concept bases for building technology that allows knowledge reuse and sharing. As information exchangeability and communication becomes increasingly global, multilingual lexical resources that provide transnational services are becoming increasingly important. On the other hand, multi-lingual ontologies are very important for natural language processing, such as machine translation (MT), web mining [Oyama *et al.* 2004], and cross-language information retrieval (CLIR). Generally, a multi-lingual ontology maps the keywords of one language to another language, or computes the co-occurrence of the words among languages. A key merit of a multilingual ontology is that it can achieve greater relation and structural information coverage by aligning or merging two or more language-dependent ontologies with different semantic features.

In recent years, significant effort has focused on constructing ontologies manually according to domain experts' knowledge. Manual ontology merging using conventional editing tools without intelligent support is difficult, labor intensive, and error prone. Therefore, several systems and frameworks to help knowledge engineers perform ontology merging have recently been proposed [Noy and Musen 2000]. To avoid reiteration in ontology construction, algorithms for ontology merging [UMLS http://umlsks.nlm.nih.gov] [Langkilde and Knight 1998] and ontology alignment [Vossen and Peters 1997] [Weigard and Hoppenbrouwers 1998] [Asanoma 2001] have been investigated. In these approaches, the final ontology is a merged version of the original ontologies with aligned links between them [Daudé *et al.* 2003]. Alignment is usually performed when ontologies cover domains that are complementary to each other. In the past, a domain ontology was usually constructed manually based on the knowledge or experience of experts or ontology engineers. Recently, automatic and semi-automatic methods have been developed. OntoExtract [Fensel *et al.* 2002] [Missikoff *et al*. 2002] provides an ontology engineering chain for constructing a domain ontology from WordNet and SemCor. Some recent approaches have been discussed in [Euzenat *et al.* 2004]. In [Euzenat *et al.* 2004], the alignment approaches were classified as local or global methods. Four main local methods, that is, the terminological, extensional, semantics, and structure methods, were introduced to measure the correspondence between two ontologies at the local level. Nowadays, much work is being invested in ontology construction for domain applications. Performing authoritative evaluation of ontologies is becoming a critical issue. Some evaluation methods are integrated into ontology tools to detect and prevent mistakes, which might be made in the course of developing taxonomies with frames as described in [Gómez-Pérez 2001]. They defined three main types of mistakes: inconsistency, incompleteness, and redundancy mistakes.

Although the previous research on ontology alignment has achieved much, some

important issues still require further investigation: (1) How can we to construct or extract domain concepts from a corpus? (2) Should the alignment of a cross-language or multilingual ontology be performed automatically or semi-automatically? (3) Authoritative assessment of ontology construction is desirable. In this study, the WordNet and HowNet knowledge bases were aligned to construct a bilingual universal ontology based on the co-occurrence of words in a bilingual parallel corpus. For domain-specific applications, the medical domain ontology was further extracted from the universal ontology using the island-driven algorithm and two corpora, one for the medical domain and another for the contrastive domain. Finally, axioms between medical terminology were derived based on a medical encyclopedia. A benchmark ontology based on the Unified Medical Language System (UMLS) and constructed by ontology engineers and experts was used to evaluate the constructed bilingual ontology. This paper also defines two measures, the taxonomic relation and non-taxonomic relation, as quantitative metrics for evaluating ontologies.

The rest of the paper is organized as follows. Section 2 describes the ontology construction process. Section 3 presents experimental results for the evaluation of our approach. Section 4 gives some concluding remarks.

## 2. Ontology Construction

Figure 1 shows a block diagram of the ontology construction process. There are two major stages in the proposed approach: bilingual ontology alignment and domain ontology extraction.

### 2.1 Bilingual Ontology Alignment

In this approach, a bilingual ontology is constructed by aligning Chinese words in HowNet with their corresponding synsets defined in WordNet according to the co-occurrence of the words in a bilingual parallel corpus. The hierarchical structure of the ontology is actually a conversion of HowNet. One of the important parts of HowNet consists of definitions of lexical entries. In HowNet, each lexical entry is defined as a combination of one or more primary features and a sequence of secondary features. The primary features indicate the entry's category, for example, the relation "is-a" in a hierarchical structure. Based on the entry's category, the secondary features make the entry's sense more explicit, but they are non-taxonomic. Totally, 1,521 primary features are divided into 6 upper categories: Event, Entity, Attribute Value, Quantity, and Quantity Value. These primary features are organized into a hierarchical structure.

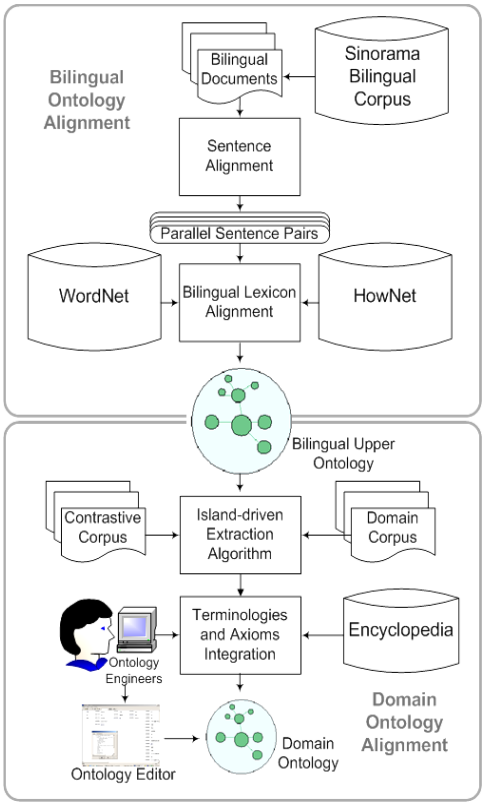



***Figure 1. Ontology construction framework***

In the alignment process, the Sinorama [Sinorama 2001] database, containing over 6,500 documents with 48,000,000 words from 1976 to 2000 in Chinese and English, is adopted as the bilingual parallel corpus. This corpus is then used to compute the conditional probability of the words in WordNet, given the words in HowNet. Then, a bottom up algorithm is used to perform relation mapping. In WordNet, a word may be associated with many synsets, each corresponding to a different sense of the word. To find a relation between two different words, all the synsets associated with each word are considered [Fellbaum 1998]. In HowNet, each word is composed of primary features and secondary features. The primary features indicate the word's category. The goal of this approach is to increase the amount of relation and structural information coverage by aligning their semantic features in WordNet and HowNet.

Equation (1) shows the alignment between the words in HowNet and the synsets in WordNet. Given a Chinese word, $CW_i$, the probability of the word being related to synset, $synset^k$, can be obtained via its corresponding English synonyms, $EW_j^k$, $j = 1, \ldots, m$, which

are the elements in $synset^k$. The probability is estimated as follows:

$$\begin{aligned}\Pr(synset^k \mid CW_i) &= \sum_{j=1}^{m} \Pr(synset^k, EW_j^k \mid CW_i) \\ &= \sum_{j=1}^{m} (\Pr(synset^k \mid EW_j^k, CW_i) \times (\Pr(EW_j^k \mid CW_i)),\end{aligned} \tag{1}$$

where

$$\Pr(synset^k \mid EW_j^k, CW_i) = \frac{N(synset_j^k, EW_j^k, CW_i)}{\sum_l N(synset_j^l, EW_j^k, CW_i)}. \tag{2}$$

In the above equation, $N(synset_j^k, EW_j^k, CW_i)$ represents the number of co-occurrences of $CW_i$, $EW_j^k$, and $synset_j^k$. The probability $\Pr(EW_j^k \mid CW_i)$ is set to one when at least one of the primary features, $PF_i^l(CW_i)$, of the Chinese word $CW_i$ defined in HowNet matches one of the ancestor nodes of synset , $synset_j^k(EW_j)$, except for the root nodes in the hierarchical structures of the noun and verb. Otherwise, the probability $\Pr(EW_j^k \mid CW_i)$ is set to zero:

$$\Pr(EW_j \mid CW_i) = \begin{cases} 1, \ if \ \left(\bigcup_l PF_i^l(CW_i) - \{entity, event, act, play\}\right) \cap \\ \qquad \left(\bigcup ancestor(\bigcup_k synset_j^k(EW_j)) - \{entity, event, act, play\}\right) \neq \varnothing, \\ 0, \quad Otherwise \end{cases} \tag{3}$$

where {entity, event, act, play} is the concept set in the root nodes of HowNet and WordNet, and $\left(\bigcup_l PF_i^l(CW_i) - \{entity, event, act, play\}\right)$ represents all the primary features of the Chinese word $CW_i$ except for {entity, event, act, play}. Finally, the Chinese concept, $CW_i$, is integrated into the synset , $synset_j^k$, in WordNet as long as the probability, $\Pr(synset^k \mid CW_i)$, is not zero. Figure 2(a) shows the concept tree generated by aligning WordNet and HowNet.

## 2.2 Domain ontology extraction

Now, we will attempt to extend the ontology to domain applications. In domain-specific information retrieval, more detailed definitions and terminology are required. This paper proposes a two-stage domain ontology extraction method. This approach extracts the ontology from the cross-language ontology by using the island-driven algorithm in the first stage. The

terminology and axioms defined in a medical encyclopedia are integrated into the domain ontology in the second stage.

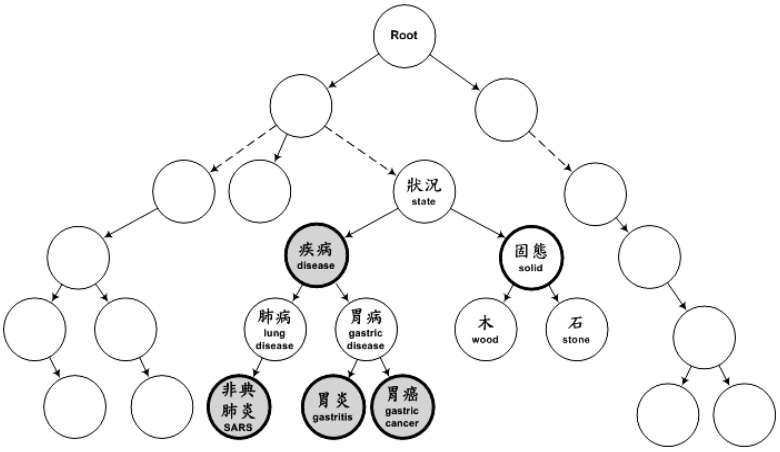



***Figure 2(a). Concept tree generated by aligning WordNet and HowNet. The nodes in bold circles represent operative nodes following concept extraction. The nodes on gray backgrounds represent operative nodes following relationn expansion.***

### 2.2.1 Extraction using the island-driven algorithm

Generally, an ontology provides consistent concepts and world representations necessary for clear communication within the knowledge domain. Even in domain-specific applications, the number of words can be expected to be huge. Synonym pruning is an effective way to perform word sense disambiguation. This paper proposes a corpus-based statistical approach to extracting a domain ontology. The steps are listed as follows:

***Step 1. Linearization:*** In this step, the tree structure in the general purpose ontology shown in Figure 2(a) is decomposed into a vertex list that is an ordered node sequence starting at the root node and ending at the leaf nodes.

***Step 2. Concept extraction from the corpus:*** The node is defined as an operative node when the tf-idf value of word $W_i$ in the domain corpus is higher than that in its corresponding contrastive (out-of-domain) corpus. That is,

$$operative\_node(W_i) = \begin{cases} 1, & if \ \ tf\text{-}idf_{Domain}(W_i) > tf - idf_{Contrastive}(W_i) \\ 0, & Otherwiae \end{cases}, \tag{4}$$

where

$$tf-idf_{Domain}(W_i) = freq_{i,Domain} \times \log \frac{n_{i,Domain} + n_{i,Contrastive}}{n_{i,Domain}},$$

$$tf-idf_{Contrastive}(W_i) = freq_{i,Contrastive} \times \log \frac{n_{i,Domain} + n_{i,Contrastive}}{n_{i,Contrastive}}.$$

In the above equations, $freq_{i,Domain}$ and $freq_{i,Contrastive}$ are the frequencies of word $W_i$ in the domain documents and its contrastive (out-of-domain) documents, respectively; $n_{i,Domain}$ and $n_{i,Contrastive}$ are the numbers of documents containing word $W_i$ in the domain documents and its contrastive documents, respectively. The nodes shown in bold circles in Figure 2(a) represent operative nodes.

***Step 3. Relation expansion using the island-driven algorithm:*** Some domain concepts are no longer operative after the previous steps have been performed due to the problem of data sparseness. According to the analysis performed during ontology construction, most of the inoperative concept nodes have operative hypernym nodes and hyponym nodes. Therefore, the island-driven algorithm is adopted to activate these inoperative concept nodes if their ancestors and descendants are all operative. The nodes shown on gray background in Figure 2(a) are activated operative nodes.

***Step 4. Domain ontology extraction:*** In the final step, the linear vertex list sequence is merged into a hierarchical tree. However, some noisy concepts defined as nodes not belonging to this domain are operative according to Equation (5). For example, the node with the concept "solid" shown in Figure 2(b) is an operative noisy concept. Accordingly, the second goal is to filter out the nodes with operative noisy concepts. In this step, noisy concepts without ancestors or descendants belonging to the domain are removed. Finally, the domain ontology is extracted, and the final result is shown in Figure 2(b).

### 2.2.2 Axiom and terminology integration

In practice, specific domain terminology and axioms should be derived and introduced into an ontology for domain-specific applications. There are two approaches to integrating terminology and axioms into an ontology: the first one is manual editing performed by ontology engineers, and the second is automatic integration from a domain encyclopedia.

For medical domain applications, 1,213 axioms were derived here from a medical encyclopedia with terminology related to diseases, syndromes, and the clinic information. Figure 3 shows an example of an axiom. In this example, the disease "diabetes" is tagged as level "A," which means that this disease occurs frequently. The degrees for the corresponding syndromes indicate the causality between the disease and the syndromes. The axioms also provide two fields, "department of the clinical care" and "the category of the disease," for

medical information retrieval or other medical applications.

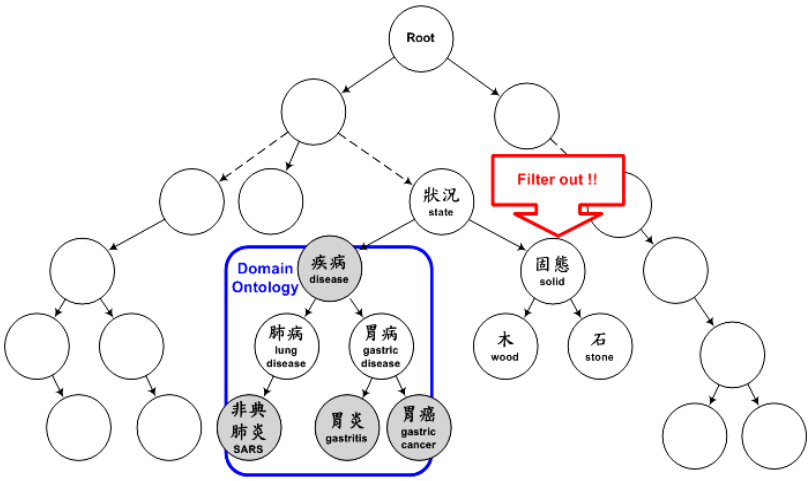



***Figure 2(b). The domain ontology after isolated concepts are filtered out***

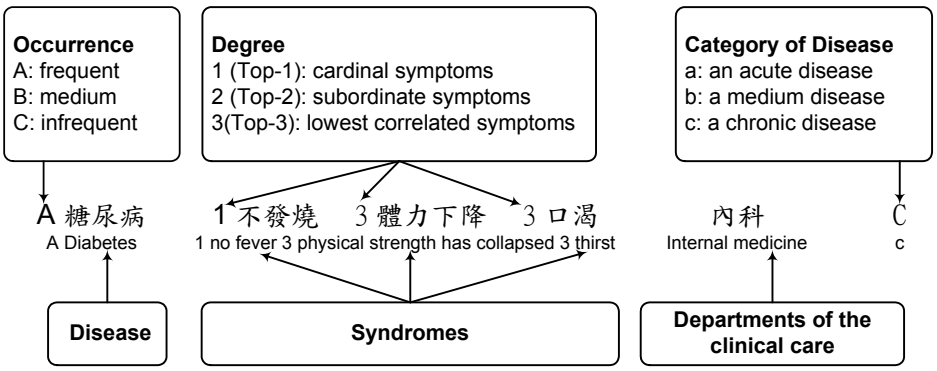



***Figure 3. One example of an axiom***

## 3. Evaluation

For quantitative evaluation of the ontology, two types of evaluation, conceptual evaluation and domain application evaluation, were adopted to evaluate the coincidence between the extracted domain ontology and the manually designed ontology. Furthermore, a medical web mining system was implemented to evaluate the practicability of the bilingual ontology.

## 3.1 Conceptual Evaluation

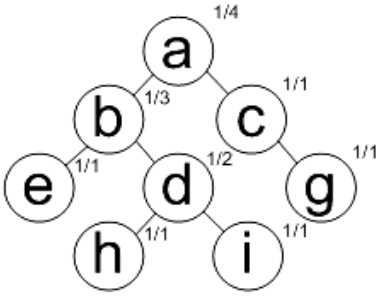
The benchmark ontology was created as a test-suite of reusable data which could be employed by ontology engineers for benchmarking purposes. The benchmark ontology was constructed by domain experts, including two doctors and one pharmacologist, based on the Unified Medical Language System (UMLS). The domain experts integrated the Chinese concepts without changing the contents of UMLS.

The construction of an ontology is generally evaluated using a two-layer measure, consisting of lexical and conceptual layers [Eichmann *et al.* 1998]. Evaluation in the conceptual layer seems to be more important than that in the lexical layer when the ontology is constructed by aligning or merging several well-defined source ontologies. There are two conceptual relation types of evaluation: taxonomic and non-taxonomic evaluation.

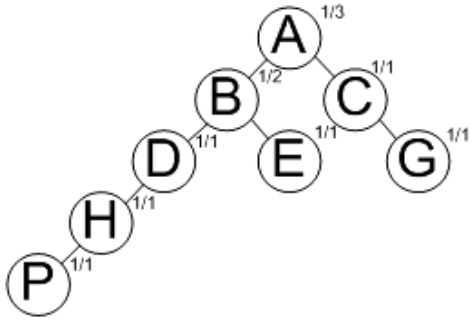
### 3.1.1 Evaluation of taxonomic relations

Evaluation of taxonomic relations is based not only on lexical similarity but also on hierarchical information according to the basic ontology definition. In this approach, obtaining the metric is a five-step process.

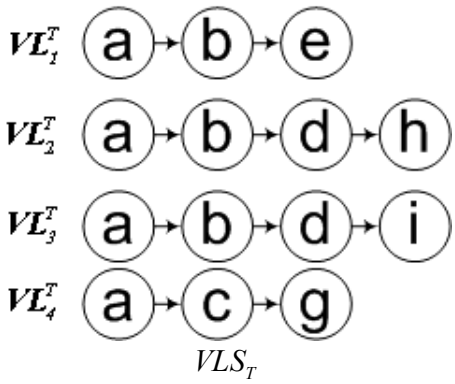
***Step1. Linearization:*** In this step, the tree structure is decomposed into a vertex list as described in Section 2.2. The ontology, $O_T$, and the benchmark, $O_B$, are shown in Figures 4(a) and 4(b), respectively. After linearization is performed, the vertex list sets $VLS_T$ and $VLS_B$ are obtained as shown in Figure 4(c) and Figure 4(d), where $VLS_T = \{VL_1^T, ....., VL_p^T\}$; $VLS_B = \{VL_1^B, ....., VL_q^B\}$; $VL_i^O$ represents the i-th vertex list of ontology *O*, and *p* and *q* are the numbers of vertex lists for the target ontology and the benchmark ontology, respectively.



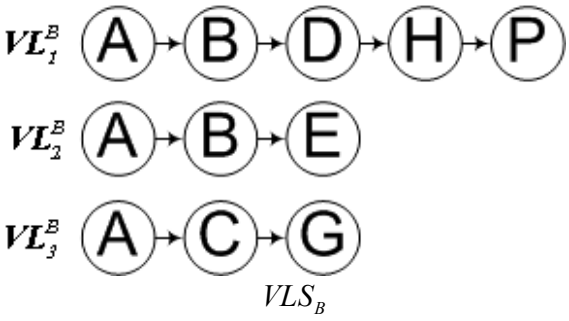

(a)The taxonomic hierarchical representation of target ontology $O_T$




(b) The taxonomic hierarchical representation of benchmark ontology $O_B$




(c) The taxonomic vertex list representation of the target ontology




(d) The taxonomic vertex list representation of the benchmark ontology

***Figure 4. Linearization of the target and benchmark ontologies***

***Step 2. Normalization:*** Since the frequencies of concepts in the vertex lists are not identical, normalization factors are introduced. For the target ontology, the set of factor vectors adopted for normalization is $NF^T = \left\{nf_1^T, nf_2^T, nf_3^T, nf_4^T, nf_5^T, ....., nf_m^T\right\}$, and for the benchmark ontology it is $NF^B = \left\{nf_1^B, nf_2^B, nf_3^B, nf_4^B, ....., nf_n^B\right\}$, where $nf_i^o$ is the normalization factor for the i-th concept of ontology *O*. It is defined as the reciprocal of the number of vertex lists:

$$nf_i^O = \frac{1}{NV_i^O}, \tag{5}$$

where $NV_i^O$ represents the number of vertex lists containing concept *i* in ontology *O*.

***Step 3. Similarity estimation of two vertex lists:*** As the Figure 5 shows, the pairwise similarity of two vertex lists for the target ontology and benchmark ontology can be obtained using the Needleman/Wunsch techniques as described in the following steps:

***1. Initialization:*** Create a matrix with *m*+1 columns and *n*+1 rows, where *m* and *n* are the numbers of nodes in the vertex lists of the target ontology and benchmark ontology, respectively. The first row and first column of the matrix can both be initially set to 0. That is,

$$Sim(m,n) = 0, \ if \ \ \mathrm{m} = 0 \ \ or \ \ \mathrm{n} = 0. \tag{6}$$

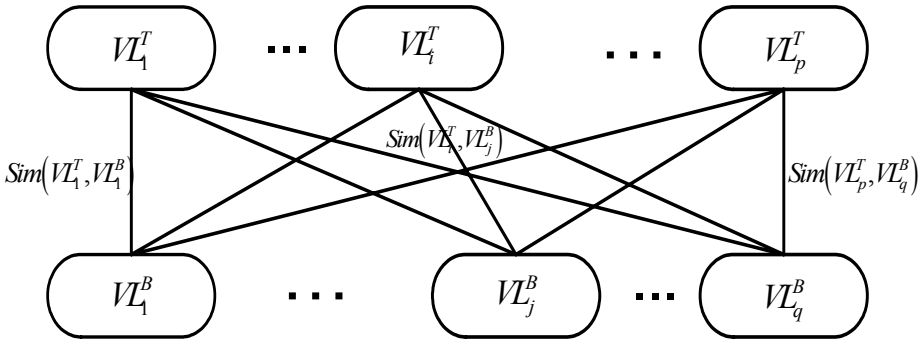



***Figure 5. Pairwise similarity between the target ontology and benchmark ontology***

***2. Matrix filling:*** Assign values to the remaining elements in the matrix according to the following equation:

$$Sim(V_m^{T_i}, V_n^{B_j}) = \max \begin{cases} Sim(m-1, n-1) + \frac{1}{2}\left(nf_{m-1}^{T_i} + nf_{n-1}^{B_j}\right) \times Sim_{lexicon}\left(V_{m-1}^{T_i}, V_{n-1}^{B_j}\right), \\ Sim(m-1, n) + \frac{1}{2}\left(nf_{m-1}^{T_i} + nf_n^{B_j}\right) \times Sim_{lexicon}\left(V_{m-1}^{T_i}, V_n^{B_j}\right), \\ Sim(m, n-1) + \frac{1}{2}\left(nf_m^{T_i} + nf_{n-1}^{B_j}\right) \times Sim_{lexicon}\left(V_m^{T_i}, V_{n-1}^{B_j}\right). \end{cases} \quad (7)$$

There are some synonyms belonging to the same concept in one vertex. Thus, the lexical similarity can be defined as

$$Sim_{lexicon}\left(V_{m-1}^{T_i}, V_n^{B_j}\right) = \frac{\left|\text{Synonyms defined in } V_{m-1}^{T_i} \text{ and } V_n^{B_j}\right|}{\left|\text{Synonyms defined in } V_{m-1}^{T_i} \text{ or } V_n^{B_j}\right|}. \quad (8)$$

***3. Traceback:*** Determine the actual alignment with the maximum score, $Sim(V_m^{T_i}, V_n^{B_j})$; therefore, the pairwise similarity is defined as follows:

$$Sim\left(VL_i^T, VL_j^B\right) \equiv \max_{m.n} Sim\left(V_m^{T_i}, V_n^{B_j}\right). \quad (9)$$

***Step 4. Pairwise similarity matrix estimation:*** The pairwise similarity matrix is obtained after $p \times q$ iterations using the vertex list similarity defined in Step3. $p$ and $q$ are the numbers of vertex lists for the target ontology and benchmark ontology, respectively. Each element of the pairwise similarity matrix in Equation (10) is obtained from Equation (9):

$$PSM(O_T, O_B) \equiv \begin{bmatrix} Sim(VL_1^T, VL_1^B) & \cdots & Sim(VL_1^T, VL_q^B) \\ \vdots & \ddots & \vdots \\ Sim(VL_p^T, VL_1^B) & \cdots & Sim(VL_p^T, VL_q^B) \end{bmatrix}_{p \times q}. \quad (10)$$

***Step 5. Evaluation of the taxonomic hierarchy:*** The total similarity between the target ontology and benchmark ontology, defined as the average similarity of all the vertex lists, is estimated as follows:

$$Sim_{taxonomic}(O_T, O_B) = \frac{1}{p} \sum_{i=1}^{p} \max_{1 \le j \le q} \left\{ Sim(VL_i^T, VL_j^B) \right\} \quad (11)$$

### 3.1.2 Evaluation of non-taxonomic relations

Some relations defined in the ontology are non-taxonomic such as synonyms. In fact, lexical similarity is applied to measure the conceptual similarity. Lexical similarity is computed using the following equation:

$$Sim_{lexicon}\left(V_s^{T_i}, V_t^{B_j}\right) = \frac{\left|\text{Words defined in } V_s^{T_i} \text{ and } V_t^{B_j}\right|}{\left|\text{Words defined in } V_s^{T_i} \text{ or } V_t^{B_j}\right|}. \quad (12)$$

Therefore, evaluation of all of the whole non-taxonomic relations is performed according to the following equation:

$$Sim_{non-taxonomic}(O_T, O_B) = \frac{1}{p \times q} \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{s} \sum_{t} Sim_{lexicon}\left(V_s^{T_i}, V_t^{B_j}\right). \quad (13)$$

### 3.1.3 Evaluation results

Using the benchmark ontology and evaluation metrics described in the previous sections, we obtained the evaluation results shown in Table 1. The matching ratios between the constructed ontology and benchmark ontology were 57% and 68% for taxonomic and non-taxonomic relations, respectively. From the experimental results, the following phenomena were discovered: first, the number of words mapped to the same concept in the upper layer of the ontology was larger than that in the lower layer because the terminology usually appeared in the lower layer. Owing to the lack of an authoritative benchmark, the metrics could not provide an ideal measure. The main weakness was the difference between the target and benchmark ontologies, especially the terminology used. Introducing concept or word frequency measures may lead to a significant improvement.

***Table 1. Matching ratio between the target ontology and benchmark ontology***

| | |
|---|---|
| **Taxonomic relation matching ratio** | ***57%*** |
| **Non-Taxonomic relation matching ratio** | ***68%*** |

## 3.2 Evaluation of domain application

To assess the performance of the ontology, a cross-language medical domain web-mining system was implemented. For domain concept extraction, a corpus was collected from several websites. A total of 2,322 web pages were collected as a medical domain corpus, and 8,133 web pages as a contrastive domain corpus. Besides the training corpus, 1,212 web pages different from the training sets and the test queries were also collected for the purpose of system evaluation. Forty users, who did not take part in system development, were asked to provide a set of queries given the collected web pages. After post-processing was performed, the duplicate queries and the queries that were out of the medical domain were removed. Finally, 3,207 test queries using natural language were obtained.

The baseline system is based on the Vector-Space Model (VSM). That is, a sequence of words is treated as a bag of words regardless of the word order. For a word sequence from a user's input, $q = \{q_1, q_2, \cdots\cdots q_n\}$, and a word sequence in a web page, $d = \{d_1, d_2, \cdots d_n\}$, the similarity is defined as the cosine function as follows:

$$Sim_{VSM}(D_i, q) = \cos(d, q) = \frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} d_i^2} \times \sqrt{\sum_{i=1}^{n} q_i^2}}, \tag{14}$$

where $D_i$ is the i-th document in the web page and $q$ is the user's query. This approach to key term expansion based on a synonym set is also adopted in the baseline system.

The conceptual relations and axioms defined in the medical ontology were integrated into the baseline as the ontology-based system. The medical web search engine was developed based on the constructed medical domain ontology consists of a relation inference module and axiom inference module. The functions of and techniques used with these modules are described in the following.

### 3.2.1 Relation inference module

For semantic representation, traditionally, keyword-based systems face two problems. First, ambiguity usually results from the polysemy of words. The domain ontology gives clear descriptions of the concepts. In addition, not all of the synonyms of a word should be expanded without any constraints being applied. Secondly, the relations between the concepts should be expanded and weighted in order to include more semantic information for semantic inferences. We treat each user's input and the content of a web page as a sequence of words.

The similarity between an input query and a web page is defined as the similarity between the two bags of words based on key concepts in the ontology [Yeh *et al.* 2004].

$$Sim_{relation}(D_i, q) = Sim_{relation}(d, q) = Sim_{relation}(d_1, d_2, \ldots, d_L, q_1, q_2, \ldots q_k)$$
$$= \begin{cases} 1 & d_l \text{ and } q_k \text{ are identical} \\ \sum_{k=1}^{K}\sum_{l=1}^{L}\left(\frac{1}{2^r}\right) & d_l \text{ and } q_k \text{ are hypernyms and } r \text{ is the number of levels in between} \\ \sum_{k=1}^{K}\sum_{l=1}^{L}\left(1-\frac{1}{2^t}\right)^2 & d_l \text{ and } q_k \text{ are synonyms and } t \text{ is the number of their common concepts} \\ 0 & \text{Other} \end{cases} \quad (15)$$

### 3.2.2 Axiom inference module

Some axioms, such as "result in" and "result from," that are expected to affect the performance of a web search system in a medical domain are defined in order to describe the relationships between syndromes and diseases. We collected data about syndromes and diseases from a medical encyclopedia and tagged the diseases with three levels according to their frequency of occurrence and tagged syndromes with four levels according to their significance with respect to a specific disease. The "result in" relation score is defined as $RI(D_i, q)$ if a disease occurs in the input query and its corresponding syndromes appear in the web page. Similarly, if a syndrome occurs in the input query and its corresponding disease appears in the web page, the "result from" relation score is defined as $RF(D_i, q)$. The relation score is estimated as described in [Yeh *et al.* 2004]:

$$\begin{aligned} Axiom(D_i, q) &= \max\{RI(D_i, q), RF(D_i, q)\} \\ &= \max\{RI(d_1, d_2, \ldots, d_p, q_1, q_2, \ldots, q_R), RF(d_1, d_2, \ldots, d_p, q_1, q_2, \ldots q_R)\} \\ &= \max\left\{\sum_{p=1,r=1}^{P,R} a_{pr}^{RI}, \sum_{p=1,r=1}^{P,R} a_{pr}^{RF}\right\}, \end{aligned} \quad (16)$$

where $a_{pr}^{RI} = 1/2^{n-1}$ if disease $d_p$ results in syndrome $q_r$ and $q_r$ is the top-n feature of $d_p$. Similarly, $a_{pr}^{RF} = 1/2^{n-1}$ if syndrome $d_p$ results from disease $q_r$ and $d_p$ is the top-n feature of $q_r$. The similarity between the i-th web page and query $q$ is defined as

$$Sim_{axiom}(D_i, q) = \frac{Axiom(D_i, q)}{\sum_i Axiom(D_i, q)}. \quad (17)$$

### 3.2.3 Weight determination using the 11-avgP score

The medical domain web search system is modelled using a linear combination of a relational inference model and axiom inference model. The normalized weight factor, $\alpha$, is employed for the purpose of concept expansion as follows:

$$Sim(D_i, q) = (1-\alpha)Sim_{relation}(D_i, q) + \alpha \times Sim_{axiom}(D_i, q). \quad (18)$$

An experiment was conducted to evaluate the estimation of the combination weights for each model. The results are shown in Figure 6. A performance measure called 11-AvgP [Eichmann and Srinivasan 1998] was used to summarize the precision and recall rates. The best 11-AvgP score was obtained when the weight $\alpha$ was set to 0.428.

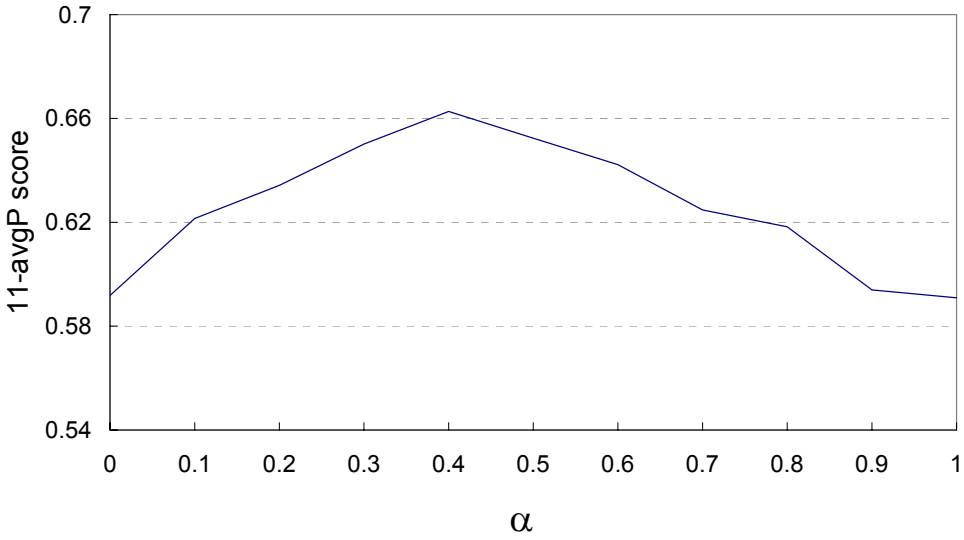



***Figure 6. The 11-avgP score with different values of α***

### 3.2.4 Evaluation of different inference modules

In the following experiments, web pages were separately evaluated by focusing on one inference module based on the domain-specific ontology at a time. That is, the mixture weight was set to 1 for one inference module, and the other weight was set to 0 in each evaluation. For comparison purposes, the keyword-based VSM approach and the ontology-based system were also evaluated, and the results are shown in Figure 7. The precision and recall rates were used as the evaluation measures. The ontology-based approach combines of concept inferences and axiom inferences as described in the previous sections. The results shown in Table 2 reveal that the ontology-based system outperformed the baseline system in synonym

expansion. Instead of keywords, the concepts defined in the ontology play an important role in term expansion for a specific domain. In addition, relation axioms are important and can be effectively used in domain applications; that is to say, the inference axioms provide semantic relationships between words.

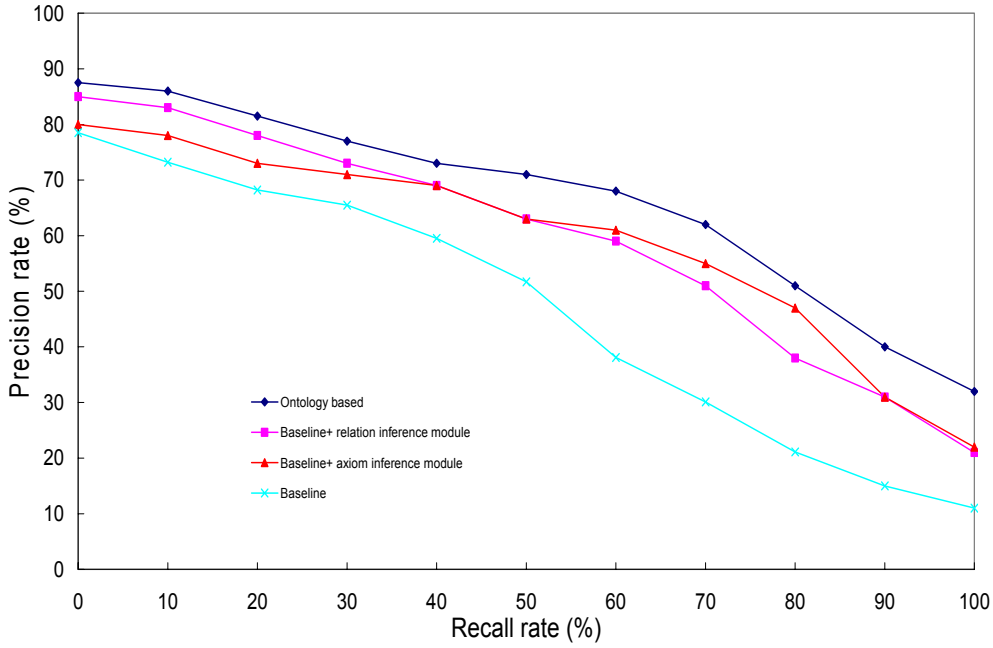



***Figure 7. The precision rates and recall rates achieved with the proposed method and the baseline system***

***Table 2. Precision rates (%) at the 11-point recall level***

| Recall Level | 0 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline system (Precision) | 78 | 73 | 68 | 65 | 60 | 52 | 38 | 30 | 21 | 15 | 11 |
| Ontology based (Precision) | 87 | 86 | 82 | 77 | 73 | 71 | 68 | 62 | 51 | 40 | 32 |

## 4. CONCLUSIONS

A novel approach to automated ontology alignment and domain ontology extraction from two knowledge bases has been presented in this paper. In this study, a bilingual ontology has been developed from two well established knowledge bases, WordNet and HowNet, based on the co-occurrence of words in a parallel bilingual corpus. A domain-dependent ontology has been further extracted from the universal ontology using the island-driven algorithm and a domain corpus as well as a contrastive corpus. In addition, domain-specific terms and axioms have also been added to the domain ontology. A metric based on the similarity measure for ontology evaluation has also been proposed. The experimental results show that the proposed approach can extract an aligned bilingual domain-specific ontology which mostly coincides with a corresponding manually designed ontology. We have also applied the obtained domain-specific ontology to web page search in a medical domain. The experimental results show that the proposed approach outperformed the synonym expansion approach.

## REFERENCES

Asanoma, N., "Alignment of Ontologies: WordNet and Goi-Taikei," *Proc. of WordNet and Other Lexical Resources Workshop Program,* NAACL2001, 2001, pp. 89-94

Daudé, J., L. Padró. and G. Rigau, "Validation and Tuning of Wordnet Mapping Techniques," *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*. Borovets, Bulgaria, 2003.

Eichmann, D., M. Ruiz and P. Srinivasan, "Cross-language information retrieval with the UMLS Metathesaurus," *Proc. of ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1998), 1998, pp. 72-80.

Euzenat, J., T. Le Bach, J. Barrasa, P. Bouquet, J. De Bo, R. Dieng, M. Ehrig, M. Hauswirth, M. Jarrar, R. Lara, D. Maynard, A. Napoli, G. Stamou, H. Stuckenschmidt, P. Shvaiko, S. Tessaris, S. Van Acker and I. Zaihrayeu, State of the Art on Ontology Alignment. Knowledge Web Deliverable D2.2.3, INRIA, Saint Ismier, 2004.

Fensel, D., C. Bussler, Y. Ding, V. Kartseva, M. Klein, M. Korotkiy, B. Omelayenko and R. Siebes, "Semantic Web Application Areas," *Proc. of the 7th International Workshop on Applications of Natural Language to Information Systems*, Stockholm - Sweden, June 27--28, 2002.

Fellbaum, F. C., "WordNet an electronic Lexical Database," The MIT Press, 1998. pp. 307-308

Gómez-Pérez, A., "Evaluating ontologies: Cases of Study," *IEEE Intelligent Systems and their Applications: Special Issue on Verification and Validation of ontologies*. vol. 16, no 3. March 2001, pp. 391-409.

HowNet, http://www.keenage.com/

Langkilde, I. and K. Knight, "Generation that Exploits Corpus-Based Statistical Knowledge," *Proc. of COLING-ACL* 1998, pp. 704-710

Levensthein, V., "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics–Doklady*, vol.10, no. 8, 1966, pp.707–710.

Missikoff, M., R. Navigli and P. Velardi, "An Integrated Approach for Web Ontology Learning and Engineering," *IEEE Computer*, November 2002, pp. 60-63.

Noy, N. F. and M. Musen, "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment," *Proc. of the National Conference on Artificial Intelligence*. AAAI2000. 2000, pp.450-455

Oyama, S., T. Kokubo and T. Ishida, "Domain-Specific Web Search with Keyword Spice," *IEEE Transactions on Knowledge and Data Engineering,*vol 16, nO. 1, 2004, pp. 17-27.

Sinorama Magazine and Wordpedia.com Co.. Multimedia CD-ROMs of Sinorama from 1976 to 2000, Taipei, 2001.

UMLS, http://www.nlm.nih.gov/research/umls/

Vossen, P. and W. Peters, "Multilingual design of EuroWordNet," *Proc. of the Delos workshop on Cross-language Information Retrieval.* 1997.

Weigard, H. and S. Hoppenbrouwers, "Experiences with a multilingual ontology-based lexicon for news filtering," *Proc. of the 9th International Workshop on Database and Expert Systems Applications*. 1998, pp.160-165.

WordNet, http://www.cogsci.princeton.edu/~wn/

Yeh, J.F., C.H. Wu, M.J. Chen and L.C. Yu, "Automated Alignment and Extraction of Bilingual Domain Ontology for Medical Domain Web Search," *Proc. Of the 3rd SIGHAN Workshop on Chinese Language Learning,* ACL2004, Barcelona, 2004, pp.65-71

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 53-94 


# Chinese Main Verb Identification: From Specification to Realization[1]

**Bing-Gong Ding[*], Chang-Ning Huang[+] and De-Gen Huang[*]**

**Abstract**

Main verb identification is the task of automatically identifying the predicate-verb in a sentence. It is useful for many applications in Chinese Natural Language Processing. Although most studies have focused on the model used to identify the main verb, the definition of the main verb should not be overlooked. In our specification design, we have found many complicated issues that still need to be resolved since they haven't been well discussed in previous works. Thus, the first novel aspect of our work is that we carefully design a specification for annotating the main verb and investigate various complicated cases. We hope this discussion will help to uncover the difficulties involved in this problem. Secondly, we present an approach to realizing main verb identification based on the use of chunk information, which leads to better results than the approach based on part-of-speech. Finally, based on careful observation of the studied corpus, we propose new local and contextual features for main verb identification. According to our specification, we annotate a corpus and then use a Support Vector Machine (SVM) to integrate all the features we propose. Our model, which was trained on our annotated corpus, achieved a promising F score of 92.8%. Furthermore, we show that main verb identification can improve the performance of the Chinese Sentence Breaker, one of the applications of main verb identification, by 2.4%.

**Keywords:** Chinese Main Verb Identification, Text Analysis, Natural Language Processing, SVM

---

[1] The work was done while the author was visiting Microsoft Research Asia.

[*] Department of Computer Science, Dalian University of Technology, 116023, China
Email: dbg_dlut@hotmail.com, dlhuangdg@263.net

[+] Microsoft Research Asia, Beijing, 100080, China
Email: cnhuang@msrchina.research.microsoft.com

## 1. Introduction

The main verb is the verb corresponding to the main predicate-verb in a sentence. Our task is to identify the main verb of the sentence, which is a critical problem in natural language processing areas. It is a prerequisite for diverse applications such as dependency parsing [Zhou 1999], sentence pattern identification [Luo 1995], Chinese sentence breaker, and so on.

Unlike western languages, Chinese grammar has little inflection information. Chinese verbs appear in the same form no matter whether they are used as nouns, adjectives, or adverbs. Below are some examples[2].

***Example* 1**

他 /r(ta1) 深 /d(shen1) 得 /v(de2) 学 生 /n(xue2sheng1) 的 /u(de) **喜 爱 /vn(xi3ai4)** 。/ww

*(He is deeply **loved** by his students.)*

***Example* 2**

乡 镇 企 业 /n(xiang1zhen4qi3ye4) 都 /d(dou1) 很 /d(hen3) **盛 行 /v(sheng5xing2)** 。/ww

*(The Township Enterprises are very **popular**.)*

***Example* 3**

毫 不 /d(hao2bu4) **放 松 /v(fang4song1)** 地 /u(de) 继 续 /v(ji4xu4) 推 进 /v(tui1jin4) 党风/n(dang3feng1) 廉政/n(lian2zheng4) 建设/vn(jian4she4) ，/ww

*(Never **relaxedly** advance the cultivation of party conduct and construction of a clean government.)*

In the Example 1 sentence, the word in bold, “喜爱” (love), is a verbal noun. In the Example 2 sentence, “盛行” (popular) is modified by “很” (very), so it functions as an adjective. In the Example 3 sentence, “放松” (relax) is followed by “地” (de)[3], so “放松” functions as an adverbial. Thus, if one wants to identify the main verb in a Chinese sentence, one faces a more

[2] If not specially pointed out, the following examples come from the PK corpus, which was released by the Institute of Computational Linguistics, Peking University, and is available at http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp. The corpus contains one month of data from *People's Daily* (January 1998). It has been both word segmented and part-of-speech tagged. “/r”, “/v” etc. are the part-of-speech tags. “/ww” denotes the end of the sentences.

[3] In Chinese, “地”(de) is used after an adjective or phrase to form an adverbial adjunct before the verb.

difficult task than in an English sentence since one cannot use morphological forms as clues. The second characteristic of Chinese verbs is that they have no specific syntactic function. Verbs can be used as subjects, predicates, objects, adverbial modifiers, or complements. So there are occasions when verbs are used consecutively. This can be shown by the following examples.

***Example* 4**

**转移/v(zhuan3yi2)** 不/d(bu4) **等于/v(deng3yu2)** **压缩/v(ya1suo1)** 。/ww

*(Shifting does not mean compressing.)*

***Example* 5**

大多数/m(da4duo1shu4) 人/n(ren2) 更/d(geng4) **反对/v(fan3dui4)** **提前/v(ti2qian2)** **举行/v(ju3xing2)** **大选/v**[4]**(da4xuan3)** 。/ww

*(Most people were opposed to holding the election ahead of time.)*

***Example* 6**

**恶化/v(e4hua4)** 的/u(de) 经济/n(jing1ji4) **得到/v(de2dao4)** **改善/v(gai3shan4)** 。/ww

*(The deteriorated economy was improved.)*

In the above three sentences, the verb "转移" (shift) is used as the subject. Verbs like "等于" (mean), "反对" (oppose), and "得到" (get) are used as predicate-verbs. "压缩" (compress), "提前" (ahead of time), "举行" (hold), "大选" (election), and "改善" (improve) are used as objects. "恶化" (deteriorate) is used as an adjective modifier. Note that in Example 5, four verbs are used consecutively.

Therefore, the essence of the main verb identification problem is to identify the main verb among several verbs in a sentence that have no inflections at all, which is determined by the characteristics of Chinese grammar.

Although the lack of regular morphological tense markers renders main verb identification complicated, finding the main verb cannot be bypassed since it plays a central role in Chinese grammar [Lv 1980]. For example, suppose one is building a sentence pattern identification system. There are several reasons why we should identify the main verb first.

---

[4] In our corpus annotation, we tend to follow the annotation of the Peking Corpus and try to set aside part-of-speech annotation, which still needs discussion among researchers. For example, some researchers may argue that "大选(da4xuan3)" should be annotated as a noun. Since the original annotation of "大选" in the Peking corpus is "/v", we have not revised its part-of-speech tag to "/n".

- It has been shown that most sentences have verbs as predicates. So once the verb-predicate sentence pattern has been analyzed, almost all the sentence patterns can be analyzed [Lv *et al.* 1999]. Our investigation on the annotated corpus also produced the same results for Wu's assertion as reported in [Lv *et al*. 1999] (see section 3.3 for the reference).
- A sentence pattern identification system generally needs to identify the subject, object, adverbial modifier, and complement. All these syntactic parts are related to the main verb [Luo 1995].
- Many sentence patterns are embodied by a set of verbs. By identifying these main verbs first, we can classify the sentence patterns. For example, in the pivotal sentence, the main verbs tend to be "使" (shi3, have (sb. to do sth.)), "让" (rang4, let), "叫" (jiao4, ask), "请" (qing3, invite), "派" (pai4, send) etc. Another example is a sentence that has a clause as its object; in this case verbs such as "觉得" (jue4de2, feel), "希望" (xi1wang4, hope), "认为" (ren4wei2, think), "是" (shi4, be) are more likely to be main verbs.

The points mentioned above are particularly related to Chinese sentence pattern identification, but analogous arguments can easily be made for other applications. See for example, the discussion in [Zhou 1999] about the subject-verb and object-verb dependency relations and section 6 with regard to the Chinese Sentence Breaker.

Recently researchers have arrived at a consensus that large annotated corpora are useful for applying machine learning approaches to solve different NLP problems. When constructing a large corpus, such as the Penn TreeBank [Xia *et al.* 2000; Marcus *et al.* 1993] or Chinese chunking [Li *et al.* 2004], the design of the specification is the basis part of the work. With this idea in mind, we propose the use of main verb specification to cover the various linguistic phenomena and provide a mechanism to ensure that the inter-annotator consistency is as high as possible. The second motivation of our new specification is as follows: The definition problem involved in automatically identifying the main verb from the computational point of view has not been tackled in detail. To our knowledge, only Luo [1995] has studied a relatively simple definition. Since there has not been sufficient discussion of the specification of main verb, it is difficult to push the research of main verbs forward. Finally, while we were designing our specification, we found that there exist different complicated cases with respect to main verb definition (see section 3.2.3 for details). Thus, the first step in our work was to devolop a more clear definition of a main verb and tries to investigate its ambiguities. This was the real foundation of our work.

Previous studies focused on exploring different statistical and heuristic features in order to identify predicates. Heuristic rules [Luo 1995] and statistical methods like the Decision Tree [Sui and Yu 1998b] have been used to identify predicates. But they either use one of the methods or just use them separately [Gong *et al.* 2003]. We believe it is better to combine the

heuristic and statistical features together. In this paper, we treat the Main Verb Identification (MVI) task as a binary classification problem of determining whether the VP is the MVP or not. We define the main VP (MVP) as the VP chunk in which the head word is the main verb. Here, a verb chunk, VP, is composed of a head verb with its pre-modifier or the following verb-particles, which form a morphologically derived word sequence [Li *et al.* 2004]. The head word of the VP is the verb that dominates the VP. For example, if the main verb is "返回" (fan3hui2, return), then the chunk "连忙/d 返回/v" (lian2mang2 fan3hui2, immediately return) is the MVP, in which the head verb is "返回". We can have a one-to-one mapping between main verbs and MVPs. Therefore, identifying the main verb is equal to identifying the MVP with additional available chunking information. So in the following, "MVP Identification" and "Main Verb Identification" are interchangeable.

We employ one of the most successful machine learning approaches, the Support Vector Machine (SVM), as the classifier. Our method combines lexicalized knowledge with statistical information. We evaluated the performance of our MVI system on the PK corpus, which is an annotated test set. The MVP recall and precision rates reached 92.1% and 93.6% respectively. The main aspects of our research are as follows:

- We investigated in detail the distribution of simple sentence structure and main verbs. After that, we tried to develop our specification and conducted a pilot study on the complicated aspects of the main verb definition.
- Because shallow parsing provides useful information such as chunks and chunk type information, we propose conducting MVI on the results of chunking [Li *et al.* 2004]. Our experiments show the MVI performance based on chunking is better than that of part-of-speech.
- We propose new features based on careful observations of the training corpus. The features are divided into two categories, local and contextual features. Among them, VP position, VP length, Probability of head verbs being MVPs, and Anti-patterns are all new features that we propose. Although they are simple, they work well in MVI.

The rest of the paper is structured as follows. Section 2 presents related works. Section 3 describes in detail the specification of the main verb and how the MVI is handled by our approach. Section 4 gives experimental results. Section 5 presents error analysis and discussion. Section 6 presents an application of main verb identification, the Chinese Sentence Breaker. Finally, we draw conclusions and suggest future work in section 7.

## 2. Related Works

The Chinese language is a typical SVO sequence language, in which 'V' is the main verb in the sentence [Jin and Bai 2003]. The problem of main verb identification has been studied

extensively by Chinese linguists for a long while [Lv 1980; Ding *et al.* 1961; Zhang 1982; Huang 1987; Fan 1995; Liu *et al.* 2002]. Since the definition of the main verb is related to different verb-predicate sentence patterns, linguists usually describe the different kinds of main verbs in the context of verb-predicate sentence patterns.

[Liu *et al.* 2002] divided verb-predicate sentences into five types: predicates that (1) take no object, (2) take a single object, (3) take double objects, (4) include an adverbial modifier, or (5) include complements. Based on this classification scheme, seven specific verb-predicate sentence patterns were also proposed and discussed individually, including "是" (shi4, to be) sentences, "有" (you3, have) sentences, series-verb sentences, pivotal sentences, existential sentences, "把" (ba3) sentences, and "被" (bei4, passive voice) sentences[5].

[Fan 1995] introduced a verb-predicate sentence pattern framework that includes seven subdivided sentence patterns, which overlap with Liu's classification. For example, SV (Subject-Verb), SVO (Subject-Verb-Object), SZV (Subject-Adverbial -Verb), and SVB (Subject-Verb-Complement) patterns in Fan's framework are similar to (1), (2), (4), and (5) in Liu's work. Other sentence patterns include SVL (Subject-Coordination), SCT (Subject-Series-Verb), and SVD (Subject-Duplicate-Verb). Detailed information can be found in his book. The reader should be aware that an SVL like "他一边走一边说" (He talks while walking) or "我们爱祖国爱人民" (We love both our motherland and our people) is equivalent to a series-verb instead of a sentence with verb-coordination in our definition (see section 3.1 for details). Fan's and Liu's works differ in that Fan tries to incorporate more sentence patterns into a single framework. For example, Fan further subdivides SZV into eight specific verb-predicate sentence patterns, like "被" (bei4, passive voice) sentences, "使" (shi3, let) sentences, "从" (cong2, from) sentences, etc. Fan also further subdivides SVB into seven constructions, like the "verb-resultative construction," "verb-得 construction," etc.

A particular feature of Huang's work [1987] is the examples he provides from real texts. The sentence patterns listed in his work are similar to those in [Liu *et al.* 2002].

[Zhang 1982] divided verb-predicate sentences into eleven types: verb sentences, verb-object phrase sentences, verb-compliment phrase sentences, modifier-verb phrase sentences, series-verb phrase sentences, pivot-verb phrase sentences, series-verb combined with pivot-verb phrase sentences, "把" sentences, "被" sentences, the negative form of verb-predicate sentences and the interrogative form of verb-predicate sentences. Similar to [Liu *et al.* 2002] and [Fan 1995], Zhang regards the adverbial-modifier as the basis for subdividing the verb-predicate sentence pattern. However, the author in [Lv 1980] did not use this kind of basis for classification. In addition, unlike [Lv 1980] and [Liu *et al.* 2002], Zhang

[5] "是"(shi4, to be), "有"(you3, have), "把"(ba3, ba), and "被"(bei4, passive voice) sentences are Chinese sentences which contain the above words.

uses the whole verb-object phrase, the verb-complement phrase, and the modifier-verb phrase to subdivide the verb-predicate. However, in our specification, longer phrases or whole phrases, such as whole verb-object phrases, are recursively defined. This categorization scheme cannot be used to subdivide our verb-predicate sentence pattern since a shallow parser cannot provide such information.

The findings in [Lv 1980] were the earliest and most widely ones accepted by other linguists. According to the different sentence structures, the author in [Lv 1980] introduced 13 types of verb-predicate sentence patterns. See Table 1.

***Table 1. Verb-predicate sentence patterns in [Lv]***

| |
|---|
| 1. Transitive Verb Sentence |
| 2. Intransitive Verb Sentence |
| 3. Double Object Sentence |
| 4. A sentence whose object is a verb |
| 5. A sentence whose object is a clause |
| 6. A sentence whose object is number |
| 7. A sentence whose object is placed before the predicate |
| 8. “把 (ba3)” Sentence |
| 9. Passive Voice Sentence |
| 10. Complement Sentence |
| 11. Existential Sentence |
| 12. Series Verb Sentence |
| 13. Pivotal Sentence |

Transitive Verb Sentence

| | Subject | Adverbial modifier | **Verb** | Accusative Object | Non Accusative Ojbect | Auxiliary |
|---|---|---|---|---|---|---|
| A | 你<br>她<br>你<br>她 | 从前<br><br><br>最近 | **学过**<br>**唱过**<br>**会写**<br>**吃** | 英语 | <br>女高音<br>这种笔<br>食堂 | 吗？<br><br>吗？<br>了 |
| B | 通县<br>这<br>晚上<br>这位同志 | 已经 | **属于**<br>**成为**<br>**不如**<br>**姓** | 北京市<br>制度<br>早晨<br>李 | | |

***Figure 1. One example of a verb-predicate sentence pattern***

For each type of the sentence, for example the Transitive Verb Sentence shown in Figure 1, the author of [Lv 1980] provides the predicate in the sentence pattern.

From the above discussion, we can conclude that when linguists describe and further subdivide verb-predicate sentences, an important basis of their work is the object of the predicate. For example, among the thirteen kinds of verb predicates in [Lv 1980], the first eight kinds of sentence patterns are subdivided according to the type of object. However, our work is different from theirs because we pay closer attention to main verb types in verb-predicate sentences than to object types. The reason for this is shown in the following example.

> [MVP 通知/v(tong1zhi1, inform)][NP 他们/r(ta1men2, them)][VP 准备/v(zhun3bei4, prepare for)][MP 三/m(san1, three) 天/n(tian1, days)] 的/u(de) [NP 干粮/n(gan1liang2, solid food)] 。/ww

See the above example cited from [Meng *et al.* 2003]. In this example, the sentence is explained as being a pivotal sentence like a) in [Meng *et al.* 2003]. Obviously the above sentence takes more than one parse, such as b), c), and d), if syntactic information only is available.

a) **Pivotal sentence**: [Piv-O 他们] [Piv-V 准备] 三 天 的 干粮
Note: "他们" is the pivotal object, which acts as both the object of "通知" and the subject of "准备".

b) **Series verb**: [Object 他们] [2nd-V 准备] 三 天 的 干粮
Note: "通知" and "准备" are two series verbs. "他们" acts as the object of "通知".

c) **Clause as object**: [Object 他们 准备 三 天 的 干粮]
Note: The whole clause "他们准备三天的干粮" acts as the object of "通知".

d) **Double objects**: [Obj1 他们] [Obj2 准备 三 天 的 干粮]
Note: "通知" takes double objects including "他们" and "准备三天的干粮".

Since it is hard to employ consistent annotation in such sentences and we prefer that our annotation be theoretically neutral, in our specification, we subdivide a verb-predicate sentence into four types, including simple verb-predicate sentences, series-verb sentences, pivotal sentences, and sentences with verb-coordination, instead of using the objects of their predicates.

The Chinese Penn TreeBank (CTB) is a large-scale bracketed corpus of hand-parsed sentences in Chinese [Xia *et al.* 2000; Xue and Xia 2000]. The annotation of the Chinese Penn Treebank is more complete because they annotate everything, whereas currently we only annotate verb predicates. Compared with the "Guideline for Bracketing in the Chinese Penn TreeBank" [Xue and Xia 2000], our specification is different in that the goal of the CTB is to

annotate linguistically-standard and non-controversial **parse trees**, while the goal of our MVP annotation is based on **chunking** which is relatively easily parseable. For this reason, the guideline of CTB is not entirely identical to our specification. Other differences are listed as follows.

- The annotation of CTB is based on sentences that end with periods, exclamation marks, or question marks. Our specification defines the main verbs of Chinese simple sentences (see section 3.2.1 for the reference).
- Since we only focus on the output of chunking instead of whole parsed trees as in CTB, the MVP in our specification is a verb chunk with the main verb, while the predicate in CTB may be a whole phrase. For example, in CTB, we have the following:

  (IP (NP-PN-SBJ (NR 张三 zhang1san1, Zhangsan))
  (VP (VV 应该 ying1gai1, should)
  (VP (VV 参加 can1jia1, join)
  (NP-OBJ (NN 会议 hui4yi4, meeting)))))

  "In the above example, the lowest level VP (VP 参加会议) is the predicate," whereas based on our parsed chunk results, [VP 应该/v 参加/v] is annotated as an MVP in this sentence according to our specification.
- In CTB, "…a VP is always a predicate, -PRD is assumed….." However, in our specification, we only tag the main verb, that is, the verb corresponding to the main predicate-verb in the sentence. This annotation scheme is consistent to the sentence analysis methodology of Chinese linguists [Lv 1980].
- CTB also tags non-verbal predicates, such as ADJP/NP etc. In our specification, we don't consider this case since our focus is verb-predicate sentences.

Linguists provide a grammatical view of Chinese sentences by analyzing them. Identifying the main verb automatically is a task faced by many computational linguists. Most of their works have focused on the identification process instead of on the definition of the main verb. Previous works on MVI can be grouped into three categories: heuristic methods [Luo 1995; Sui and Yu 1998a]; statistical methods [Chen and Shi 1997; Sui and Yu 1998b]; first heuristic and then statistical methods [Gong *et al.* 2003].

Heuristic methods were introduced in the early stage of MVI research. Some proposed approaches depend on linguists' knowledge; for example, Luo [1995] used hand-crafted rules to identify predicates. The rules are related to auxiliary words, such as "的(de)" or "得(de)",

or to numerical or temporal words. Other approaches employ a bilingual corpus to extract rules, for example, Sui and Yu's [1998a] method. However a bilingual corpus is not always available.

Statistic methods were proposed in [Chen and Shi 1997] and [Sui and Yu 1998b]. Both of these works are based on verb sub-categorization information. But their categorization frameworks are different. Chen and Shi's work [1997] uses only part-of-speech information to decide on the main verb. Sui and Yu [1998b] use not only sub-categorized part-of-speech information but also lexicalized context information, such as "的". Both static and the context features are integrated into a decision tree model.

[Gong *et al.* 2003] first used rules to filter quasi-predicates. The features used include the part-of-speech of the quasi-predicate, the contextual part-of-speech, and the contextual words like "的". Then each feature's weight is calculated from training data. The combined weights are used to determine the predicates in the sentences.

The works noted above except that in [Chen and Shi 1997] presume that the sentence boundary has been given. All of them detect predicates in simple sentences. However, they have a deficiency in that in real text, the sentence boundaries are not provided naturally. Another difference is that the above works identify verb predicates, nominal predicates and adjective predicates. In our work, we focus on verb-predicate since both previous [Lv *et al.* 1999] and our own observations show that the sentences with verb-predicates make up the most part in corpus.

Another point is that some of the above works use correct verb sub-categorization information as input [Chen and Shi 1997; Sui and Yu 1998b]. They do not provide main verb identification evaluation results, where verb sub-categorization needs to be done automatically as a preprocessing step performed on raw text. Although the task of verb sub-categorization has long been studied in the Chinese community, the performance achieved has not been satisfactory. Thus in our work, we make use of more reliable knowledge; for example, we will provide a closed set of specific verbs whose objects can include multiple clauses, rather than sub-categorization information in general.

Finally, it is difficult to compare our results with the results of related works because the test corpora used may be quite different and there are also some differences in the definitions of the main verb. Thus, we hope that our introduction of a clear specification and corpus for main verb identification will enable future researchers to compare their results with ours.

## 3. Our Solution

### 3.1 Motivation for Developing Another Type of Specification

One reason for designing a specification is to ensure consistency of the corpus. In the "guideline of bracketing the Chinese", Xue and Xia [2000] explain this issue as follows:

*"Without doubt, consistency is one of the most important considerations in designing the corpus. . . .Many things can be done to ensure consistency, one of them is to make sure that the guidelines are clear, specific and consistent. . . .We also try to ensure that the guidelines cover all the possible structures that are likely to occur in the corpus. . . ."*

The above description indicates that a clear and wide coverage specification will ensure consistency of the annotated corpus. However, such a specification is not available publicly for main verb identification. To our knowledge, Luo [1995] was the first and the only one to propose a relatively simple definition. There are several deficiencies, however, in his specification. First, the definition is based on verb sub-categorization, which has been long criticized by linguistic community. Secondly, some parts of the definition are relatively simple and unclear. For example, "the verbs that have the subcategorized part-of-speech vgo or vgs etc. will be main verb in general cases; the verbs that have the part-of-speech vgn or vgv etc. will be main verbs in some cases or the modifiers of predicate-verb in other cases." But the author does not explain in which cases this assertion is true. Finally the proposed verb analysis using rules of exclusion does not cover some commonly used sentence patterns, such as series-verb sentences or verb-coordination sentences.

Thus, we propose another type of specification with the following characteristics.

- In order to ensure that the most important syntactic relations are covered, we base our main verb definition on various verb-predicate sentences.
- For specific purposes, our definition makes use of more reliable knowledge, such as a closed set of certain verbs whose objects can include multiple clauses rather than sub-categorization information in general.
- To deal with ambiguous syntactic constructions, we adopt a scheme in which we preserve the basic information and make the structures easily converted to structures following other annotation scheme. A similar scheme was used in [Xia *et a.l* 2000] and [Lai and Huang 2000].
- A lot of different complicated cases are studied, and the findings help make the specification's description clearer.

## 3.2 Design Specification

In this paper, we propose to define the main verb based on a simple sentence structure for the following reasons.

- A simple sentence is a sentence with only one predicate, and in our definition each predicate includes only one main verb if any. This guarantees that the main verb will have a unique operational definition.
- Chinese linguists have provided simple sentence structures in details, which have less disagreement between them. Since main verbs are related to simple sentence structures, we suppose there will be less disagreement in main verb definition with the help of simple sentence structures.

Because our annotation is based on a simple sentence, we firstly define the simple sentence and then the predicate, especially the predicate-verb if one exists, of each simple sentence. Then, we discuss in detail on the complicated aspects of our spec design and corpus annotation. This discussion will help to uncover the difficult point of the main verb identification.

### 3.2.1 Sentence Definition

Chinese sentences are of two types: simple sentences and complex sentences. The boxes above the dashed line in Figure 2 show the widely accepted sentence pattern classification [Lv 1980; Ding *et al.* 1961]. In our specification, since we pay more attention to main verb types in verb-predicate sentences instead of object types, we subdivide verb-predicate sentences into four types as shown below the dashed line in Figure 2.

**Definition 1:**

*A simple sentence* is a sentence with only one predicate-verb.

The predicate of a simple sentence can be a verb, an adjective, a noun, or a subject-predicate in Chinese [Liu *et al.* 2002]. Accordingly, simple sentences are categorized as verb-predicate, adjective-predicate, noun-predicate and subject-predicate sentences, respectively. Here, a subject-predicate sentence has a subject-predicate phrase as its predicate. For example, in the sentence “他(ta1) 肚子(du4zi1) 疼(teng2)” (He has a stomach-ache), “肚子疼” is a subject-predicate phrase acting as the predicate, while “他” is the subject of the sentence. In our specification, we only focus on simple sentences with verb-predicates.

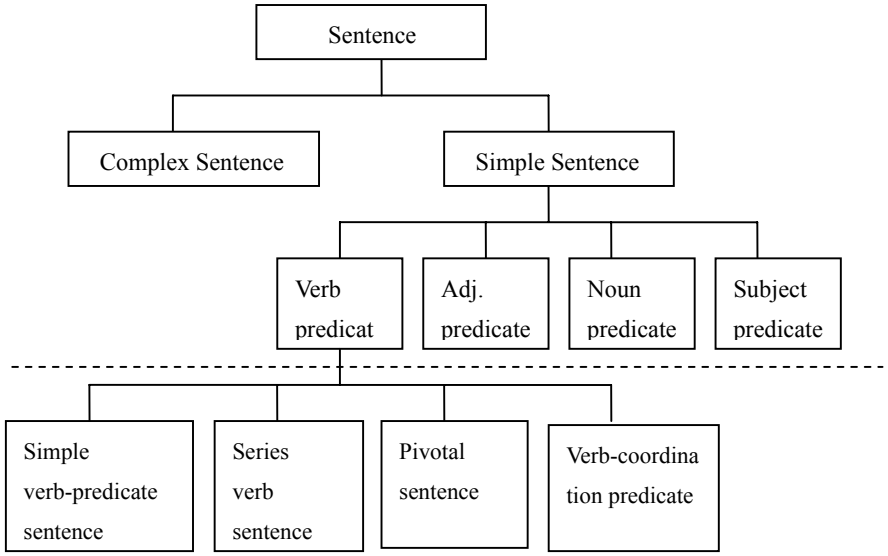



***Figure 2. Sentence pattern classification***

**Definition 2:**

*A complex sentence* is made up of two or more simple sentences. The simple sentences in one complex sentence can not be included each other.

**Definition 3:**

In a complex sentence，each *sub-sentence* is a sentence, which can be either a complex sentence or a simple sentence.

Another related topic that should be introduced is punctuation at the end of sentence. In general, "。|？|！|；" are punctuation used at the end of a sentence in Chinese. Sometimes "，|：|——|……" can also be seen as the end of a sentence if it has the main verb. See example 31 in section 6.

### 3.2.2 Main Verb Definition

**Definition 4:**

*The main verb* is the predicate-verb, if one exists, in a simple sentence. It corresponds to a tensed verb in English.

In this paper, we will only discuss the main verb in a verb-predicate sentence. Each verb-predicate sentence contains only one main verb, which is the predicate-verb of the sentence. Verb-predicate sentences can be classified into four types shown in Figure 2. Some examples of verb-predicate sentences are shown below.

***Example* 7** (simple verb-predicate sentence)

[NP 张/nr(zhang1) 晓伟/nr(xiao3wei3)] **[VP 坚决/ad(jian1jue2) 不/d(bu4) 收/v(shou1)]** 。/ww[6]

*(Zhang xiaowei **resolutely refused** to accept. )*

***Example* 8** (pivotal sentence)

**[VP 必须/d(bi4xu1) 先/d(xian1) 请/v(qing3)]** [NP 外国/n(wai4guo2) 专家/n(zhuan1jia1)] [VP 运行/v(yun4xing2)] [VP 管理/v(guan3li3)] 。/ww

*([One] **must first invite** the foreign expert to run and manage [it].)*

***Example* 9** (series-verb sentence)

[NP 张/nr(zhang1) 晓伟/nr(xiao3wei3)] **[VP 连忙/d(lian2mang2) 返回/v(fan3hui2)]** [NP 大/a(da4) 水潭/n(shui3tan2) 边/n(bian1)] [VP 去/v(qu4) 找/v(zhao3)] ，/ww

*(Zhang xiaowei **immediately returned to** the big puddle to search for [it])*

***Example* 10** (sentences with verb-coordination predicate)

[NP 交通/n(jiao1tong1) 肇事/vn(zhao4shi4)] [SP 后/f(hou4)] ，/w [NP 肇事/vn(zhao4shi4) 司机/n(si1ji1)] **[VP 伪造/v(wei3zao4)]** 、/w [VP 破坏/v(po4huai4)] [SP 现场/s(xian4chang3)] 。/ww

*(After the traffic accident, the trouble-making driver **falsified** and destroyed the scene.)*

In the above examples, the main verb in each sentence has been underlined. Without doubt, in simple verb-predicate sentences, the main verb is the predicate verb. In a series-verb sentence, a pivotal sentence, or a sentence with a verb-coordination predicate, the first predicate-verb of that construction is defined as the main verb of the sentence.

A serious concern with main verb definitions is the treatment of different syntactic constructions in different researchers' works. For instance, there is another point of view that both of the verbs in a verb-coordination sentence can be main verbs. However since there exist different levels of verb coordination, such as word level, phrasal level, and even clause level coordination [Xue and Xia 2000], we adopt a scheme similar to that used in [Lai and Huang 2000; Xia *et al.* 2000]. What we do is label the first verb as the main verb, preserve the VP information, and leave deeper analysis of verb-coordination for future work. From another point of view, it is easier to convert our annotation to other specifications with the preserved

[6] Chunks tags here are annotated according to our chunk spec [Li *et al.* 2004].

information.

### 3.2.3 Complicated Cases in Main Verb Annotation

Sentences in running text are complicated. To maintain inter-annotator consistency during corpus annotation, we not only perform cross-validation but also examine the phenomena that appear in our corpus annotation. This helps us to understand the problem of main verb identification. In the following, we classify the complicated cases into six types.

**1) Verbs in a non verb-predicate sentence**

Verbs or verb chunks (VPs) in a non verb-predicate sentence, whose predicates are formed by an adjective, nominal, or subject-predicate phrase, should not be treated as main verbs or MVPs. See Example 11 below.

***Example* 11**

[NP 我/r(wo3)] **[VP 吃/v(chi1)]** 的/u(de) [NP 邱县/ns(qiu1xian4) 饭/n(fan4)] ，/ww **[VP 喝/v(he1)]** 的/u(de) [NP 邱县/ns(qiu1xian4) 水/n(shui3)] ，/ww **[VP 当/v(dang1)]** 的/u(de) [NP 邱县/ns(qiu1xian4) 官/n(guan1)] ，/ww

*(What I ate [was] Qiuxian's meal. What I drank [was] Qiuxian's water. What I worked as [was] a Qiuxian's officer.)*

Note: These three sentences are sentences with predicates that are formed by subject-predicate phrases. All of them share the same subject, "[NP 我/r]" (I). "[VP 吃/v] 的/u [NP 邱县/ns 饭/n]" (what I ate) is a subject-predicate phase, in which the 的-structure "[VP 吃/v] 的/u" (ate + de) acts as a nominal subject, while [NP 邱县/ns 饭/n] (Qiuxian's meal) is a nominal-predicate. Thus, no main verbs can be found in these three sentences.

***Example* 12**

[NP 人们/n(ren2men2)] [VP 生活/v(sheng1huo2)] [ADJP 很/d(hen3) 苦/a(ku3)] 。/ww

*(People's lives are very bitter.)*

Note: This is a subject-predicate sentence in which the subject-predicate phrase [VP 生活/v] [ADJP 很/d 苦/a] acts as the predicate of the sentence. Thus, "生活"(life) should not be tagged as an MVP.

In example 12, annotators tend to tag "生活" (life) as a MVP because they incorrectly analyze

verb-predicate sentences and subject-predicate sentences with VPs.

**2) Auxiliary Verbs**

Auxiliary verbs are a special subdivision of verbs. Typically, they are placed before a verb, e.g., "会跳舞" (hui4tiao4wu3, be able to dance). In our specification, there is a closed set of 26 auxiliary verbs, including 能 (neng2, can), 会 (hui4, be able to), 可以 (ke3yi3, may), 应该 (ying1gai1, should) etc. However, these auxiliary verbs in the PK corpus share the same part of speech tag: "v".

As for the question of whether the auxiliary verbs can be used as main verbs, there is disagreement among Chinese linguists. Some suppose that auxiliary verbs can be treated as predicate verbs [Zhu 1982] while others propose that auxiliary verbs have the same syntactic functions of adverbial modifiers [Hong 1980]. Thus, we propose that auxiliary verbs should be annotated on a case by case basis.

● Auxiliary verb in a VP chunk

In our chunk specification [Li *et al.* 2004], we treat an auxiliary verb as a pre-modifier of an adjoining main verb. See in Example 13, the annotation of MVPs is not affected since the auxiliary verb is chunked with the main verb.

*Example* 13

[NP 欧盟/j(ou1meng2) 国家/n(guo2jia1)] **[MVP 也/d(ye3) 不/d(bu2) 会/v(hui4) 大力/d(da4li4) 干预/v(gan1yu4)]** 。/ww

*(The countries of the European Union will not intervene energetically, either.)*

In the above example, the main verb is "干预" (intervene), while the preceding auxiliary verb "会" is treated as a pre-modifier of "干预".

● Auxiliary verb outside a VP chunk

An auxiliary verb can be a single chunk of a VP that is separated from its modifying VP by a following prepositional phrase, noun phrase. Or the auxiliary verb is followed by VP coordination. In this case, we annotate the VP of the auxiliary verb as a MVP. Perhaps some will argue that the main verb can be a verb followed an auxiliary verb. In our annotation scheme, we want to annotate the sentences consistently. For example, in the sentence "[NP 价格/n(jia4ge2)] **[MVP 要/v(yao4)]** [ADJP 低/a(di1)] [MP 一些/m(yi4xie1)] ，/ww" (*The price **is** a little lower.),* there are no other verbs in the sentence, and the verb "要" is a MVP. Thus, there is no need to decide whether the verb "要" is a common verb or an auxiliary verb. From another point of view, if some researchers prefer to treat an auxiliary verb as a

Non-MVP, it is easy to convert our annotation in order to accommodate their specification. Some examples are listed as follows.

***Example* 14**

[NP 国家/n(guo2jia1)] 的/u(de) [NP 事/n(shi4)] **[MVP 要/v(yao4)]** [NP 大家/r(da4jia1)] [VP 关心/v(guan1xin1)] ，/ww

*(The businesses of the country **need** people's attention.)*

Note: In this example，there is a NP instead of a PP following the auxiliary verb "要".

***Example* 15**

**[MVP 能够/v(neng2gou4)]** [PP 把/p(ba3)] [NP 一般/a(yi4ban1) 号召/vn(hao4zhao1)] [PP 与/p(yu3)] [NP 个别/a(ge4bie2) 指导/vn(zhi3dao3)] [VP 结合/v(jie2he2) 起来/v(qi3lai2)] ，/ww

*([One] **is able to** combine the general calling with an individual guide.)*

***Example* 16**

**[MVP 应该/v(ying1gai1)]** [ADVP 坚决/ad(jian1jue2)] [VP 反对/v(fan3dui4)] 和/c(he2) [VP 制止/v(zhi4zhi3)] 。/ww

*([One] **should** firmly oppose and prevent [it].)*

Note: In the above sentence, the auxiliary verb "应该" (should) modifies a verb coordination phrase "[VP 反对/v] 和/c [VP 制止/v]" (oppose and prevent).

**3) "PP+XP+VP" sequences**

In real text, there are a lot of prepositional sequences like "[PP 从/p(cong2, from)] +… + [VP 起步/v(qi3bu4, beginning)]", "[PP 从/p(cong2, from)] + … +[VP 看/v(kan4, watch)]", "[PP 按/p(an1, according to)] + … +[VP 计算/v(ji4suan4, calculate)]", "[PP 以/p(yi3, according to)] + …+[VP 为由/v(wei2you2, excuse)]". We call these sequences PP+XP+VP sequences. One issue to be considered is whether the VP in the sequence is the object of the preposition (PP).

There is a limited number of cases where PP can include the following VP as a part of its object. See Example 17 in [Liu *et al.* 2002]. In this case, we do not annotate the VP as a MVP since the VP acts as the head of a verb phrase, which in turn acts as the object of the PP. The prepositions that can have a verb (or VP) or a clause as their object are also summarized in a

closed set, including “为了”(wei4le, for), “随着”(shui2zhe, with), “关于”(guan1yu2, about) etc.

***Example* 17**

[PP 关于/p(guan1yu2)] [NP 怎么样/r(zen3me1yang4)] [VP 学好/v(xue2hao3)] [NP 汉语/nz(han4yu3)] ，/w [NP 阿里/ns(a1li3)] **[MVP 谈/v(tan2) 了/u(le)]** [ADJP 很/d(hen3) 多/a(duo1)] 。/ww

*(Ali **talked** a lot about how to learn Chinese well.)*

Note: “学好汉语” (learn Chinese well) is a verb phrase in the object of the preposition “关于”. Thus “学好” (learn) should not be tagged as the main verb of the sentence.

However, in most situations, we cannot include the VP in the object of the PP. Nor can the VP be treated as the MVP since it is more likely to be parenthesis[7] in Chinese. See Example 18 below.

***Example* 18**

[PP 按/p(an1)] [NP 可比/vn(ke3bi3) 口径/n(kou3jing4)] **[VP 计算/v(ji4suan4)]** ，/w [TP 去年/t(qu4nian2)] "/w [NP 两/m(liang3) 税/n(shui4)] "/w [MVP 实际/ad(shi2ji4) 完成/v(wan2cheng2)] [MP ４０８３亿/m(yi4) 元/q(yuan2)] ，/ww

*(Calculated from constant requirements, “two taxes” actually are collected 408,300 million yuan last year.)*

Like the above example, we summerized 14 similar structures like [PP 按/p(an1, according to)]+XP+[VP 计算/v(ji4suan4, calculate)], [PP 从(cong2, from)]+XP+[VP 看/v(kan4, watch)] etc. VPs in these structures are not treated as MVPs.

Otherwise, in a PP+XP+VP sequence, VPs can be viewed as MVPs if those verbs are verbs whose objects can include multiple clauses. See Example 19 in [Liu *et al.* 2002].

[7] Parenthesis is a grammatical phenomenon in Chinese grammar. For example, 据了解, 据介绍, 我看, 我说 are all examples of parenthesis. In our spec, we should not tag a VP like “了解” or “介绍” as a MVP in these parentheses.

***Example* 19**

[PP 从/p(cong2)] [NP 孩子/n(hai2zi1)] [SP 嘴里/s(zui3li3)] **[MVP 知道/v(zhi1dao4)]** ，/w [NP 他/r(ta1)] [NP 姐姐/n(jie3jie3)] [VP 是/v(shi4)] [NP 个/q(ge4) 转业军人/n(zhuan3ye4jun1ren2)] 。/ww

*(From the child's mouth , [we]* ***know that*** *his elder sister is a former member of the military who has transferred to civilian work.)*

Note: Although the VP "知道" (know) follows the preposition "从" (from), "知道"(know) is a verb whose object can include multiple clauses. Thus, "知道" (know) should be treated as the MVP of the sentence. The following clause "他姐姐是个转业军人" (his elder sister is a former member of the military who has transferred to civilian work.) is the object of "知道" (know).

**4) Verb "有"**

"有"(have) can be used as a MVP in the following three sentence patterns: a "有-sentences", which has the basic possession sense, e.g., 我有一本书 (I have a book), series-verb sentences, and pivotal sentences [Liu *et al.* 2002]. In most of the above cases, "有" is annotated as the main verb. However, some "有" sentences should not be treated as series-verb or pivotal sentences, nor should "有" be treated as the predicate verb in these sentences. See example 20.

***Example* 20**

**[VP 有/v(you3)]** [MP 一/m(yi2) 次/q(ci4)] [NP 灵感/n(ling2gan3)] **[MVP 来/v(lai2) 了/y(le)]** ，/ww

*(Once upon a time, the inspiration came.)*

***Example* 21**

**[VP 有/v(you3)]** [NP 风险/n(feng1xian3)] [NP 我/r(wo3)] **[VP 来/v(lai2)担/v(dan1)]**。/ww

*(I will take the risk.)*

Note: This is a sentence with a predicate of a subject-predicate phrase, where the verb-object phrase "有/v 风险/n" (risk) is the subject of the sentence.

**5) Verb "是"**

Ambiguity is encountered in "是" (is) sentences when verbs are in the subjects of "是". If the VPs are inside the subject of the "是-sentence", we cannot annotate such VPs as MVPs no

matter whether there is punctuation like "，" immediately before "是" or not. See example 22.

***Example* 22**

[NP 买家/n(mai3jia1)] **[VP 不/d(bu2) 怕/v(pa4)]** [NP 赝品/n(yan4pin3)] ，/w **[MVP 也/d(ye3) 是/v(shi4)]** [PP 为了/p(wei4le)] [MP 一个/m(yi2ge4)] [NP "/w 钱/n(qian2) "/w 字/n(zi4)] 。/ww

*(It **is also** for the reason of "money" that the buyer is not afraid of forgeries.)*

Note: Although we find the punctuation "，" before "是", the whole clause, "[NP 买家/n] [VP 不/d 怕/v] [NP 赝品/n]" (the buyer is not afraid of forgeries), acts as the subject of the "是-sentence". Thus, the VP "不/d 怕/v" (is not afraid of) inside it should not be tagged as a MVP.

**6) Multiple clauses in a subject or object**

We should note that there are many long sentences in texts whose subjects or objects include multiple clauses. These clauses are similar to English ones, and the verbs are nearly a closed set. It includes, for example, "觉得" (feel), "希望" (hope), "认为" (think), and "以为" (suppose) which are listed in our specification. The problem with annotating this kind of sentence stems from the ambiguous subject or object boundaries. See example 23.

***Example* 23**

[NP 张三/nr(zhang1san1)] **[VP 承认/v(cheng2ren4)]** [NP 李四/nr(li3si4)] **[VP 是/v(shi4)]** [MP 一个/m(yi2ge4)] [ADJP 重要/a(zhong4yao4)] 的/u(de) [NP 谈判/vn(tan2pan4) 因素/n(yin1su4)] ，/ww

This sentence has two readings.

1) [VP 承认/v] (admit) is the main verb, and the following clause [NP 李四/ns][VP 是/v]…[NP 谈判/vn 因素/n] (Li is the negotiation factor) is the object of [VP 承认/v]. An English translation of this sentence is "*Zhang admitted that Li is an important negotiation factor.*"
2) [VP 是/v] (is) is the main verb, and the clause [NP 张三/nr] [VP 承认/v] [NP 李四/nr] (Zhang admit Li) is the subject. The English gloss of this sentence is "*[The fact] that Zhang admitted Li is an important negotiation factor*."

Example 23 shows ambiguity with respect to the subject boundary. Example 24 below shows

ambiguity with respect to the object boundary.

***Example* 24a**

[NP 中/j(zhong1)] 、/w [NP 俄/j(e2)] 、/w [NP 法/j(fa3) 等/u(deng3) 国/n(guo2)] [VP 认为/v(ren4wei2)] [VP 可以/v(ke3yi3) 结束/v(jie2shu4)] [PP 对/p(dui4)] [NP 伊拉克/ns(yi1la1ke4)] 的/u(de) [VP 核查/v(he2cha2)] ，/w [NP 美国/ns(mei3guo2)] [VP 则/d(ze2) 坚决/ad(jian1jue2) 反对/v(fan3dui4)] 。/ww

This sentence also has two readings.

1) Both of the clauses following [VP 认为/v] (think) are its objects. In this case, the sentence can be translated as "*Countries such as Chinese, Russia and France thought that the investigation on Iraq could be finished, and [they also thought] that the United States firmly opposed it*."
2) Only the clause immediately following [VP 认为/v] (think) is its object. The next sentence is an independent one. In this case, the sentence can be translated as "*Countries such as Chinese, Russia and France thought that the investigation on Iraq can be finished. [However], the United States firmly opposed it.*"

The two readings of Example 23 are reasonable. But only the reading 2) of Example 24a is reasonable according to the context. However, for a computer, it is hard to make decision here since 1) in Example 24 is also a reasonable parsing candidate if the computer does not have the additional knowledge. For these ambiguities, we apply an annotation scheme similar to that in CTB [Xue and Xia 2000]. If the syntactic ambiguity can be resolved with the knowledge of the context, then we annotate the correct reading. The proposed annotation of Example 23 is based on the context. The proposed annotation of Example 24a is as follows:

***Example* 24b**

[NP 中/j] 、/w [NP 俄/j] 、/w [NP 法/j 等/u 国/n] **[MVP 认为/v]** [VP 可以/v 结束/v] [PP 对/p] [NP 伊拉克/ns] 的/u [VP 核查/v] ，/ww [NP 美国/ns] **[MVP 则/d 坚决/ad 反对/v]** 。/ww

*(Countries such as Chinese, Russia and France thought that the investigation on Iraq could be finished. [However], the United States was firmly opposed to [ it].)*

In the above example, if there is no punctuation immediately after the predicate-verb, the predicate-verb is annotated as a MVP, and the first sentence will end after the punctuation following the first clause. This means that the VP in the first sub-sentence should not be tagged as a MVP at all. The remaining sub-sentences will annotate their predicate-verbs as MVPs and are broken one by one. Also, if some linguists prefer the clause "[NP 美国/ns] [MVP 则/d 坚决/ad 反对/v]"(the United States firmly opposed) as the object of [VP 认为] (think), then they can carry out another task to identify this kind of object since none of the syntactic information of this sentence is lost.

### 3.2.4 Assignment of Descriptors

Three annotation descriptors are needed: "MVP", "/ww" and "#/ww". The chunk labels are pre-annotated before MVP annotation is performed. The combined label "MVP" indicates the main verb chunk of a sentence. "/ww" and "#/ww" stands for the end of a sentence, where "#/ww" is used to indicate that the sentence lacks of an ending punctuation.

## 3.3 MVP Statistic

Based on the main verb definition given above, we investigated the distribution of simple sentence types in the annotated PK corpus, which has a total of 100, 417 tokens[8]. The sentences in the corpus were manually annotated with the sentence end tag "/ww" defined above. We got 8, 389 sentences of this kind.

In Figure 3, we show the distribution of three sentence types, that is, sentences with MVPs, sentences without MVPs but with one or more VPs, and sentences without any VPs at all. Sentences with MVPs are given in Examples 7 to 10. Sentences without MVP but with VPs are ones like "[NP 人们/n(ren2men2)] [VP 生活/v(sheng1huo2] [ADJP 很/d(hen3) 苦/a(ku3)] 。"(People's lives are hard). Sentences without VPs are ones like "[NP 劳动/vn(lao2dong4) 经验/n(jing1yan4)] [ADJP 少/a(shao3)] 。" (Work experience is rare).

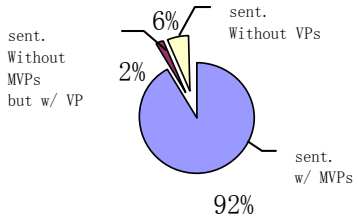



***Figure 3. Distribution of sentences w/o MVPs***

[8] Here *tokens* include words, punctuation mark in the entire corpus.

From the above figure, we can see sentences with MVPs make up most of the sentences, approximately 92%. This result agrees with Wu's assertion in [Lv *et al*. 1999]. Among these 92% sentences, we find that about 80% of the MVPs are the first VPs in the sentences.

Figure 4 shows the distribution of the remaining 8% of the sentences, totally 671 sentences without any MVPs. The non-predicate sentences are sentences like [NP 照片/n 人物/n] 的/u [NP 故事/n] #/ww. (The story of the people in pictures). These sentences come from the titles of texts or headlines of news reports.

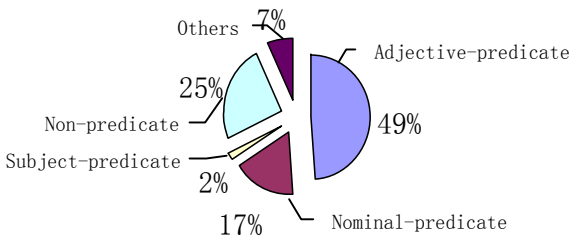



***Figure 4. Distribution of sentences without MVPs***

Since the MVP sentences amount for most of the sentences (i.e., 92% of all the sentences in the PK corpus), our study focused on identifying the verb predicates in the sentences. We will explore them in more detail below.

## 3.4 A Model for Chinese Main Verb Identification

Our aim is to conduct main verb identification on a binary classifier. For each VP, we determine whether it is an MVP or not. The Support Vector Machine (SVM) is one of the most successful binary classifiers. This method has been used in many domains of NLP, such as part-of-speech tagging [Nakagawa *et al.* 2001], Name Entity recognition [Isozaki and Kazawa 2002], Chunking [Li *et al.* 2004] and Text categorization [Joachims 2002]. To our knowledge, the use of SVM to identify Chinese main verbs has not been studied previously. Moreover, there are indications that the differences among various learning techniques tend to get smaller as the size of the training corpus increases [Banko and Brill 2001].

We follow the definition of SVM in [Vapnik 1995]. Suppose the training samples are $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where each $x_i$ $(1 \le i \le N)$ represents an input vector defined on an n-dimensional space, and each dimension is a feature we define in the following sections. $y_i \in \{1, -1\}(1 \le i \le N)$ indicates whether it is MVP or not. The separating hyperplane is defined by

$$\vec{w} \cdot \vec{x} + b = 0 \quad \vec{w} \in R^n, b \in R .$$

SVM searches for the hyperplane that separates a set of training examples that contain two distinct classes with the maximum margin. We use SVM$^{light}$ [Joachims 1999] as our implementation tool.

A processing cycle can arise here. Because most of the related works are based main verb identification in sentences with pre-determined sentence boundaries, sentence boundary labeling must be done before tagging. But if sentence boundary labeling is done before tagging, where does the predicate information come from? So instead of doing sentence boundary labeling beforehand, we first detect the predicate without using sentence boundary information. It is for this reason that we want to break the sentence into simple sentences that by definition require main verbs. This procedure is similar to the work in [Chen and Shi 1997]. Firstly, we break the sentence into process units. They are word sequence separated by punctuation marks, such as "。！？，", but we do not know if they are sentence ending labels or not. Secondly, our algorithm determines whether the VPs are MVPs in these units. If the value is negative, the VP is not a MVP and vice versa. Finally, if more than two MVPs are identified in a processing unit, we rank these MVPs according to the classifier's output (value of the decision function) and choose the one with the highest rank as the MVP. The chunk information is obtained from our chunking system.

Building an effective SVM classifier involves choosing good features. We break up the features used in our research into two categories, local and contextual. The first set of features is derived from the surface information of VPs. Since these features are based on chunks themselves, they are called local features. The second set of features is derived from the context information of VPs, while also incorporating some lexical knowledge and patterns. Thus, we call these features contextual features. Our model is based on the level of chunking because our experiments show that this is better than basing the model on parts of speech.

In the following sections, we will describe the feature set in detail.

### 3.4.1 Local Features

Local features are explored based on careful observation of the training corpus. All of them are new features we have proposed. Although they are simple, they work well in MVI since they represent the characteristics of the VPs themselves. Our model captures three local features: 1) the VP position, 2) the VP length 3) and the probability of head verbs being MVPs. Here, VP position and VP length are feature groups. Each feature group is made up of several binary features. This means for each VP, if one feature in the group is set to 1, other features in the same feature group are set to zero.

***VP position*** is a feature group. Totally, there are six binary features in this group. This means that the phrasal position number of a VP appears in the process unit, which starts with 1. For example, if the VP is the first VP in the process unit, the value of the first feature is 1, and the other feature values are set to zero. If the position value of the VP in the process unit is larger than 5, then the value of the sixth feature is 1, and the other features are set to zero. Figure 5 shows the VP position distribution.

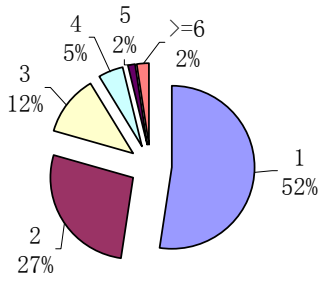



***Figure 5. VP position distributions***

In the figure, we show the distribution of up to six binary features because the percentage of VPs with position values of 6 or less is 98%.

Also based on our statistics for the training data, about 80% of the MVPs are the first VPs in the sentences. So we use this feature as the base-line feature (refer to section 4.2).

***VP length*** is also a feature group. Totally there are six binary features, chosen based on our intuition that the longer a VP is, the more likely it is a MVP since it has more modifiers. VP length is measured in terms of the number of words in a VP. Thus, the $i^{th}$ feature in the feature group stands for a VP with a length of i (starting from 1). The sixth feature means the VP has a length larger than 5. See the VP length distribution shown in Figure 6.

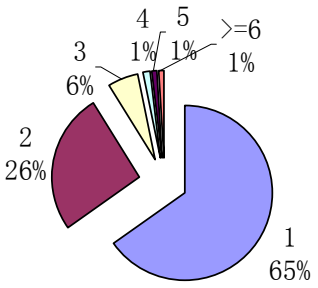



***Figure 6. VP length distribution***

In the above figure, we show the distribution of up to six binary features since the percentage of VPs with lengths smaller than 6 is about 99%. For example, the VP "坚决/ad 不/d 收/v" has a length of three. See in table 2, the third feature is set to 1, and other feature values are set to zero.

***Table 2. VP length feature table***

| *Feature Number* | *1* | *2* | *3* | *4* | *5* | *6* |
|---|---|---|---|---|---|---|
| Feature Value | 0 | 0 | 1 | 0 | 0 | 0 |

***Probability of head verbs being MVPs*** is a real value feature. Our statistics show that some VPs are MVPs, like "是 (is)" and "认为 (think)". This feature is estimated beforehand as follows based on the training corpus.

$$\text{MVP\_P(x)}=\frac{\text{C(x in MCVP)}}{\text{C(x)}},$$

where C(x in MVP) is the number of occurrences of verb x as a MVP and C(x) is the total occurrences of verb x in the training data.

### 3.4.2 Contextual Features

So far, we have introduced features that are based on the characteristics of a VP itself. One problem with these features is that they only use the surface information of a VP, not its contextual information. Related works [Sui and Yu 1998b; Gong *et al.* 2003] have shown contextual features are helpful in MVI. Thus, we also incorporate contextual features into our model. One difference is that in our work, we added new features included in our specification, such as "PP+XP+VP" sequences, into our model. In addition, we integrate them into our SVM model instead of dealing with this problem in two steps as in [Gong *et al.* 2003].

***Pattern features*** are one type of binary feature. The patterns we define include features like "的" (de) and "得" (de) that were also used in [Gong *et al.* 2003; Sui and Yu 1998b]. One difference is that we only consider "的" when it is next to a VP. In addition, we find that in about 92% of cases, the verb "是" followed by "的" is used as a MVP, so we treat this word differently from other verbs. These pattern features are very precise based on our statistics on training corpus.

***Table 3. Pattern feature table***

| | |
|---|---|
| Pattern Features | VP+SP |
| | VP+NO_CHUNK_UNIT |
| | PP+VP |
| | "《" + VP + "》" |
| | "的" + VP |
| | "、" + VP |

In Table 3, VP+SP means a SP chunk followed a VP chunk. NO_CHUNK_UNIT indicates the out-of-chunk units as defined in our chunk system, including "等" (etc.), "之/r" (zhi), "的"(de) etc. These pattern features indicate the contexts of MVPs. They share the same formulation shown below:

$$f^i(VP) = \begin{cases} 1, \text{if VP corresponds to a defined pattern} \\ 0, \text{otherwise} \end{cases}.$$

***Anti-features*** include words and patterns in which VPs can not be used as MVPs. We define an anti-feature as a binary feature. If a VP meets this requirement, the f(VP) = 1; otherwise, f(VP) = 0. If a VP appears in an anti-pattern, it will be masked, and other features will not be added. Anti-features are mostly derived through our careful observation of the specification.

1) Lexical anti-feature to exclude MVP

As described in our specification, "据了解" (ju4liao2jie3, it is reported), "据介绍"(ju4jie2shao4, it is introduced), "我看" (wo3kan4, I see), "我说" (wo3shuo1, I say) are all examples of parenthesis. VPs like "了解" (report)" and "介绍" (introduce) in such contexts are not used as MVPs. In addition, based on the statistic of the training data, some words are typically not used as MVPs, like "新年伊始" (xin1nian2yi1shi3, the beginning of New Year), "解放思想" (jie3fang4si1xiang3, emancipate the mind), etc. Lexical anti-features of the above two types are set to 1. This kind of information is stored in a list of words explored in our specification.

2) Frame anti-feature to exclude MVP

The VPs in frame-like structures like [PP 在/p(at)]+...+[SP 上/s(above)] are not MVPs. Because of the right boundary of such a long prepositional phrase is hard to identify and to avoid ambiguities, our chunk system only finds the PP chunks of frame-like construction with explicit boundaries and length constraints, such as [PP 在/p (at) … 中/f (middle)] [Li *et al.* 2004]. We have to detect Non-MVPs in longer prepositional phrases. Statistics show that based on the current PP chunk tags, some right boundaries of longer prepositional phrases can be recognized.

Take the PP "在"(zai4) as an example. We collect all the SP chunks as its right boundary candidates in the training data. Among the resulting 111 SP chunks, only 1 SP chunk is not a right boundary. So the pattern [PP 在/p] + SP is very precise to form a longer PP. From another point of view, we use such kind of patterns to perform a rough PP boundary recognition. For example, [PP 当/p(dang1)] [NP 他们/r(ta1men2)] [VP 来到/v(lai2dao4)] [NP 另/r(ling4) 一个/m(yi1ge4) 风景点/n(feng1jing3dian3] [VP 要/v(yao4) 拍照/v(pai1zhao4)] [SP 时/Ng(shi2)] ，/w (When they come to the spots of interest to take photos) is a PP, and [VP 来到/v] (come) and [VP 要/v 拍照/v] (to take photos) are masked as Non-MVPs. Table 4 lists three types of anti-patterns. It should be noted that the frame structures are not limited to PPs. These structures are selected from the statistics of the training corpus.

***Table 4. Frame anti-feature types***

| Frame Anti-pattern type | Examples |
|---|---|
| Only chunk type | [PP 在/p] + [SP */*] ; |
| | [VP 有/v] + [MP */*] |
| Only lexical type | [PP 当/p] + [SP 之际/Ng] |
| Both chunk and lexical types | [PP 将/p] + [NOCHUNK_WORD 的/u] [NP */*]; |
| | [VP 找到/v] + [NP */*] [SP 时/Ng] |

Here, the first chunk is the trigger chunk. That is, if we encounter such a chunk, we trigger the pattern matching module, and all the VP chunks are blinded. That is, we set f(VP) = 1 if the chunks match one of the patterns from MVP identification. See the example above where feature values of "来到" (lai2dao4, come) and "要/v 拍照/v" (yao4 pai1zhao4, to take photos) are both set to 1. Totally, we have 62 patterns. Among them, 52 patterns have PP trigger chunks. Ten patterns have VP trigger chunks. Similar patterns can be designed according to "有" sentences or "是" sentences in our specification.

In our implementation, we have a module that we use to convert the corpus into the proper input format for SVM$^{light}$ [Joachims 1999]. Each of the above features corresponds to one dimension of the feature vector. In the next section, we will discuss the evaluation results.

## 4. Experiments

We evaluated our MVI approach using manually annotated data, which was a subset of the PK corpus. The PK corpus was released by the Institute of Computational Linguistics, Peking University. The corpus contains one month's data from the *People's Daily* (January 1998). This corpus has already been segmented and part-of-speech tagged. Its specification has been published in [Yu *et al.* 2002]. Totally, there are about 40 part-of-speech tags including noun (/n), verb(/v), adjective(/a), name entity tags, verbal noun (/vn) etc. The details of our MVI training and testing corpus are shown in Table 5.

***Table 5. Training and testing corpus***

| **Data** | **# of Chunks** | **# of Tokens** | **# of Simple Sent.** | **# of Whole Sent.** | **Ave. Simple Sent. Length** | **Ave. Whole Sent. Length** |
|---|---|---|---|---|---|---|
| Train | 72, 645 | 100, 417 | 8, 389 | 3, 784 | 11.97 | 26.54 |
| Test | 19, 468 | 26, 334 | 2, 456 | 1, 047 | 10.72 | 25.15 |

Here, *tokens* include words and punctuation marks in the entire corpus. *Chunk* marks are annotated according to our chunk spec [Li *et al.* 2004] with 11 chunk tags. *Simple sentences* are annotated as described in our spec. *Whole Sentences* are sentences that use "。！？" as sentence endings. As the table shows, following annotation of simple sentences, the average length of a simple sentence was less than two times the length of a whole sentence. Notice that we do not use sentence-ending label information in our algorithm.

### 4.1 Evaluation Metrics

The evaluation metrics used here are the traditional Precision, Recall and F Measure:

$$\text{Precision (P)} = \frac{\#\text{ of Correct MVPs in system output}}{\#\text{ of Total MVPs in system output}};$$

$$\text{Recall (R)} = \frac{\#\text{ of Correct MVPs in system output}}{\#\text{ of Total MVPs in answer}};$$

$$F = 2 * P * R/(P + R).$$

We also compared our evaluation metrics with the Sentence Accuracy Rate (SAR) used in related works:

$$\text{SAR} = \frac{\#\text{ of correct tagged verb - predicate sentences}}{\#\text{ of total verb - predicate sentences}}.$$

We propose using the Precision/Recall evaluation metrics for three reasons: Firstly, these evaluation metrics are more widely used than a single percent-correct score. Secondly, we don't deal with sentences whose predicates are adjective phrases or noun phrases. So if we include these sentences into the total number of sentences in our calculations, the performance will suffer and the result will not reflect the performance of identifying MVPs. Thirdly, in our approach, we do not discriminate verb-predicate sentences with other sentences. However, in order to show the soundness of our technical approach, we also provide the SAR, and we manually calculate the number of verb-predicate sentences.

## 4.2 Impact of Different Features on MVP Identification Results

We investigated the contributions of different features as shown in Table 6.

***Table 6. Impact of different features***

| Model | Precision | Recall | F-Measure | SAR |
|---|---|---|---|---|
| **Baseline** (*VP Position*) | 76.7 | 87.5 | 81.74 | 78.1 |
| **Baseline** (*VP Position*) + **Other Local Features** (*VP Length; Probability of head verbs being MVPs*) | 82.89 | 88.8 | 85.74 | 80 |
| **Baseline** (*VP Position*) + **Other Local Features** (*VP Length; Probability of head verbs being MVPs*) + **One Contextual Feature** (*Pattern Features*) | 90.23 | 89.66 | 89.94 | 85.1 |
| **Baseline** (*VP Position*) + **Other Local Features** (*VP Length; Probability of head verbs being MVPs*) + **Contextual Features** (*Pattern Features; Anti-features*) | 93.6 | 92.1 | **92.8** | **88.6** |

1) Local features improved the performance by 4%. One of the problems with local features is data sparsity, because the real value feature, that is, "Probability of head verbs being MVPs" is estimated from the whole training data. There are occasions when verbs in the testing corpus have not been encountered before. Thus, we will investigate the use of smoothing technology in future research.

2) Pattern features of the contextual type are very useful for MV identification and here increased the performance further by 4.2% from 85.74% to 89.94%. The lexicalized contextual features like "的", punctuation like "《》、" really helps to improve the performance.

3) Anti-features also contributed about 3% to the performance based on a comparison of the results obtained with and without anti-features. The reason is that anti-features can exclude VPs that have no chance of being MVPs.

We also provide the SAR results in the table. However, they are not comparable essentially because of different test data and amounts of data used in other works. The above results show that the SVM provides a flexible statistical framework for incorporating a wide variety of knowledge, including local and contextual features, for MVP identification.

After we tested the impact of different features on the performance of MVP identification, we wanted to know whether our annotated corpus was large enough to achieve acceptable performance. We used the best feature set according to the results of the above experiment. The performance achieved showed that the current training size had almost reached the saturation point.

## 4.3 Impact of Chunk Information on MVP Identification

Since we annotate MVPs based on chunk levels, we wanted to know how this shallow syntactic information affected the MVP identification. So we devised the following two experiments.

1. We firstly tested the performance of main verb identification based on POS, which does not include any shallow syntactic information. We stripped all the chunk tags in the corpus and used a simple rule to tag the predicate verb based on MVP chunks. That is, the headword of MVP chunk is the main verb of the sentence. For example, a) sentence is mapped to b) in the following.

a) [NP 公园/n(gong1yuan2)] **[MVP 时时/d(shi2shi2) 梦想/v(meng4xiang3) 着/u(zhe)]** [VP 有/v(you3)] [NP 条件/n(tiao2jian4)] [VP 繁育/v(fan2yu4) 出/v(chu1)] [NP 小√a(xiao3) 虎/n(hu3)] 。

b) 公园/n 时时/d 梦想/v_$ 着/u 有/v 条件/n 繁育/v 出/v 小√a 虎/n 。

*(People in the park have always dreamed that it is possible to breed tigers)*

Here /v_$ denote "梦想" (dream) is the main verb of sentence because it is the head verb of [MVP 时时/d 梦想/v_$ 着/u] (always dream of).

In this way, from the MVP training and testing corpus, we got the training and testing corpus with main verbs tagged. In our algorithm, we use the correct part-of-speech tags as input for main verb identification.

When we identify main verbs based on part-of-speech tags, all the features except for VP length are mapped to verb features. For example, the feature "VP position" is mapped to "verb position" an so on. The feature "Probability of verbs being MVPs" is revised to obtain the following formulation:

$$\mathrm{MV\_P}(X) = \frac{\mathrm{C(x\ is\ main\ verb)}}{\mathrm{C}(x)},$$

where C(x is the main verb) is the number of occurrences of the verb x as the main verb, and C(x) is the total occurrences of verb x in the training data.

"Only lexical type" among the frame anti-pattern features shown in Table 4 is also modified to use part-of-speech tags without chunk tags. The others are not modified since they have general chunk information, such as [SP */*], which cannot be directly converted to POS.

2. We also stripped all the chunk tags in the corpus, but this time we used our chunk system [Li *et al.* 2004] to re-chunk the data based on part-of-speech tags. Our chunk system is built on HMM. TBL-based error correction is used to further improve chunking performance. The average chunk length was found to be about 1.38 tokens and the F measure of chunking reached 91.13%. Inevitably, our chunk system will incur errors. Based on this noisy data, we use the same feature set to identify the MVPs. In this experiment, we wanted to know how the chunk errors would affect the MVP identification results.

The experimental results for the above two cases are shown in the Table 7.

***Table 7. Impact of chunk information***

| Model | Precision | Recall | F-Measure |
|---|---|---|---|
| POS | 84.98 | 84.04 | 84.5 |
| POS+Chunk1 | 88.56 | 89.27 | 88.9 |
| POS+Chunk2 | 93.6 | 92.1 | 92.8 |

The *POS* model row shows the first set of experiment results discussed above, that is, the results of identifying main verbs without using any chunk information. The *POS+Chunk1* model row shows the second set of experimental results: identifying MVPs with noisy chunk

information. The *POS+Chunk2* model row shows the results of identifying MVP with correct chunk information.

As the above table shows, the model trained on part-of-speech tags had the worst performance. This is because that the model lacks both chunk length information and part of the frame anti-pattern information. For example, if VP chunk is available, the VP chunk length can be calculated. Thus, the observation that VPs are precise (more than 85%) to be MVPs when their lengths are longer than 4 can improve the performance of main verb identification. Further more, the model trained on part-of-speech tags tends to tag the first verb as the main verb.

We performed the error analysis on results of POS+Chunk1 model. We wanted to see how many errors resulted in chunking errors in the table 8 below.

***Table 8. Error types for the POS+Chunk1 Model***

| **Error Types** | **Total Number** | **Caused By Chunk Errors** | **Caused By MVP Tag Errors** |
|---|---|---|---|
| Miss | 242 | 80 (33.1%) | 162 (66.9%) |
| False | 260 | 88 (33.8%) | 172 (66.2%) |
| All | 502 | 168(33.5%) | 334 (66.5%) |

In the table, Chunk Error means errors are caused by the chunk output, such as over-combining, under-combining etc. MVP Tag Error means we have correct chunks but MVP tagging is incorrect. It can be seen that more than one-third of the errors are caused by the chunk errors. The POS+Chunk2 model had the best performance since it uses shallow syntactic information and no errors appear in chunks.

## 5. Error Analyses and Discussion

The errors appearing in test data fall into the following categories.

### 5.1 Ambiguity of VP in Subject

Disambiguating VPs in subjects and predicates is a difficult problem. Since main verb identification is not based on syntactic and semantic parsing, we can only find the surface features of sentences. Thus, while the current algorithm correctly handles Example 25 and Example 27, it fails to handle Example 26 and Example 28.

Example 25 can be handled because the VP length feature helps. However, in some cases, the VP length will not help. Example 26 is a typical sentence in which the MVP should be “提醒” (ti2xing3, remind). The whole phrase “[SP 街上/s(jie1shang4)]…[NP 爆竹声/n(bao4zhu2sheng1]” (The sound of firecrackers …in the street) acts as the subject of “提醒”

(remind). The double objects of the main verb are "我" (I) and "[TP １日/t] [VP 是/v] [TP 新年/t]"(January the first is a new year's day). Both of the VPs in the subject are longer than the VP "提醒" (remind). Although we can exclude the second VP as the MVP (the pattern feature "VP+的" helps), it is rather difficult to exclude the first VP simply based on surface information. What leads to more ambiguity is "是" (is) in the object which also has a large probability of being a MVP. From the above analysis, it is currently difficult for our algorithm to detect "提醒" (remind) as a MVP.

***Example* 25**

**Correct:**

**[VP 没有/v(mei3you3)]** [NP 这/r(zhe4) 点/q(dian3) 精神/n(jing1shen2)] **[MVP 就/d(jiu4) 不/d(bu2) 配/v(pei4)]** [NP 电力/n(dian4li4) 人/n(ren2)] [NP 这/r(zhe4)] [ADJP 光荣/a(guang1rong2)] 的/u(de) [NP 称号/n(cheng1hao4)] 。/ww

*(Without that spirit, you will **not deserve to** have the glorious title "electronic people".)*

This example can be handled because the VP length feature helps.

***Example* 26**

**Correct:**

[SP 街上/s(jie1shang4)] **[VP 不时/d(bu4shi2) 地/u(de) 响起/v(xiang3qi3)]** [MP 一阵阵/m(yi2zhen4zhen4)] [PP 在/p(zai4)] [NP 北京/ns(bei3jing1)] **[VP 已/d(yi3) 听/v(ting1) 不/d(bu4) 到/v(dao4)]** 的/u(de) [NP 爆竹声/n(bao4zhu2sheng1] **[MVP 提醒/v(ti2xing3)]** [NP 我/r(wo3)] [TP １日/t(ri4)] **[VP 是/v(shi4)]** [TP 新年/t(xin1nian2)] 。/ww

*(The sound of the firecracker in the street every now and then, which haven't been heard already in Beijing, **remind** me that January the first is a new year's day.)*

**System Output:**

[SP 街上/s] **[MVP 不时/d 地/u 响起/v]** [MP 一阵阵/m] [PP 在/p] [NP 北京/ns] **[VP 已/d 听/v 不/d 到/v]** 的/u [NP 爆竹声/n] **[VP 提醒/v]** [NP 我/r] [TP １日/t] **[VP 是/v]** [TP 新年/t] 。/ww

In Example 27 and Example 28, since there is not enough information to determine that "是" is not in the object of "承认", the algorithm fails to find that "是" is the main verb.

***Example* 27**

**Correct:**

[NP 各方/r(ge4fang1)] **[MVP 承认/v(cheng2ren4)]** [NP 波黑/ns(bo1hei1)] **[VP 是/v(shi4)]** [MP 一个/m(yi1ge4)] [ADJP 统一/a(tong3yi1)] 的/u(de) [NP 主权/n(zhu3quan2) 国家/n(guo2jia1)] ，/ww

*(Each side **admits** that Bosnia-Herzegovena is a unified, sovereign country.)*

This example can be handled because the VP position helps.

***Example* 28**

**Correct:**

**[VP 承认/v(cheng2ren4)]** [NP 错误/n(cuo4wu4)] **[MVP 是/v(shi4)]** [MP 一/m(yi1) 种/q(zhong3)] [NP 好/a(hao3) 习惯/n(xi2guan4)] 。/ww

*(It is a kind of good habit to be able to acknowledge making mistakes.)*

**System Output:**

**[MVP 承认/v]** [NP 错误/n] **[VP 是/v]** [MP 一/m 种/q] [NP 好/a 习惯/n] 。/ww

## 5.2 Long Adjective Modifier

In Chinese parsing, the left boundary of "的" is a typical ambiguity problem. This problem also arises in main verb identification. The algorithm falsely identifies VPs in adjective modifiers as MVPs. See the following examples.

***Example 29***

**Correct:**

**[VP 积淀/v(ji1dian4)]** [PP 在/p(zai4) 大众/n(da4zong4) 血液/n(xue4ye4) 中/f(zhong1)] 的/u(de) [NP 传统/n(chuan2tong3) 文化/n(wen2hua4) 基因/n(ji1yin1)] [ADVP 也/d(ye3)] [PP 在/p(zai4) 传承/v(chuan2cheng2) 中/f(zhong1)] **[MVP 发生/v(fa1sheng1)]** [NP 种种/q(zhong3zhong3) 变异/n(bian4yi4)] 。/ww

*(The genes of the traditional culture which have been settling in the blood of the masses **undergo** various mutations when passing on.)*

**System Output:**

**[MVP 积淀/v]** [PP 在/p 大众/n 血液/n 中/f] 的/u [NP 传统/n 文化/n 基因/n] [ADVP 也/d] [PP 在/p 传承/v 中/f] **[VP 发生/v]** [NP 种种/q 变异/n] 。/ww

Note: The main verb of the whole sentence should be "发生". The verb phrase "[VP 积淀/v] [PP 在/p 大众/n 血液/n 中/f]" acts as a pre-modifier of the head noun "[NP 传统/n 文化/n 基因/n]". Thus, "积淀" should not be identified as a MVP in the whole sentence.

***Example* 30**

**Correct:**

[PP 于/p(yu2)] [TP 7月/t(qi1yue4) 5日/t(wu3ri4)] **[MVP 作出/v(zuo4chu1)]** **[VP 确定/v(que4ding4)]** [NP 肇事人/n(zhao4shi4ren2)] [NP 张/nr(zhang1) 成聚/nr(cheng2ju4)] **[VP 负/v(fu4)]** [NP 事故/n(shi4gu4) 全部/m(quan2bu4) 责任/n(ze2ren4)] ，/w [NP 受害人/n(shou4hai4ren2)] [NP 张/nr(zhang1) 平/nr(ping2)] **[VP 不/d(bu2) 负/v(fu4)]** [NP 责任/n(ze2ren4)] 的/u(de) [NP 交通/n(jiao1tong1) 事故/n(shi4gu4) 责任/n(ze2ren4) 认定书/n(ren4ding4shu1)] 。/ww

*(On July 5$^{th}$, the officer wrote the Traffic Accident Responsibility Assertion Book, in which the traffic troublemaker, Zhang Chenju, takes all the responsibility while the victim, Zhangpin is not responsible.)*

**System Output:**

[PP 于/p] [TP 7月/t 5日/t] **[MVP 作出/v]** **[VP 确定/v]** [NP 肇事人/n] [NP 张/nr 成聚/nr] **[VP 负/v]** [NP 事故/n 全部/m 责任/n] ，/ww [NP 受害人/n] [NP 张/nr 平/nr] **[MVP 不/d 负/v]** [NP 责任/n] 的/u [NP 交通/n 事故/n 责任/n 认定书/n] 。/ww

Note: The main verb of the whole sentence is "作出". The adjective modifier of NP "交通/n 事故/n 责任/n 认定书/n" (Traffic Accident Responsibility Assertion Book ) consists of two sub-sentences, in which the VPs "负" (take) and "不负" (not take) act as main verbs. Thus, "负" (take) and "不负" (not

take) should not be annotated as MVPs in the whole sentence.

## 6. Application

Labeling sentence boundaries is a prerequisite for many natural language processing tasks, including information extraction, machine translation etc. However, as Yu and Zhu [2002] pointed out, the problem is that "We have discussed a lot of word segmentation problems. But limited work has been done on Chinese sentence segmentation and it is still a difficult problem for computers." Without predicate information, it is difficult to predict sentence boundaries. Thus, we first identify the main verb and then label the sentence boundaries. The tagged results of simple sentence boundary labeling are like the following examples.

***Example* 31**

[NP 母爱/n(mu3ai4)] , /w [MVP 作为/v(zuo4wei2)] [NP 人类/n(ren2lei4)] [MP 一/m(yi1) 种/q(zhong3)] [ADJP 崇高/a(chong2gao1)] 的/u(de) [NP 爱/vn(ai4)] ，/ww [MVP 是/v(shi4)] [MP 一/m(yi4) 棵/q(ke1)] [NP 人类/n(ren2lei4) 精神/n(jing1shen2) 大树/n(da4shu4)] ，/ww [NP 她/r(ta1)] [MVP 永久/d(yong3yuan3) 地/u(de) 枝繁叶茂/i(zhi1fan2ye4mao4)] 。/ww

*(Mother's love is a kind of lofty love of mankind. It is a big tree of the human spirit. It will have a permanent foundation with luxuriant foliage and spreading branches.)*

There are three simple sentences in this example. Our task is to use MVP information to break the sentence up into simple sentences. Here, when we refer to a "sentence," we mean a verb-predicate sentence. Since there are no MVPs in non- verb-predicate sentence, we cannot use the MVP information to break up these sentences.

We compared two sentence-breaking models. First, in the base line, we tagged all the commas as sentence ending punctuation if the sentences had at least one VP. Second, we tagged all the commas as sentence ending punctuation if the sentences had MVPs. This was an end-to-end evaluation because MVP identification was used as preprocessing step before sentence breaking.

The evaluation metrics we used were as follows:

$$\text{Precision} = \frac{\#\text{ of correct sentence} - \text{stoppunc.in system output}}{\#\text{ of sentence} - \text{stoppunc.in system output}};$$

$$\text{Recall} = \frac{\#\text{ of correct sentence} - \text{stoppunc.in system output}}{\#\text{ of sentence} - \text{stoppunc.in answer}};$$

$F = 2 * P * R/(P + R)$ .

***Table 9. Performance of Chinese Sentence Breaker***

| **Model** | **Precision** | **Recall** | **F** |
|---|---|---|---|
| Baseline | 86.74 | 94.57 | 90.48 |
| Tag with MVP | 94.22 | 91.34 | 92.76 |

From the above table, one can see that the simple sentence breaker improved the performance by about 2.4% with the help of MVP identification. Errors in the tagging of stop-punctuation were mostly caused by errors in the tagging of MVPs.

## 7. Conclusion and Future Work

Main verbs are useful for dependency parsing, sentence pattern identification, and Chinese sentence breaking. Chinese linguists have done research on predicate-verbs for a long time and provided a grammatical view of analyzing the Chinese sentence. However, automatically identifying main verbs is quite another problem. Most of the previous works by computational linguists have focused on the identification process instead of the definition of a main verb. In this paper, we have discussed in detail the whole process of automatically identifying Chinese main verbs from specification to realization.

The contributions of our work are as follows.

1) We have thoroughly investigated main verbs from both linguists' point of view and the computational point of view. Based on this investigation, we have presented our specification as well as a corpus annotation method. The advantage of our specification is that the main verbs of different verb-predicate sentences are included. More specific and reliable knowledge is applied in our main verb definition. Various complicated cases have been studied, and abundant examples from real text have been provided.
2) We have presented our results of identifying main verbs based on chunking levels. The experimental results show that the performance of our approach is better than that of the

approach based on part-of-speech tags. We have also proposed an end-to-end evaluation based on the use of a Chinese simple sentence breaker.

3) New local and contextual features investigated in our specification and statistics have been incorporated into our identification algorithm and used to achieve promising results.

In future work, we would like to find more effective features from lexical knowledge and solve the data sparse problem that is encountered in feature selection. We also are interested in developing more applications based on MVP information, such as an application for extracting the verb-subject or verb-object dependency relations.

**Acknowledgements**

We would like to thank Prof. Jianyun Nie, Dr. John Chen, Dr. Ming Zhou, and Dr. Jianfeng Gao for their valuable suggestions and for checking the English in this paper. We thank the anonymous reviewers of this article for their valuable comments and criticisms. We would also like to thank Juan Lin for her help with our specification design and corpus annotation.

## References

Banko, M. and E. Brill, "Scaling to very very large corpora for natural language disambiguation," *Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics,* 2001, pp. 26-33.

Chen, X.H. and D.Y. Shi, "To Mark Topic and Subject in Chinese Sentences," *Proceedings of 4th National Computational Linguistics,* Tsinghua University Press, 1997m, pp. 102-108.

Ding, S.S., S.X. Lv, R. Chen, D.X. Sun, X.C. Guan, J. Fu, S.Z. Huang and Z.W. Chen, "xiandai hanyu yufa jianghua," (Modern Chinese Grammar Talk), The Commercial Press, 1961.

Fan, X., "sange pingmian de yufa guan," (The Grammar View of Three Levels), Publishing house of Beijing Language Institute, 1995.

Gong, X.J., Z.S. Luo and W.H. Luo, "Recognizing the Predicate Head of Chinese Sentences," *Journal of Chinese Information Processing,* vol. 17, no. 2, 2003, pp. 7-13.

Hong, X.H., "hanyu cifa jufa chanyao," (The Brief Introduction to Chinese Word and Syntax), Jilin People's Press, 1980.

Huang, Z.K., "xiandai hanyu changyong jushi," (The Daily Sentence Pattern of Modern Chinese), People's Education Publishing House, 1987.

Isozaki, H. and H. Kazawa, "Efficient Support Vector Classifiers for Named Entity Recognition," P*roceedings of the 17th International Conference on Computational Linguistics,* Taipei, Taiwan, 2002, pp. 390-396.

Jin, L.X. and S.Z Bai, "The Characteristics of Modern Chinese Grammar and the Research Standards," *Chinese Language Learning,* no. 5, Oct, 2003, pp. 15-21.

Joachims, T., "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning," Schölkopf B., C. Burges and A. Smola (editor), MIT-Press, 1999, pp.169.

Lai, T. B. Y. and C.N. Huang, "Dependency-based Syntactic Analysis of Chinese and Annotation of Parsed Corpus," *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics,* Hong Kong, China, 1-8 October 2000, pp. 255-262.

Li, H.Q., C.N. Huang, J.F. Gao and X.Z. Fan, "Chinese Chunking with Another Type of Spec," *The Third SIGHAN Workshop on Chinese Language Processing,* Barcelona, July 24-26, 2004, pp. 41-48.

Li, H.Q., C.N. Huang, J.F. Gao and X.Z Fan, "The use of SVM for Chinese new word identification," *The First International Joint Conference on Natural Language Processing.* March 22-24, 2004, pp. 497-504.

Liu, Y.H., W.Y. Pan and W. Gu, "xiandai hanyu shiyong yufa," (Practical Chinese Grammar Book), The Commercial Press. 2002.

Liu, Q. and S.W. Yu, "Discussion on the Difficulties of Chinese-English Machine Translation," *International Conference on Chinese Information Processing,* January 1998, pp.507-514.

Luo, Z.S., C.J. Sun and C. Sun, "An Approach to the Recognition of Predicates in the Automatic Analysis of Chinese Sentence Patterns," *Proceedings of 3th National Computational Linguistics,* 1995, pp. 159-164.

Lv, S.X., "xiandai hanyu babaici," (Eight Hundred Words of Modern Chinese) The Commercial Press, 1980.

Lv, S.X. and Q.Z. Ma (editor), "yufa yanjiu rumen," (Elementary Study of Chinese Grammar). The Commercial Press. 1999.

Marcus, M., B. Santorini and M. Marcinkiewicz, "Building a Large Annotated Corpus of English: the Penn Treebank," *Computational Linguistics*, 19(2), 1993, pp. 313-330.

Meng, C., H.D. Zheng, Q.H. Meng and W.L. Cai, "hanyu dongci yongfa cidian," (The Chinese Verb Usage Dictionary). The Commercial Press. 2003.

Nakagawa, T., T. kudoh and Y. Matsumoto, "Unknown Word Guessing and Part-of-speech Tagging Using Support Vector Machines," P*roceedings of the 6th NLPRS,* 2001, pp. 325-331.

Sui, Z.F. and S.W. Yu, "The Research on Recognizing the Predicate Head of a Chinese Simple Sentence in EBMT," *Journal of Chinese Information Processing,* vol.12, no. 4, 1998a, pp. 39-46.

Sui, Z.F. and S.W. Yu, "The Acquisition and Application of the Knowledge for Recognizing the Predicate Head of a Chinese Simple Sentence," *Journal (Natural Sciences) Of Peking University,* vol. 34, no. 223, 1998b, pp. 221-230.

Vapnik, V. N., "The nature of Statistical Learning Theory," Springer, 1995.

Xia, F., M. Palmer, N.W. Xue, M.E. Okurowski, J. Kovarik, F.D. Chiou, S.Z. Huang, T. Kroch and M. Marcus, "Developing Guidelines and Ensuring Consistency for Chinese Text Annotation," *Proceedings of the second International Conference on Language Resources and Evaluation (LREC 2000),* Athens, Greece, 2000, pp. 3-10.

Xue, N.W. and F. Xia, "The Bracketing Guidelines for the Penn Chinese Treebank," http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf. (3.0). Oct. 2000.

Yu, S.W. and X.F. Zhu, "Chinese Information Processing and Its Methodology," Applied Linguistics, May 2002, pp. 51-58.

Yu, S.W., H.M. Duan, X.F. Zhu and B. SWEN, "The Specification of Basic Processing of Contemporary Chinese Corpus," *Journal of Chinese Information Processing,* vol.16, issue 5, pp. 49-64 & issue 6, pp. 58-64, 2002.

Zhang, Z.G., "xiandai hanyu," (Modern Chinese,Volume Two), People's Education Publishing House, 1982.

Zhou, M., "J-Beijing Chinese-Japanese Machine Translation System," *1999 Joint Symposium on Computational Linguistics (JSCL-1999).* Tsinghua University Press, 1999, pp. 312-319.

Zhu, D.X., "yufa jiangyi,"(Grammar Tutorial), The Commercial Press, 1982.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 95-122 


# Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria

**Thomas C. Chuang[*] and Kevin C. Yeh[+]**

**Abstract**

We present a new approach to aligning sentences in bilingual parallel corpora based on punctuation, especially for English and Chinese. Although the length-based approach produces high accuracy rates of sentence alignment for *clean* parallel corpora written in two Western languages, such as French-English or German-English, it does not work as well for parallel corpora that are noisy or written in two disparate languages such as Chinese-English. It is possible to use cognates on top of the length-based approach to increase the alignment accuracy. However, cognates do not exist between two disparate languages, which limit the applicability of the cognate-based approach. In this paper, we examine the feasibility of exploiting the statistically ordered matching of punctuation marks in two languages to achieve high accuracy sentence alignment. We have experimented with an implementation of the proposed method on parallel corpora, the Chinese-English Sinorama Magazine Corpus and Scientific American Magazine articles, with satisfactory results. Compared with the length-based method, the proposed method exhibits better precision rates based on our experimental reuslts. Highly promising improvement was observed when both the punctuation-based and length-based methods were adopted within a common statistical framework. We also demonstrate that the method can be applied to other language pairs, such as English-Japanese, with minimal additional effort.

**Keywords:** Sentence Alignment, Cognate Alignment, Machine Translation

[*] Department of Computer Science, Vanung University, No. 1 Van-Nung Road, Chung-Li Tao-Yuan, Taiwan, ROC
E-mail: tomchuang@cc.vit.edu.tw
[+] Department of Computer Science, National Tsing Hua University, 101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC

## 1. Introduction

Bilingual corpora are very important for building natural language processing systems [Moore 2002; Gey *et al.* 2002], including data-driven machine translation [Dolan *et al.* 2002], computer-assisted revision of translations [Jutras 2000], and cross-language information retrieval [Chen and Gey 2001]. In order to develop NLP systems, it is useful to align bilingual corpora at the sentence level with very high precision [Moore 2002; Chuang *et al.* 2002, Kueng and Su 2002]. With aligned sentences, further analysis such as phrase and word alignment analysis [Ker and Chang 1997; Melamed 1997], bilingual terminology [Déjean *et al.* 2002] and collocation [Wu 200] extraction analysis can be performed. Yang, C, Li, K. [2003] proposed an alignment method for bilingual title pairs on the Web for automatic generation of bilingual parallel corpora. The hybrid dictionary approach [Collier *et al.* 1998], text-based alignment [Kay and Röscheisen 1993], part of speech-based alignment [Chen and Chen 1994], and the lexical method [Chen 1993] are other examples of sentence alignment methods. While these methods presume little or no prior knowledge of source and target languages, they are relatively complex and require significant amounts of parallel text and language resources.

Much work reported in the computational linguistics literature has focused on aligning English-French and English-German sentences. While the length-based approach [Gale and Church 1993; Brown *et al.* 1991] to sentence alignment produces surprisingly good results for French and English with success rates well over 96%, it does not work well for the alignment of English and Chinese sentences. Simard, Foster, and Isabelle [1992] proposed using cognates on top of the length-based approach to improve the alignment accuracy. They use an operational definition of cognates, which include digits, alphanumerical symbols, punctuation, and alphabetical words. Several other measures of cognateness have also been suggested [Melamed 1999; Danielsson and Muhlenbock 2000; Ribeiro *et al.* 2001], but none of them are sufficiently reliable, and all of them are tailored to close Western language pairs.

Simard, Foster, and Isabelle [1992] pointed out that cognates in two close languages, such as English and French, can be used to measure the likelihood of mutual translation. Those cognates include alphabetic words, numeric expressions, and punctuation that are almost identical and readily recognizable by computers. However, for disparate language pairs, such as Chinese and English, that lack a shared Roman alphabet, it is not possible to rely on such cognates to achieve high-precision sentence alignment of noisy parallel corpora.

Research on sentence alignment of English and Chinese texts [Wu 1994], indicates that the lengths of English and Chinese texts are not as highly correlated as they are with French and English, leading to lower precision rates (86.4-95.2%) for length based aligners. A comparison of the correlation between German-English and Chinese-English bilingual corpora is depicted in Figure 1, where 138 German- English and 151 Chinese-English aligned

sentences are analyzed. The correlations are 0.99 and 0.95 for the German-English and Chinese-English cases, respectively. The expected ratios and the corresponding standard deviations are (0.92, 0.1124) and (4.614, 0.84) for the German-English and Chinese-English cases, respectively.

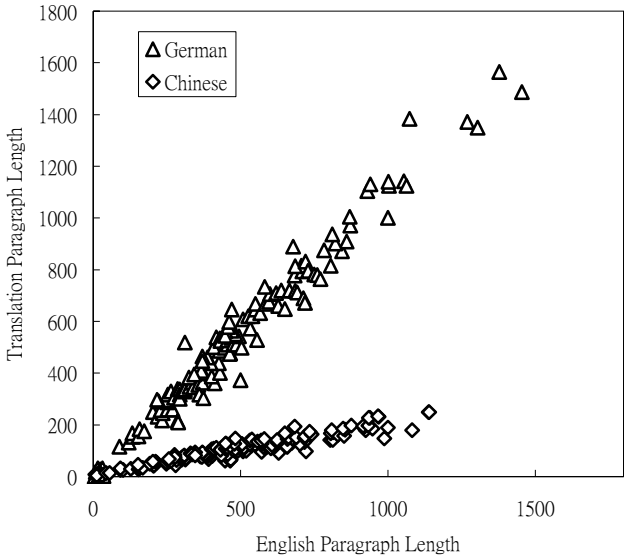



***Figure 1. The relationships between German-English [Gale and Church 1993] and Chinese-English bilingual paragraph lengths [Chuang et al. 2002]. The correlations are 0.99 and 0.95 for the German-English and Chinese-English cases, respectively. The expected ratios and the corresponding standard deviations are (0.92, 0.1124) and (4.614, 0.84) for the German-English and Chinese-English cases, respectively.***

Furthermore, for English-Chinese alignment tasks, no orthographic, phonetic, or semantic cognates, that are readily recognizable by computer exist. Therefore, the inexpensive cognate-based approach is not applicable to Chinese-English tasks. We are thus motivated to find alternative evidence that two blocks of texts are mutual translations. It turns out that punctuation can be telling evidence, if we do more than hard matching of punctuation and take into consideration intrinsic sequencing and the statistical distribution of punctuation in ordered comparison. What is attractive about this approach is that it easily leads to sub-sentential alignment, which has been shown to be useful for statistical machine translation.

Section 2 of this paper, we provide some information about the similarity in the use of punctuation in Chinese and English literature and also the differences. Our conclusion is that using punctuation as cognates to align disparate parallel texts will fail to provide adequate alignment results. Section 3, we define a punctuation compatibility factor as an indicator of mutual translation. A translation model that employs a punctuation probability function is proposed. In Section 4, we present experiments based on our novel approach of using the statistical properties of punctuation in parallel texts being analyzed to perform bilingual sentence alignment. We demonstrate that one can use punctuation alone to develop a high-precision sentence alignment program for distant parallel texts like those in Chinese-English corpora. Additionally, we examine the performance of sentence alignment by using punctuation in combination with length. In Section 5, we demonstrate that the proposed method is a very cost effective approach that can be effectively applied to other disparate bilingual languages like English-Japanese without *a priori* language knowledge of them. A brief conclusion is provided in Section 6.

## 2. Punctuation across languages

According to the Longman Dictionary of Applied-Linguistics [Richards *et al.* 1985], a *cognate* is "a word in one language which is similar in form and meaning to a word in another language because both languages are related." Although the ways in which different languages around the world use punctuation vary, symbols such as commas and full stops are used in most languages to demarcate writing, while question and exclamation marks are used to show interrogation and emphasis. However, these forms of punctuation can often look different or be used in different ways.

The traditional Chinese writing system does not have punctuation, and it is up to the reader to demarcate the text while reading. With the influx of Western culture in the eighteenth century, punctuation systems similar to the one used with Roman script was adopted in China and Japan. The punctuation includes the period, comma, colon, dash, etc. Although most of those forms of punctuation look similar to Roman ones, they are usually coded as double-bytes and tend to be used differently. The full stop in Chinese and Japanese is a small empty circle, quite different in appearance from the Roman period. Quotes are also very different, shaped like a Greek letter Γ, upright or upside down. There are forms of punctuation that have no counterparts in Roman text. For instance, " 、 " is the pause symbol, which is used somewhat like the comma but only when separating items in a list. On the other hand, there are several uses of the Roman comma which do not occur in Chinese texts. A few examples are given below:

(Parenthetical expressions)

(1e) Evolution, as far as we know, doesn't work this way.

(1c) 我們所知道的進化論不是如此的。

(Appositives)

(2e) His father, Tom, is a well-known scholar.

(2c) 他的父親湯姆是一位有名的學者。

Yang [1981] described more punctuation marks in Chinese used in various ways that are similar or dissimilar to English punctuation. In summary, although Simard et al. [1992] considered the various forms of punctuation in English and French to be cognates, in general, punctuation forms are not cognates for many other language pairs.

In both Chinese and English texts, the average ratio of the punctuation count to the total number of tokens available is low (less than 15%). But punctuation provides valid additional evidence, which can help one achieve a high degree of alignment precision. Our method can easily be generalized to other language pairs since minimal a priori linguistic knowledge is required.

## 3. Punctuation and Sentence Alignment

### 3.1 Punctuation Marks in English and Chinese

In this section, we will describe how punctuation in two languages can be used to measure the likelihood of mutual translation in sentence alignment. We will use an example in the following to illustrate the method. A formal description also follows:

Example 3 shows a Chinese sentence and its translation counterpart of two English sentences in a parallel corpus.

(3c) 逐漸的，打鼓不再能滿足他，「打鼓原是我最喜歡的，後來卻變成邊打邊睡，一個月六萬元的死工作」，薛岳表示。

(3e) Over time, drums could no longer satisfy him. "Drumming was at first the thing I loved most, but later it became half drumming, half sleeping, just a job for NT$60,000 a month," says Simon.

If we keep punctuations in the above examples in the original order and strip everything else out, we have ten pieces of punctuation from the English part (3e) and eight from the Mandarin part (3c) as follows:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **(4c)** | ， | ， | 「 | ， | | ， | 」 | ， | 。 |
| **(4e)** | , | . | " | , | , | , | , | " | . |

They can be arranged into different match types as shown below.

| Match type | (4c) | (4e) |
|---|---|---|
| 1-1 | ， | , |
| 1-1 | ， | . |
| 1-1 | 「 | " |
| 1-1 | ， | , |
| 0-1 | | , |
| 1-1 | ， | , |
| 2-2 | 」 ， | , " |
| 1-1 | 。 | . |

***Figure 2. The correspondence between two punctuation strings***

There are several frequently used punctuation forms in Chinese text that are not available in English text, for example, the punctuation forms "、" and "。". These punctuation forms often correspond to the English punctuation forms "," and ".", respectively. It is not difficult to see that the two punctuation strings above match up quite nicely, indicating that the corresponding texts are mutual translations. Roughly, the first two commas in Chinese correspond to the first two English punctuation marks (comma and period), while the Chinese open quote in the third position corresponds to the English open quote also in the third position. The two Chinese commas inside the quotes correspond to two of the four commas within the quotes in English. The two consecutive marks (」，) correspond to (,"), forming a 2-2 match. These correspondences can be unraveled via a dynamic programming procedure, much like sentence alignment. See Figure 2 for more details.

It is apparent that the punctuation in the two strings match up very consistently, and that the matching is somewhat continuous with respect to the alignment of regular words surrounding the punctuation (see the double-lined links in Figure 3 for details). Therefore, the example gives a convincing indication that the correspondence between punctuation across two languages can provide telling evidence that two texts are mutual translations.

## 3.2 Punctuation marks as Good Indicators of Mutual Translation

Based on our initial observation, the portion of the identifiable punctuation matches between two parallel texts in Chinese and English is over 50%. Examining Figure 2, we can identify institutively the matches between the Chinese punctuation and the equivalent English punctuation marks: (「) corresponds to (“), etc. This implies that although direct match information is useful, there is still a large discrepancy in the punctuation mappings between Chinese and English. We, therefore, define here a punctuation compatibility factor that can be used to further analyze the relationship between the punctuation found in parallel texts. The punctuation compatibility factor as an indicator of mutual translation is defined as

$$\gamma = \frac{c}{\max(n, m)}, \tag{1}$$

where $\gamma$ = the punctuation compatibility factor,
$c$ = the number of direct punctuation matches,
$n$ = the number of Chinese punctuation marks,
$m$ = the number of English punctuation marks.

We took aligned English-Chinese sentences that had the same punctuation count (which is the denominator of Equation 1), take ten for example, in order to determine how well punctuation works as an indicator of mutual translation of English and Chinese sentences. We also took the same English sentences and matched them up with randomly selected Chinese sentences to calculate the compatibility of punctuation marks in unrelated texts.

The results obtained indicated that the average compatibility of pairs of sentences, which were mutual translations, was about 0.67 (with a standard deviation of 0.170), while the average compatibility of random pairs of bilingual sentences was 0.34 (with a standard deviation of 0.167).

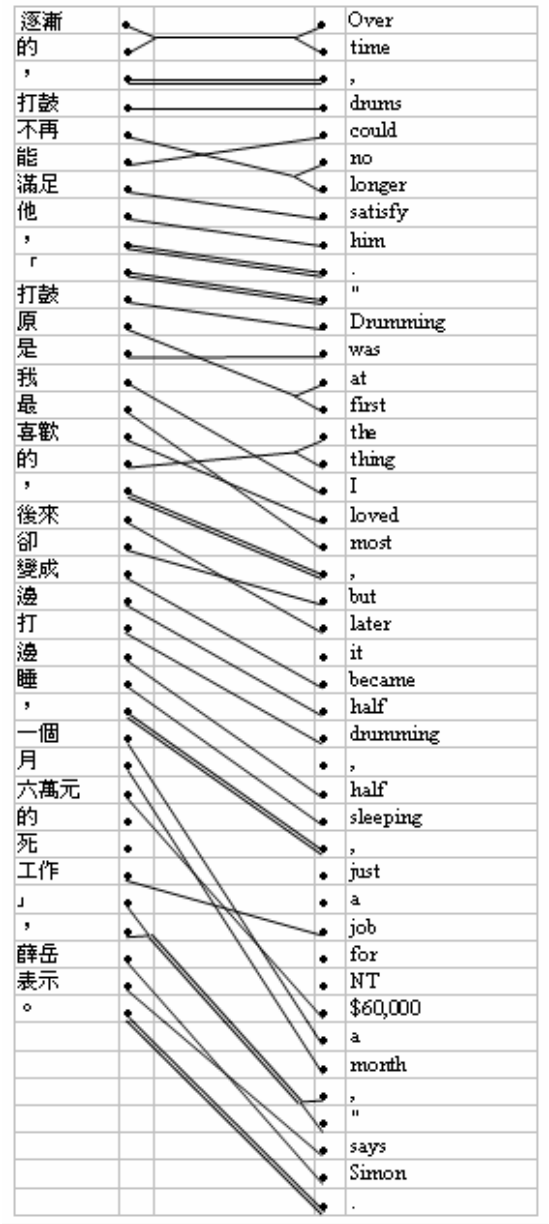



***Figure 3. English punctuation across aligned sentences***

Figures 4 through 6 show the compatibility results based on punctuation counts of eight, ten and twelve respectively. These graphs were constructed by analyzing around 50,000 aligned sentences found in the Sinorama Magazine (1990-2000). 521, 259, and 143 sentences were selected to obtain values of n and m equal to 8, 10, and 12, respectively. The solid lines simply connect data points for easier observation.

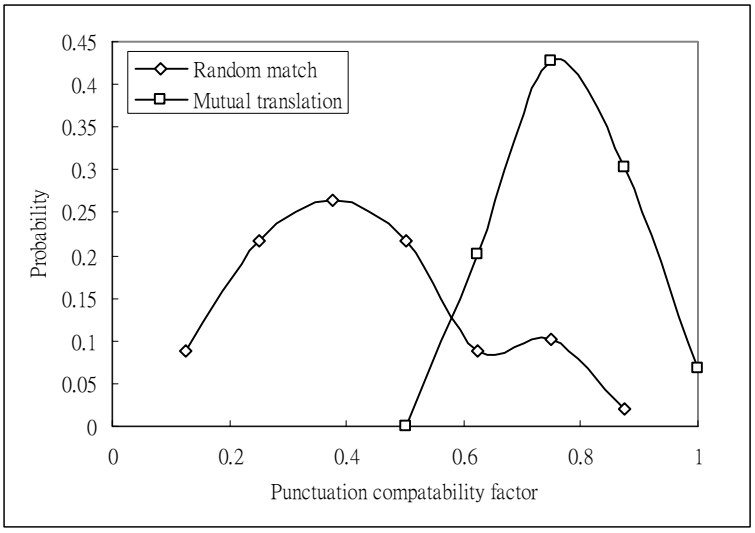



***Figure 4. Compatibility of translation pairs vs. random pairs with n=m=8***

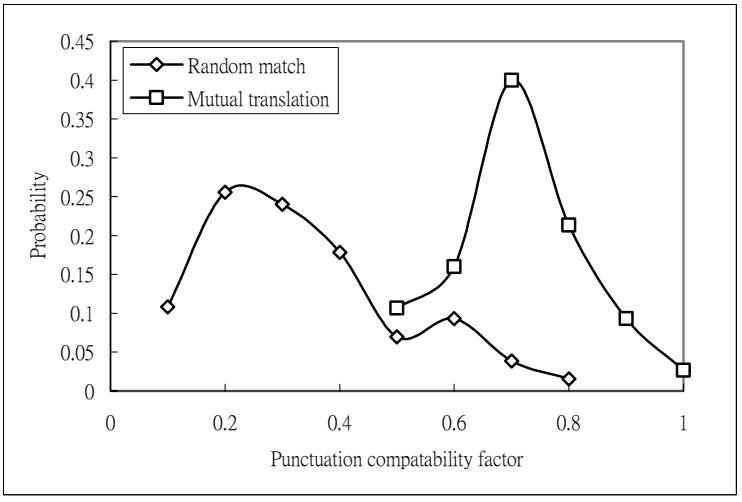



***Figure 5. Compatibility of translation pairs vs. random pairs with n=m= 10***

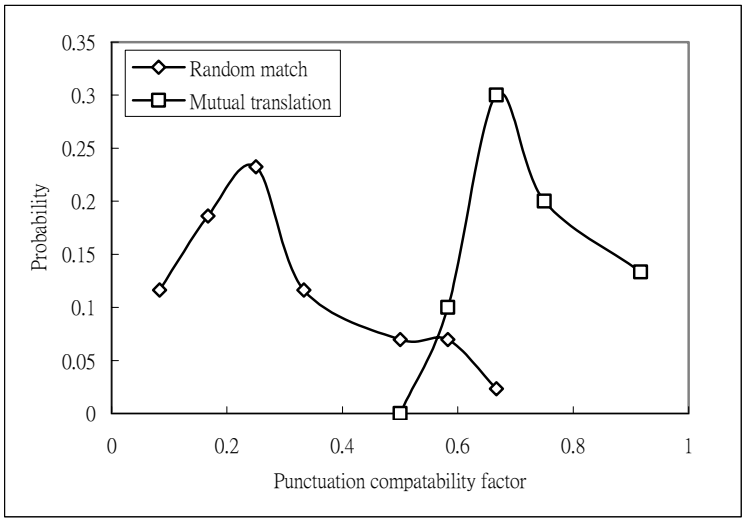



***Figure 6. Compatibility of translation pairs vs. random pairs with n=m = 12***

Intuitively, as the number of punctuation marks increases, the reliability of the compatibility function does also. Overall, if the punctuation marks are softly matched in ordered comparison across the two languages, they indeed provide useful information for effective sentence alignment. Analyzing the Sinorama corpus, we found that the percentage of matched sentences having the same number of punctuation marks was 21.42%. We selected and analyzed aligned sentences having different numbers of punctuation marks to get more insight into the distinction between matched and random sentences. The analysis also helped us to determine the proper use of the binomial distribution function for sentence alignment. Sentences with eight, ten, and twelve punctuation marks were arbitrarily chosen for analysis. It appears that the distinction between mutual translations and unrelated texts indeed becomes more prominent for sentences that have larger numbers of punctuation marks.

## 3.3 Punctuation Alignment Model

Instead of one-to-one hard matching of punctuation marks in parallel texts as used in the cognate approach of Simard et al. (1992), we allow no match and one-to-several matching of punctuation matches. Our model of the probability of punctuation alignment is very similar to the word alignment model proposed by Brown et al. (1991). In order to perform soft matching of punctuation, we define the probability that a sequence of punctuation marks $CP_{\mathrm{i}} = Cp_1 Cp_2 Cp_3 \cdots Cp_i$ in a sentence in the source language (L1) translates into a sequence of punctuation marks $EP_{\mathrm{i}} = Ep_1 Ep_2 Ep_3 \cdots Ep_i$ in a sentence in the target language (L2) as $P(EP_i, CP_j)$. We choose the punctuation alignment that maximizes the probability overall

possible alignments, given a pair of punctuation sequences corresponding to a pair of parallel sentences, i.e.,

$$\arg\max_{\mathrm{A}} P(\mathrm{A}|CP_i, EP_j) \quad ,$$

where A is a punctuation alignment. Assuming that the probabilities of the individually aligned punctuation pairs are independent and applying the Bayes' rule, we can make the following approximation:

$$P(EP_i, CP_j) \approx \prod_{A} P(Cp_k, Ep_k) \cdot P(|Cp_k|, |Ep_k|) \tag{2}$$

where $|Cp_k|$ and $|Ep_k|$ = the number of punctuation marks in $CP_i$ and $EP_j$ ,respectively, which ranges from 0 to 2,

$P(Cp_k, Ep_k)$ = the probability of translating $Cp_k$ into $Ep_k$ , and

$P(|Cp_k|, |Ep_k|)$ = the probability of translating $|Cp_k|$ punctuations in L1 into $|Ep_k|$ punctuation in L2.

We observe that in most cases, the links of punctuation do not cross each other, much like the situation with sentence alignment. Therefore, it is possible to use the dynamic programming procedure to softly match punctuation across languages.

In order to explore the relationship between punctuation in pairs of Chinese and English sentences that are mutual translations, we selected a small set of manually aligned texts and investigated the characteristics and the statistics associated with the punctuation. Information from around 500 manually analyzed sentences was then used as the initial parameters to bootstrap a larger corpus. An unsupervised EM algorithm and dynamic programming were used to optimize the punctuation correspondence between a text and its translation counterpart. The steps in the standard EM algorithm which we used included initializing model parameters with manually analyzed punctuation matching probabilities, assigning probabilities to missing punctuation, estimating model parameters from completed data, and iterating the process until convergence was reached. The EM algorithm converged quickly after the second iteration of training.

We observed that, in most cases, the links of punctuation did not cross each other, much like the situation with sentence alignment. Therefore, we were motivated to use the dynamic programming procedure to *softly match* punctuation across the languages by finding the Viterbi path using the punctuation translation function $P(Cp_k, Ep_k)$ and fertility function $P(|Cp_k|, |Ep_k|)$. The translation probability functions corresponding to 1-1, 2-2, 1-0, and 0-1 English-Chinese punctuation matches are shown in Tables 1 to 4, respectively. It should be noted that the calculated probability was the conditional probability of each punctuation mark, therefore, the sum of the probability in each table does not equal to one.

***Table 1. The frequency counts and the conditional probabilities of 1-1 English-Chinese punctuation matches, sorted according to count***

| $E_p$ | $C_p$ | Match type | Count | Prob. |
|---|---|---|---|---|
| , | ， | 1-1 | 541 | 0.809874 |
| . | 。 | 1-1 | 336 | 0.657528 |
| " | ﹁ | 1-1 | 131 | 0.34203 |
| . | ， | 1-1 | 113 | 0.221133 |
| " | ﹂ | 1-1 | 112 | 0.292423 |
| " | 「 | 1-1 | 65 | 0.16971 |
| " | 」 | 1-1 | 59 | 0.154044 |
| , | 、 | 1-1 | 56 | 0.083832 |
| , | 。 | 1-1 | 41 | 0.061377 |
| ! | ！ | 1-1 | 38 | 0.883508 |
| . | … | 1-1 | 30 | 0.058708 |
| ? | 。 | 1-1 | 17 | 0.447277 |
| : | ， | 1-1 | 12 | 0.666302 |
| ; | 、 | 1-1 | 11 | 0.422925 |
| , | 「 | 1-1 | 10 | 0.01497 |
| ? | ？ | 1-1 | 9 | 0.236794 |
| . | 、 | 1-1 | 7 | 0.013698 |
| " | ， | 1-1 | 7 | 0.018276 |
| ; | ， | 1-1 | 7 | 0.269134 |
| . | ； | 1-1 | 6 | 0.011742 |
| " | ： | 1-1 | 6 | 0.015666 |
| , | ： | 1-1 | 5 | 0.007485 |
| ? | ， | 1-1 | 5 | 0.131552 |
| , | ； | 1-1 | 4 | 0.005988 |
| . | — | 1-1 | 4 | 0.007828 |
| : | ： | 1-1 | 4 | 0.222101 |
| ; | 。 | 1-1 | 4 | 0.153791 |
| ) | ） | 1-1 | 4 | 0.997159 |
| , | · | 1-1 | 3 | 0.004491 |
| . | · | 1-1 | 3 | 0.005871 |
| . | ﹁ | 1-1 | 3 | 0.005871 |
| ! | 。 | 1-1 | 3 | 0.069751 |
| ? | — | 1-1 | 3 | 0.078931 |

**Table 2. The frequency counts and conditional probabilities of 2-2 English-Chinese punctuation matches, sorted according to count**

| $E_p$ | $C_p$ | Match Type | Count | Prob. |
|---|---|---|---|---|
| ," | 」’ | 2-2 | 6 | 0.956403 |
| ." | 」’ | 2-2 | 3 | 0.916449 |
| ?" | ——— | 2-2 | 2 | 0.611063 |
| .! | …！ | 2-2 | 1 | 0.785235 |
| !" | ）） | 2-2 | 1 | 0.785235 |
| ?" | 」。 | 2-2 | 1 | 0.305531 |
| ?? | ——— | 2-2 | 1 | 0.785235 |

**Table 3. The frequency counts and conditional probabilities of 1-0 English-Chinese punctuation matches, sorted according to count**

| $E_p$ | $C_p$ | Match Type | Count | Prob. |
|---|---|---|---|---|
| , | | 1-0 | 106 | 0.3655 |
| . | | 1-0 | 66 | 0.2276 |
| " | | 1-0 | 59 | 0.2034 |
| ) | | 1-0 | 23 | 0.0793 |
| ( | | 1-0 | 20 | 0.0691 |
| : | | 1-0 | 7 | 0.0241 |
| ? | | 1-0 | 5 | 0.0172 |

**Table 4. The frequency counts and conditional probabilities of 0-1 English-Chinese punctuation matches, sorted according to count**

| $E_p$ | $C_p$ | Match Type | Count | Prob. |
|---|---|---|---|---|
| | ’ | 0-1 | 229 | 0.389455 |
| | ——— | 0-1 | 58 | 0.098639 |
| | 。 | 0-1 | 52 | 0.088435 |
| | 」 | 0-1 | 50 | 0.085034 |
| | 、 | 0-1 | 45 | 0.076531 |
| | 「 | 0-1 | 41 | 0.069728 |
| | … | 0-1 | 39 | 0.066326 |
| | ？ | 0-1 | 14 | 0.02381 |
| | 「——— | 0-1 | 14 | 0.02381 |
| | ： | 0-1 | 9 | 0.015306 |
| | ——」 | 0-1 | 7 | 0.011905 |

The punctuation match types (also known as the fertility functions) obtained through training are summarized in Table 5. Notice that the counts shown in the table are not integers because the results of EM training were adjusted using the Good Turing Smoothing Method to improve them.

***Table 5. The punctuation fertility functions***

| Punctuation Match Type | Count | Probability |
|---|---|---|
| 0-1 | 588.0005 | 0.225027 |
| 1-0 | 286.001 | 0.109452 |
| 1-1 | 1698.076 | 0.649852 |
| 1-2 | 2.466198 | 0.000944 |
| 2-1 | 0.965034 | 0.000369 |
| 2-2 | 37.19216 | 0.014233 |

## 3.4 Punctuation-based Sentence Alignment Model

Unlike the method Simard et al. [1992] used to handle cognates, we model the *compatibility* of punctuation across two languages using the Binomial distribution. Each punctuation mark appearing in one language either has one to three punctuation counterparts across translation or does not. For each punctuation mark, the probability of it having a translation counterpart is independent with a fixed value of $p$. Our approach differs from Simard's in the following interesting ways:

1. We use the Binomial distribution, while Simard et al. used a likelihood ratio.
2. We go beyond hard matching of punctuation marks between parallel texts. We allow a punctuation mark in one language to match up with a number of compatible punctuation marks in another. The compatibility model is similar in structure to the lexical translation probability proposed by Brown et al. [1991].
3. We take into consideration the intrinsic sequencing of punctuation marks in an ordered comparison. A flexible and ordered comparison of punctuation is carried out via dynamic programming.

Following Gale and Church [1993], we employ the Bayes Theorem to estimate the likelihood of aligning two text blocks E and C by calculating P($E$, $C$|*match*) P(*match*).We adopt the same dynamic programming method, but use punctuation marks to measure the likelihood of mutual translation instead of lengths. The proposed sentence alignment method is based on a model in which each punctuation mark in L1 is responsible for generating a number of punctuation marks with a given matching probability in L2.

We define the probability of mutual translation for a given alignment pattern P(A| C,E) as follows: Given two blocks of text $E$ and $C$, we first strip off non-punctuation therein and

determine the maximum number of punctuation marks *n* in either *E* or *C*.

We employ punctuation-based sentence alignment, which maximizes the probability of overall possible alignment, given a pair of parallel texts, i.e.,

$$\arg\max_{\mathrm{A}} P(\mathrm{A}|C,E),$$

where A is an alignment and C and E are the source and target texts, respectively.

A further approximation encapsulates the dependence of a single parameter b, which is a function of CP and EP:

$$\mathrm{P(A|Ci,Ej) = P(A|b(CP,EP))}.$$

Since it is easier to estimate the distribution for the inverted form, we apply Bayes' Rule to further simplify the calculation:

$$\mathrm{P(A|b) = P(b|A)P(A)/P(b)},$$

where P(b) is a normalizing constant that can be ignored during minimization. P (A) is the match type, and its values are shown in Table 6. We use a binominal distribution to estimate P(b):

$$P(A \mid C,E) \approx \prod_{\mathrm{A}} P(A \mid C_i, E_j)$$

$$\approx \prod_{k=1}^{t} P(A_k) \cdot \binom{n_k}{r_k} P(Cp_k, Ep_k)^{r_k} (1 - P(Cp_k, Ep_k))^{n_k - r_k} \quad , \tag{3}$$

where $n_k$ = the maximum number of punctuation marks in either the English text or the Chinese text in the k[th] sentence to be aligned;

$r_k$ = the number of compatible punctuation marks in ordered comparison;

$P(Cp_k, Ep_k)$ = the probability of the existence of a compatible punctuation mark in both languages;

$P(A_k)$ = the match type probability of aligning $E_{i,k}$ and $C_{j,k}$;

$t$ = the total number of sentences that are aligned.

From the data, we have found that about 66% of the time, a sentence in one language matches exactly one sentence in the other language (1-1). Three additional possibilities should be also considered: 1-0 (including 0-1), and many-1 (including 1-many). Chinese-English parallel corpora are quite noisy, reflecting from wider possibilities of the match types. Here, we used the same probabilistic figures as proposed by Chuang and Chang [2002]. Table 6 shows all eight possibilities used in our implementation.

***Table 6. The sentence alignment match type probability P(A)***

| P(A) | 1-1 | 1-0, 0-1 | 1-2 | 2-1 | 2-2 | 1-3 | 3-1 |
|---|---|---|---|---|---|---|---|
| Chinese-English | 0.64 | 0.0056 | 0.017 | 0.25 | - | 0.056 | - |

### 3.5 A Hybrid Punctuation-based and Length-Based Sentence Alignment Model

The length-based sentence alignment criterion involves a length-related probability distribution function $P(\delta \mid match)$, where $\delta$ is a function of the sentence length of the source language $l_c$ and the sentence length of the target language $l_e$, or $\delta = \delta(l_c, l_e)$. Since the sentence lengths of the bilingual parallel texts of interest are highly correlated, $P(\delta \mid match)$ can be estimated using the Gaussian assumption following Gale and Church [1993]. Incorporating both the length-based and punctuation-based criteria, we can modify equation (3) as follows:

$$P(A \mid C, E) \approx \prod_{k=1}^{t} P(A) \cdot P(\delta \mid match) \cdot \binom{n_k}{r_k} P(CP_k, EP_k)(1 - P(CP_k, EP_k))^{n_k - r_k} \quad . \tag{4}$$

The same dynamic programming optimization can then be used. Again, the computation and memory costs are very low when both the length-based and punctuation-based criteria are employed. The average slopes of $l_c$ and $l_e$, and the associated standard deviations are estimated in an adaptive manner for each corpus being evaluated [Chuang *et al*. 2002].

## 4. Experiments and evaluation

To explore the relationship between the punctuation marks in pairs of Chinese and English sentences that are mutual translations, we prepared a small set of 200 pairs of sentences aligned at the sentence and punctuation levels. We then investigated the characteristics of and the statistics associated with the punctuation marks. We derived estimates of the punctuation translation probabilities and fertility probabilities from the small set of hand-tagged data. This seed information was then used to train the punctuation translation model on a larger corpus via the EM algorithm. The probability of a punctuation mark having a translation counterpart was estimated as $p$ = 0.670 with a standard deviation 0.170. For random pairs of bilingual sentences, $p$ = 0.340, with a standard deviation 0.167. There appears to be marked differences between the two distributions, indicating that, indeed, soft and ordered comparison of punctuation marks across languages provide useful information for effective sentence alignment.

In order to assess the performance of punctuation-based sentence alignment, we randomly selected five bilingual articles from the Sinorama Magazine Corpus and Scientific American (US and Taiwan editions), and several chapters from the novel Harry Potter. These were subjected to an implementation of the proposed method. Some experimental results are shown in appendices A and B.

It should be noted that in Appendix A, the first English sentence and the first Chinese sentence are both title sentences, and that they are aligned based on the carriage return

deliminater, even though no punctuation marks are found in the English sentence. We found that, in general, there were more periods in the English text than that in the Chinese text for a given bilingual corpus, especially in the case of a text translated from Chinese into English. As an example, 112 periods were found in a Chinese article [Sinorama 1988], whereas only 180 periods were found in the corresponding English translation. This phenomenon can be further seen by examining the punctuation distribution. The relationship between the number of sentences and the number of punctuation marks per sentence for the English-Chinese corpus was determined by analyzing 6,103 articles, including around 130,000 sentences, from Sinorama Magazine between 1976 and 2000. 1,138,447 punctuation marks were found in the English corpus and 2,056,675 punctuation marks in the Chinese corpus. Apparently, punctuation marks are used more sparingly in Chinese sentences. As shown in Figure 7, there were two punctuation marks in most of the English sentences and three in most of the Chinese sentences. The figure also shows that a few long sentences had close to one hundred punctuation marks, but these were unuseful.

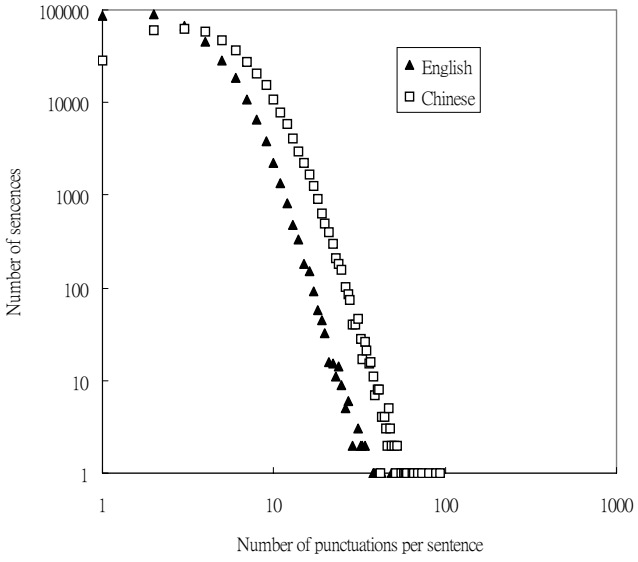



***Figure 7. The punctuation distribution for a bilingual corpus***

This observation prompted us to establish a special rule that the combination of a comma and an open quote in a Chinese sentence should be considered as being equivalent to a full stop. Applying this rule, we found that the sentence count increased from 112 to 126 for the Chinese text mentioned in the above example. This empirical rule helped to improve the

precision of sentence alignment. These special cases can be found in both Appendixes A and B.

The precision rate of the length-based approach [Gale and Church 1993] is shown in Table 7 as a baseline for comparison. The precision rate is defined as the ratio between the number of correctly matched sentences in the system output and the number of matched sentences generated from the system output. The large variation observed in the alignment precision is primary due to the disparity in the lengths and match types. The experimental results obtained with the punctuation-based approach and the combination approaches are shown in Tables 8 and 9. Overall, the punctuation-based approach outperformed the length-based approach, reducing the error rates consistently, and the improvement could exceed 50% at times.

***Table7. Baseline sentence alignment performance achieved using the length-based approach***

| Articles | No. of Chinese Sentences | No. of Errors | Percentage (%) |
|---|---|---|---|
| World in a box* | 75 | 7 | 90.7 |
| What clones* | 77 | 12 | 84.4 |
| New University** | 319 | 24 | 92.5 |
| Book I-2 *** | 439 | 16 | 96.4 |
| Book II-8 *** | 633 | 19 | 97.0 |

* Scientific American

** Sinorama

*** Harry Potter

***Table 8. Performance evaluation using punctuations***

| Article | Baseline | Precision | Improvement |
|---|---|---|---|
| World in a box* | 91.5 | 98.8 | 7.3 |
| What clones* | 86.5 | 96.6 | 10.1 |
| New University** | 93.0 | 95.3 | 2.3 |
| Book I-2 *** | 96.5 | 98.9 | 2.4 |
| Book II-8 *** | 97.1 | 98.0 | 0.9 |

***Table 9. Performance evaluation by combining length and punctuation information***

| Article | Baseline | Precision | Improvement |
|---|---|---|---|
| World in a box* | 91.5 | 100.0 | 8.5 |
| What clones* | 86.5 | 97.8 | 11.2 |
| New University** | 93.0 | 93.9 | 0.9 |
| Book I-2 *** | 96.5 | 96.7 | 0.2 |
| Book II-8 *** | 97.1 | 98.2 | 1.1 |

Additionally, we evaluated our method on a larger corpus the Scientific American Corpus. We used all of the English and Chinese articles from January 2003 to December 2003. There were 67 articles, 1523 English sentences, and 1599 Chinese sentences. Every article included both an English text and its corresponding Chinese text. The punctuation-based sentence alignment method achieved alignment precision rates of over 93%. Inferior performance was achieved when the hybrid punctuation and length-based method was used as compared with the punctuation-based method alone, as shown by the results listed in Tables 8 and 9. This phenomenon may be attributed to the strong dependence of the length-based method on the average length of the sentences being analyzed. Apparently, length-based methods do not perform well in the case of a corpus that is composed of shorter sentences. Therefore, a length-based method may achieve poorer performance when it is combined with a punctuation-based method. Consequently, caution should be exercised in interpreting these precision rates.

Our approach has been proven to be effective, and it has been used to construct a concordancer system called **TotallRecall** [Wu *et al.* 2003] in a Computer Assisted Language Learning (CALL) project. Based on the results of our experiments, it was also possible to speed up the corpus annotation and distribution efforts made by the Association for Computational Linguistics and Chinese Language Processing.

## 5. Discussion

We achieved a striking improvement over the length-based baseline for bilingual text alignment when punctuation was used alone or in combination with lexical information. Combining punctuation and length information, we could get slightly better overall performance. However, the improvement was not entirely consistent. Thus we need to experiment on a longer parallel text in order to be more certain about it.

Although word alignment links cross each other quite often, punctuation links do not. It appears that we can obtain sub-sentential alignment at the clause and phrase levels from the alignment of punctuation. For instance, after we align the punctuation in examples (3c) and

(3e), we can extract the following finer-grained bilingual analyses:

(5c) 逐漸的
(5e) Over time
(6c) 打鼓不再能滿足他
(6e) drums could no longer satisfy him
(7c) 打鼓原是我最喜歡的
(7e) Drumming was at first the thing I loved most
(8c) 後來卻變成邊打邊睡
(8e) but later it became half drumming, half sleeping
(9c) 一個月六萬元的死工作
(9e) just a job for NT$60,000 a month
(10c) 薛岳表示
(10e) says Simon

We have hand-coded a small English-Japanese punctuation mapping table and converted our alignment program to handle alignment of Japanese and English texts. It appears that the adapted program works with performance comparable to that of the original one. An example of aligning English-Japanese parallel texts based on punctuation is shown in Appendix C. Our Japanese-English program is a very preliminary one. Further and more rigorous investigation is needed.

## 6. Conclusion and Future Work

We have developed a very effective sentence alignment method based on punctuation. The probability of the finding matches between different punctuation marks in source and target texts is calculated based on a large bilingual corpus. The punctuation-based measure of mutual translation can be modeled by the binomial distribution. We have implemented the proposed method on the parallel Chinese-English Sinorama Magazine Corpus. The experimental results show that the punctuation-based approach outperforms the length-based approach with precision rates exceeding 93%.

We have also demonstrated that the alignment method can be applied to other bilingual texts, without the need for *a priori* linguistic knowledge of the languages, like Japanese and English. This general approach has been found to be fast, easy to set up, and universal. We believe that this method can be easily applied to many different languages.

A number of interesting future directions for researches present themselves. First, punctuation alignment can be exploited to constrain word alignment and reduce error rates. Second, punctuation alignment makes possible a finer-grained level of bilingual analysis of sub-sentential alignment and can provide a strikingly more effective translation memory and bilingual concordance for more effective example-based machine translation (EBMT), computer assisted translation and language learning (CAT and CALL).

## Acknowledgement

We are indebted to Dr. Jason Chang for helpful discussions and suggestions. We acknowledge the support for this study provided through grants by the National Science Council and Ministry of Education, Taiwan (NSC-2213-E-238-015, NSC 90-2411-H-007-033-MC and MOE EX-91-E-FA06-4-4), and by the MOEA under the Software Technology for Advanced Network Application Project of the Institute for Information Industry.

## Appendix A

Some experimental results of sentence alignment based on length and punctuation are presented here. Shaded parts indicate imprecision in alignment results. We calculated the precision rates by dividing the number of un-shaded sentences (counting both English and Chinese sentences) by the total number of sentences proposed. Since we did not exclude aligned pairs using a threshold, the recall rate should be the same as the precision rate. The experimental results indicate that when non 1-1 matches next to each other tend to fail the length-based aligner. However, the punctuation-based aligner appears to handle such cases more successfully.

| **Sentence alignment based on length** | | |
|---|---|---|
| Type | English text | Chinese Text |
| 11 | Take note | 「共筆」怪現象 |
| 12 | Allowing education to be led by the market may also lead to deficiencies in teaching practices. | 市場領導教育還可能引發教學上的弊病。台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路， |
| 11 | Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams. | 「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」 |
| 31 | "In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams." Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the class expenses fund. | 甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。 |
| 21 | Two years ago, the CER completed a "General Consultation Report on Educational Reform." One of its main proposals was that the past system of controlling the establishment, expansion and contraction of departments in higher education on the basis of estimates of personnel demand should be "relaxed." | 兩年前行政院教改會完成「教育改革總諮議報告書」，建議的重點之一是，過去以人力需求的推估，管制高等教育科系的設立增減，應該「鬆綁」。 |
| 11 | Education cannot be made merely to narrowly serve the economy. | 教育不能「窄化」成只爲經濟服務，但現實的狀況是， |
| 11 | Yet in reality, "the reason most parents are willing | 「大部分家長之所以肯花錢讓孩子來 |

| | | |
|---|---|---|
| | to pay to put their children through university is certainly not that they hope they will become passionate seekers after truth, but to enable them to find good careers," says Providence University president Li Chia-tung bluntly. | 念大學，絕不是希望孩子以後熱衷於真理的追求，而是爲了使孩子將來能找到好職業，」靜宜大學校長李家同明白地說。 |
| **Sentence alignment based on punctuation** | | |
| 11 | Take note | 「共筆」怪現象 |
| 11 | Allowing education to be led by the market may also lead to deficiencies in teaching practices. | 市場領導教育還可能引發教學上的弊病。 |
| 11 | Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams. | 台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路， |
| 11 | "In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams." | 「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」 |
| 21 | Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the class expenses fund. | 甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。 |
| 21 | Two years ago, the CER completed a "General Consultation Report on Educational Reform." One of its main proposals was that the past system of controlling the establishment, expansion and contraction of departments in higher education on the basis of estimates of personnel demand should be "relaxed." | 兩年前行政院教改會完成「教育改革總諮議報告書」，建議的重點之一是，過去以人力需求的推估，管制高等教育科系的設立增減，應該「鬆綁」。 |
| 11 | Education cannot be made merely to narrowly serve the economy. | 教育不能「窄化」成只爲經濟服務，但現實的狀況是， |
| 11 | Yet in reality, "the reason most parents are willing to pay to put their children through university is certainly not that they hope they will become passionate seekers after truth, but to enable them to find good careers," says Providence University president Li Chia-tung bluntly. | 「大部分家長之所以肯花錢讓孩子來念大學，絕不是希望孩子以後熱衷於真理的追求，而是爲了使孩子將來能找到好職業，」靜宜大學校長李家同明白地說。 |

## Appendix B

More English-Chinese alignment results.

| | | |
|---|---|---|
| **Sentence alignment based on length** | | |
| 31 | "The advocacy of core curriculum teaching is in itself a very important education for teachers." Lin Ku-fang says that when NHMC was set up it made broad-based education one of its founding principles, but discovered that attitudes were very hard to change, because "people today feel they are respected for their profession rather than their personality." Although when first studying an academic discipline one starts from a general outline, nonetheless one must be very well versed in a subject to teach it well. | 「通識本身的提倡，對老師就是很重要的教育，」文化評論者林谷芳指出，南華成立時就把通識教育視爲創校理念，但還是發現觀念問題最難突破，因爲「現代人常覺得自己被尊重是因爲我的專業，而不是我的人。」 |
| 12 | "There is a great sense of challenge about core curriculum teaching, but many people make the mistake of thinking it is very simple," says Lin. | 雖然一門學問最初讀時是某某學導論，但真的得弄通，才教得好，「通識挑戰意味很濃，但大家都誤以爲很簡單。」 |
| 11 | The scope of core curriculum teaching appears very broad, but it still has to start from the basics. | 通識範圍看起來很廣，但還是由基礎出發，林谷芳認爲，任何學科都可以從「人與自然」、「人與人」、「人與超自然或自我」三個層次來看。 |
| 11 | In Lin Ku-fang's view, any branch of academic learning can be viewed on the three levels of "man and nature," "man and man," and "man and the supernatural, or that which transcends self." | 就以地球科學這門專業學科爲例，人所認識的自然是專業，提升到人與生態的互動、人與未來生命處境就是通識。 |
| 10 | To take the example of earth science, a very specialized discipline, man's cognitive knowledge of nature is its specialist content, but to go a step higher and investigate the interactive relationship between man and ecology or man and the future condition of life requires a broad-based, multidisciplinary approach. | |
| **Sentence alignment based on punctuation.** | | |
| 21 | "The advocacy of core curriculum teaching is in itself a very important education for teachers." Lin Ku-fang says that when NHMC was set up it made broad-based education one of its founding principles, but discovered that attitudes were very hard to change, because "people today feel they are respected for their profession rather than their personality." | 「通識本身的提倡，對老師就是很重要的教育，」文化評論者林谷芳指出，南華成立時就把通識教育視爲創校理念，但還是發現觀念問題最難突破，因爲「現代人常覺得自己被尊重是因爲我的專業，而不是我的人。」 |
| 11 | Although when first studying an academic | 雖然一門學問最初讀時是某某學導 |

| | | |
|---|---|---|
| | discipline one starts from a general outline, nonetheless one must be very well versed in a subject to teach it well. | 論，但真的得弄通，才教得好， |
| 11 | "There is a great sense of challenge about core curriculum teaching, but many people make the mistake of thinking it is very simple," says Lin. | 「通識挑戰意味很濃，但大家都誤以爲很簡單。」 |
| 21 | The scope of core curriculum teaching appears very broad, but it still has to start from the basics. In Lin Ku-fang's view, any branch of academic learning can be viewed on the three levels of "man and nature," "man and man," and "man and the supernatural, or that which transcends self." | 通識範圍看起來很廣，但還是由基礎出發，林谷芳認爲，任何學科都可以從「人與自然」、「人與人」、「人與超自然或自我」三個層次來看。 |
| 11 | To take the example of earth science, a very specialized discipline, man's cognitive knowledge of nature is its specialist content, but to go a step higher and investigate the interactive relationship between man and ecology or man and the future condition of life requires a broad-based, multidisciplinary approach. | 就以地球科學這門專業學科爲例，人所認識的自然是專業，提升到人與生態的互動、人與未來生命處境就是通識。 |

**Appendix C**

Sentence alignment of English-Japanese parallel texts based on punctuation.

| **Sentence alignment based on punctuation** | | |
|---|---|---|
| Type | English text | Japanese Text |
| 11 | Liu Tseng-kuei, of Academia Sinica's Institute of History and Philology, once analyzed over 570 female names used during the Han dynasty in hopes it might shed some light on what the people of that time hoped to see in a woman. | 中央研究院歷史語言研究所の副研究員である劉増貴さんは、漢代において女性に何が期待されていたかを理解するために、570 名余りの漢代の女性の名前を研究したことがある。 |
| | , , . | 、、。 |
| 12 | It turns out that about two-thirds of the names examined were suitable for either women or men. | その結果、3分の2の名前が男でも女でも通用するものであることがわかった。漢代の女性の名前には実に力強いものも少なくない。 |
| | . | 、。。 |
| 21 | Wang Mang, who usurped the throne in 9 AD, named his daughter Jie ("nimble and quick"). The daughter of the emperor Huan Di (132-167 AD) was named Jian ("solid and resolute") while her mother, the empress Deng, had the even more emphatic name of Mengnu, which means "fierce woman"! | 王莽の娘の名は「倢」、後漢の桓帝の娘の名は「堅」といい、桓帝の時の皇后の名は、より直接的な「猛女」というものだったのである。 |
| | , , ( " " ) . ( ) ( " " ) , , , " " ! | 「」、「」、、「」。 |
| 11 | Says Liu, "These names show that society at that time had not yet come to hold the two sexes to such very different standards." | 「この現象は、男性と女性の道徳行爲に対する社会の要求が、あまり違わなかったことを示しています」と劉増貴さんは言う。 |
| | , " . " | 「、、、」。 |
| 11 | Although they were gradually beginning to use specifically feminine names alluding to a gentle and submissive nature, such traits as a resolute spirit and an agile, tough body were also seen as virtues in a woman. | 当時、いわゆる女性的な名前もしだいに増えており、女性を低く見るという観念も確かにあったが、それでも女性が強くたくましくあることも肯定されていたのである。 |
| | , , . | 、、、。 |
| 11 | "The notion of the ideal woman being soft and weak was not so universally accepted then as it would later come to be." | 「女性は弱くておとなしい方が良いとする考えは、後の時代のように絶対的なものではなかったようです」と劉増貴さんは言う。 |
| | " . " | 「、」。 |

## References

Brown, P. F., J. C. Lai and R. L. Mercer, "Aligning sentences in parallel corpora," *Proceedings of the 29th conference on Association for Computational Linguistics*, Berkeley, CA, USA. 1991, pp. 169-176.

Chen, A. and and F. Gey , "Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval," *TREC 2001.*

Chen, K.H. and H.H. Chen, "A Part-of-Speech-Based Alignment Algorithm," *Proceedings of 15th International Conference on Computational Linguistics,* Kyoto, 1994, pp. 166-171.

Chen, S. F., "Aligning Sentences in Bilingual Corpora Using Lexical Information," *Proceedings of ACL-93,* Columbus OH, 1993, pp. 9-16.

Chuang, T., G.N. You and J.S. Chang , "Adaptive Bilingual Sentence Alignment," L*ecture Notes in Artificial Intelligence 2499*, pp. 21-30.

Collier, N., K. Ono and H. Hirakawa, "An Experiment in Hybrid Dictionary and Statistical Sentence Alignment," *COLING-ACL* 1998, pp. 268-274.

Danielsson, P. and K. Mühlenbock, Small but Efficient,"The Misconception of High-Frequency Words in Scandinavian Translation," *AMTA 2000*, pp. 158-168.

Déjean, H., É Gaussier and F. Sadat, "Bilingual Terminology Extraction: An Approach based on a Multilingual thesaurus Applicable to Comparable Corpora," *Proceedings of the 19th International Conference on Computational Linguistics COLING 2002,* Taipei, Taiwan, pp. 218-224.

Dolan, W. B., J. Pinkham and S. D. Richardson, MSR-MT, "The Microsoft Research Machine Translation System," *AMTA 2002*, pp. 237-239.

Gale, W. A. and K. W. Church , "A program for aligning sentences in bilingual corpora," *Computational Linguistics,* vol. 19, pp. 75-102.

Gey, F.C., A. Chen, M.K. Buckland and R. R. Larson , "Translingual vocabulary mappings for multilingual information access," *SIGIR 2002*, pp. 455-456.

Jutras, J-M., An Automatic Reviser, "The TransCheck System," *Proc. of Applied Natural Language Processing*, pp. 127-134.

Kay, M. and M. Röscheisen, "Text-Translation Alignment," *Computational Linguistics,* 19:1, pp. 121-142.

Ker, S.J. and J.S. Chang, "A class-based approach to word alignment," *Computational Linguistics,* 23:2, pp. 313-344.

Kueng, T.L. and K.Y. Su, "A Robust Cross-Domain Bilingual Sentence Alignment Model," *Proceedings of the 19th International Conference on Computational Linguistics,* 2002.

Melamed, I. D., "A portable algorithm for mapping bitext correspondence," *In The 35th Conference of the Association for Computational Linguistics (ACL 1997),* Madrid, Spain,1997.

Melamed, I. D., "Bitext Maps and Alignment via Pattern Recognition," *Computational Linguistics,* 25(1), pp.107-130, March, 1999.

Moore, R.C., "Fast and Accurate Sentence Alignment of Bilingual Corpora," *AMTA* 2002, pp. 135-144.

Ribeiro, A., G. Dias, G. Lopes and J. Mexia," Cognates Alignment," In Bente Maegaard (ed.), *Proceedings of the Machine Translation Summit VIII (MT Summit VIII) – Machine Translation in the Information Age*, Santiago de Compostela, Spain, 2001, pp. 287–292.

Richards, J. et al., "Longman Dictionary of Applied Linguistics," Longman, 1985.

Simard, M., G. Foster and P. Isabelle, "Using cognates to align sentences in bilingual corpora," *Proceedings of TMI92,* Montreal, Canada, pp. 67-81.

Sinorama , The New University: Breaking down the Departmental Barriers, June, the 3rd article, 1998.

Wu, "Bilingual Collocation Extraction Based on Linguistic and Statistical Analyses," Master thesis, National Tsing Hua University, Taiwan, 2003.

Wu, D., "Aligning a parallel English-Chinese corpus statistically with lexical criteria," *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* New Mexico, USA, 1994, pp. 80-87.

Wu, J.C., K.C. Yeh, T.C. Chuang, W.C. Shei and J.S. Chang, "TotalRecall: A Bilingual Concordance for Computer Assisted Translation and Language Learning," *ACL2003 workshop.*

Yang, Y., "Researches on Punctuation Marks," Tien-Chien Publishing, Hong Kong.

Yang, C. and K. Li, "Automatic Construction of English/Chinese Parallel Corpora," *Journal of American Society of Information Science and Technology,* 54(8), 2003, pp 730-742.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp.123-144 


# Similarity Based Chinese Synonym Collocation Extraction

**Wanyin Li*, Qin Lu* and Ruifeng Xu***

**Abstract**

Collocation extraction systems based on pure statistical methods suffer from two major problems. The first problem is their relatively low precision and recall rates. The second problem is their difficulty in dealing with sparse collocations. In order to improve performance, both statistical and lexicographic approaches should be considered. This paper presents a new method to extract synonymous collocations using semantic information. The semantic information is obtained by calculating similarities from HowNet. We have successfully extracted synonymous collocations which normally cannot be extracted using lexical statistics. Our evaluation conducted on a 60MB tagged corpus shows that we can extract synonymous collocations that occur with very low frequency and that the improvement in the recall rate is close to 100%. In addition, compared with a collocation extraction system based on the Xtract system for English, our algorithm can improve the precision rate by about 44%.

**Keywords:** Lexical Statistics, Synonymous Collocations, Similarity, Semantic Information

## 1. Introduction

A collocation refers to the conventional use of two or more adjacent or distant words which hold syntactic and semantic relations. For example, the conventional expressions “warm greetings”, “broad daylight”, “思想包袱”, and “托运行李” all are collocations. Collocations bear certain properties that have been used to develop feasible methods to extract them automatically from running text. Since collocations are commonly found, they must be recurrent. Therefore, their appearance in running text should be statistically significant, making it feasible to extract them using the statistical approach.

* Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Tel: +852-27667326; +852-27667247 Fax:+852-27740842
E-mail: {cswyli, csluqin, csrfxu}@comp.polyu.edu.hk

A collocation extraction system normally starts with a so-called headword (sometimes also called a keyword) and proceeds to find co-occurring words called the collocated words. For example, given the headword "基本", such bi-gram collocations as "基本理论", "基本工作", and, "基本原因" can be found using an extraction system where "理论", "工作", and "原因" are called collocated words with respect to the headword "基本." Many collocation extraction algorithms and systems are based on lexical statistics [Church and Hanks 1990; Smadja 1993; Choueka 1993; Lin 1998]. As the lexical statistical approach was developed based on the recurrence property of collocations, only collocations with reasonably good recurrence can be extracted. Collocations with low occurrence frequency cannot be extracted, thus affecting both the recall rate and precision rate. The precision rate achieved using the lexical statistics approach can reach around 60% if both word bi-gram extraction and n-gram extraction are employed [Smadja 1993; Lin 1997; Lu *et al*. 2003]. The low precision rate is mainly due to the low precision rate of word bi-gram extraction as only about a 30% - 40% precision rate can be achieved for word bi-grams. The semantic information is largely ignored by statistics- based collocation extraction systems even though there exist multiple resources for lexical semantic knowledge, such as WordNet [Miller 98] and HowNet [Dong and Dong 99].

In many collocations, the headword and its collocated words hold specific semantic relations, hence allowing collocate substitutability. The substitutability property provides the possibility of extracting collocations by finding synonyms of headwords and collocate words. Based on the above properties of collocations, this paper presents a new method that uses synonymous relationships to extract synonym word bi-gram collocations. The objective is to make use of synonym relations to extract synonym collocations, thus increasing the recall rate.

Lin [Lin 1997] proposed a distributional hypothesis which says that if two words have similar sets of collocations, then they are probably similar. According to one definition [Miller 1992], two expressions are synonymous in a context C if the substitution of one for the other in C does not change the truth-value of a sentence in which the substitution is made. Similarly, in HowNet, Liu Qun [Liu *et al.* 2002] defined word similarity as two words that can substitute for each other in a context and keep the sentence consistent in syntax and semantic structure. This means, naturally, that two similar words are very close to each other and they can be used in place of each other in certain contexts. For example, we may either say "买书"or "订书" since "买" and "订" are semantically close to each other when used in the context of buying books. We can apply this lexical phenomena after a lexical statistics-based extractor is applied to find low frequency synonymous collocations, thus increasing the recall rate.

The rest of this paper is organized as follows. Section 2 describes related existing collocation extraction techniques that are based on both lexical statistics and synonymous collocation. Section 3 describes our approach to collocation extraction. Section 4 describes the

data set and evaluation method. Section 5 evaluates the proposed method. Section 6 presents our conclusions and possible future work.

## 2. Related Works

Methods have been proposed to extract collocations based on lexical statistics. Choueka [Choueka 1993] applied quantitative selection criteria based on a frequency threshold to extract adjacent n-grams (including bi-grams). Church and Hanks [Church and Hanks 1990] employed mutual information to extract both adjacent and distant bi-grams that tend to co-occur within a fixed-size window. However, the method can not be extended to extract n-grams. Smadja [Smadja 1993] proposed a statistical model that measures the spread of the distribution of co-occurring pairs of words with higher strength. This method can successfully extract both adjacent and distant bi-grams, and n-grams. However, it can not extract bi-grams with lower frequency. The precision rate of bi-grams collocation is very low, only around 30%. Generally speaking, it is difficult to measure the recall rate in collocation extraction (there are almost no reports on recall estimation) even though it is understood that low occurrence collocations cannot be extracted. Sun [Sun 1997] performed a preliminary *Quantitative* analysis of the strength, spread and peak of Chinese collocation extraction using different statistical functions. That study suggested that the statistical model is very limited and that syntax structures can perhaps be used to help identify pseudo collocations.

Our research group has further applied the Xtract system to Chinese [Lu *et al.* 2003] by adjusting the parameters so at to optimize the algorithm for Chinese and developed a new weighted algorithm based on mutual information to acquire word bi-grams which are constructed with one higher frequency word and one lower frequency word. This method has achieved an estimated 5% improvement in the recall rate and a 15% improvement in the precision rate compared with the Xtract system.

A method proposed by Lin [Lin 1998] applies a dependency parser for information extraction to collocation extraction, where a collocation is defined as a dependency triple which specifies the type of relationship between a word and the modifiee. This method collects dependency statistics over a parsed collocation corpus to cover the syntactic patterns of bi-gram collocations. Since it is statistically based, therefore it still is unable to extract bi-gram collocations with lower frequency.

Based on the availability of collocation dictionaries and semantic relations of words combinatorial possibilities, such as those in WordNet and HowNet, some researches have made a wide range of lexical resources, especially synonym information. Pearce [Pearce 2001] presented a collocation extraction technique that relies on a mapping from one word to its synonyms for each of its senses. The underlying intuition is that if the difference between the occurrence counts of a synonym pair with respect to a particular word is at least two, then they

can be considered a collocation. To apply this approach, knowledge of word (concept) semantics and relations with other words must be available, such as that provided by WordNet. Dagan [Dagan 1997] applied a similarity-based smoothing method to solve the problem of data sparseness in statistical natural language processing. Experiments conducted in his later research showed that this method could achieve much better results than back-off smoothing methods in terms of word sense disambiguation. Similarly, Hua [Wu 2003] applied synonyms relationships between two different languages to automatically acquire English synonymous collocations. This was the first time that the concept of synonymous collocations was proposed. A side intuition raised here is that a natural language is full of synonymous collocations. As many of them have low occurrence rates, they can not be retrieved by using lexical statistical methods.

HowNet, developed by Dong et al. [Dong and Dong 1999] is the best publicly available resource for Chinese semantics. Since semantic similarities of words are employed, synonyms can be defined by the closeness of their related concepts and this closeness can be calculated. In Section 3, we will present our method for extracting synonyms from HowNet and using synonym relations to further extract collocations. While a Chinese synonym dictionary, Tong Yi Ci Lin (《同义辞林》), is available in electronic form, it lacks structured knowledge, and the synonyms listed in it are too loosely defined and are not applicable to collocation extraction.

## 3. Our Approach

Our method to extract Chinese collocations consists of three steps.

**Step 1**: We first take the output of any lexical statistical algorithm that extracts word bi-gram collocations. This data is then sorted according to each headword, $w_h$, along with its collocated word, $w_c$.

**Step 2:** For each headword, $w_h$, used to extract bi-grams, we acquire its synonyms based on a similarity function using HowNet. Any word in HowNet having a similarity value exceeding a threshold is considered a synonym headword, $w_s$, for additional extractions.

**Step 3**: For each synonym headword, $w_s$, and the collocated word, $w_c$, of $w_h$, if the bi-gram ($w_s$, $w_c$) is not in the output of the lexical statistical algorithm applied in Step 1, then we take this bi-gram ($w_s$, $w_c$) as a collocation if the pair appears in the corpus by applying an additional search on the corpus.

### 3.1 Bi-gram Collocation Extraction

In order to extract Chinese collocations from a corpus and to obtain result in Step 1 of our algorithm, we use an automatic collocation extraction system named CXtract, developed by a research group at Hong Kong Polytechnic University [Lu *et al.* 2003]. This collocation

extraction system is based on English Xtract [Smaja 1993] with two improvements. First, the parameters ($K_0$, $K_1$, $U_0$) used in Xtract are adjusted so as to optimize them for a Chinese collocation extraction system, resulting in an 8% improvement in the precision rate. Secondly, a solution is provided to the so-called high-low problem in Xtract, where bi-grams with a high frequency the head word, $w_h$, but a relatively low frequency collocated word, $w_i$ can not be extracted. We will explain the algorithm briefly here. According to Xtract, a word concurrence is denoted by a triplet ($w_h$, $w_i$, $d$), where $w_h$ is a given headword and $w_i$ is a collocated word appeared in the corpus with a distance $d$ within the window [-5, 5]. The frequency, $f_i$, of the collocated word, $w_i$, in the window [-5, 5] is defined as

$$f_i = \sum_{j=-5}^{5} f_{i,j} \tag{1}$$

where $f_{i,j}$ is the frequency of the collocated word $w_i$ at position $j$ in the corpus within the window. The average frequency of $f_i$, denoted by $\overline{f}_i$, is given by

$$\overline{f_i} = \sum_{j=-5}^{5} f_{i,j} / 10 . \tag{2}$$

Then, the average frequency, $\overline{f}$, and the standard deviation, $\sigma$, are defined as

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i \; ; \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2} . \tag{3}$$

The *Strength* of co-occurrence for the pair ($w_h$, $w_i$), denoted by $k_i$, is defined as

$$k_i = \frac{f_i - \overline{f}}{\sigma} . \tag{4}$$

Furthermore, the *Spread* of ($w_h$, $w_i$), denoted by $U_i$, which characterizes the distribution of $w_i$ around $w_h$, is define as

$$U_i = \frac{\sum (f_{i,j} - \overline{f_i})^2}{10} . \tag{5}$$

To eliminate bi-grams which are unlikely to co-occur, the following set of threshold values is defined:

$$C1 : k_i = \frac{f_i - \overline{f}}{\sigma} \geq K_0 \tag{6}$$

$$C2 : U_i \geq U_0 \tag{7}$$

$$C3 : f_{i,j} \geq \overline{f_i} + (K_1 \cdot \sqrt{U_i}) \tag{8}$$

where the threshold value set ($K_0$, $K_1$, $U_0$) is obtained through experiments. A bi-gram ($w_h$, $w_i$, $d$) will be filtered out as a collocation if it does not satisfy one of the above conditional thresholds. Condition **C1** is used to measure the "recurrence" property of collocations when the bi-grams ($w_h$, $w_i$, $d$) with co-occurrences frequencies higher than $K_0$ times the standard deviation over the average are selected. **C2** is used to select bi-gram pairs ($w_h$, $w_i$, $d$) having a spread values that are larger than a given threshold, $U_0$. A lower U value implies that the bi-gram is evenly distributed in all 10 positions and thus is not considered a "rigid combination". **C3** is used to select bi-grams in these "certain positions". Only if certain peak positions exist, the co-occurrence bi-grams are considered collocations. The values of ($K_0$, $K_1$, $U_0$) are set to (1, 1, 10), which are the optimal parameters for English according to Xtract. For the CXtract, the values of ($K_0$, $K_1$, $U_0$) are adjusted to (1.2, 1.2, 12) which are suitable for the Chinese collocation extraction.

However, Xtract cannot extract high-low collocations when $w_h$ has a quite high frequency and its co-word $w_i$ has a relatively low frequency. For example, "棘手问题" is a bi-gram collocation. But because *freq (*棘手*)* is much lower than the *freq (*问题*),* this bi-gram collocation cannot be identified, resulting in a lower recall rate. In CXtract, an additional step is used to identify such high-low collocations by measuring the conditional probability as follows:

$$R_i = \frac{f(\mathrm{w_h}, w_i)}{f(w_i)} \geq R_0, \tag{9}$$

which measures the likelihood of occurrence of $w_h$ given $w_i$, thus discounting the absolute frequency of $w_i$. CXtract outputs a list of triplets ($w_h$, $w_i$, $d$), where ($w_h$, $w_i$,) is considered to be a collocation.

## 3.2 Construct Synonyms Set

In **Step 2** of our system, for each given headword $w_h$, we first need to find its *synonym set* $W_{syn}$, which contains all the words that are said to be the synonyms of $w_h$. As stated earlier, we estimate the synonym relation between words based on semantic similarity calculation in HowNet. Therefore, before explaining how the synonym set can be constructed, we will introduce the semantic structure of HowNet and the similarity model built based on HowNet.

### 3.2.1 Semantic Structure of HowNet

Because we hope to explore the different semantics meanings that each word carries, word sense disambiguation is the main issue when we calculate the similarity of words. For example, the word "打" used with the words "酱油" as in "打酱油" and "网球" as in "打网球" has the meanings of buy( "卖") and exercise("锻炼"), respectively. As a bilingual semantic and syntactic knowledge base, HowNet provides separate entries when the same word contains

more than one concept. Unlike WordNet, in which a semantic relation is a relation between synsets, HowNet adopts a constructive approach to semantic representation. It describes words as a set of concepts (义项) and describes each concept using a set of primitives（义元）,which is the smallest semantic unit in HowNet and cannot be decomposed further. The template of word concepts is organized in HowNet as shown below:

| |
|---|
| NO.= the record number of the lexical entries |
| W_C/E = concept of the language (Chinese or English) |
| E_C/E = example of W_C/E |
| G_C/E = Part-of-speech of the W_C/E |
| DEF = Definition, which is constructed by primitives and pointers |

For example, in the following, for the word "打", we list the two of its corresponding concepts:

NO.=000001
W_C=打
G_C=V
E_C=~酱油，~张票，~饭，去~瓶酒，醋~来了
W_E=buy
G_E=V
E_E=
DEF=buy|买

NO.=017144
W_C=打
G_C=V
E_C=~网球，~牌，~秋千，~太极，球~得很棒
W_E=play
G_E=V
E_E=DEF=exercise|锻练, sport|体育

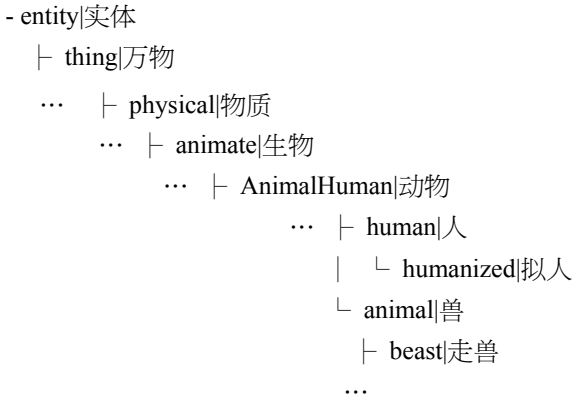
Note: Replace all the graphics above by simple text. In the above records, DEFs are where the primitives are specified. DEF contains up to four types of primitives: *basic independent primitives* (基本独立义元), *other independent primitives* (其他独立义元), *relation primitives* (关系义元), and *symbol primitives* (符号义元), where basic independent primitives and other independent primitives are used to indicate the basic concept, and the

other types are used to indicate syntactical relationships. For example, the word “生日” has all four types of primitives as shown below:

NO.=072280
W_C=生日
G_C=n
E_C=祝贺~，过~，~聚会
W_E=birthday
G_E=n
E_E=
DEF=time|时间, day|日, @ComeToWorld|问世, $congratulate|祝贺

The basic independent primitive “time|时间” defines the general classification of “birthday|生日”. The other independent primitive “day|日” indicates that “birthday|生日” is related to “day|日”. The symbol primitives “@ComeToWorld|问世” and “$congratdulate|祝贺” provide more specific, distinguishing features to indicate syntactical relationships. The pointer “@” specifies “time or space”, indicating that “birthday|生日” is the time of “ComeToWorld|问世”. Another pointer “$” specifies “object of V”, which means that “birthday|生日” is the object of “congratulate|祝贺”. In summary, we find that “birthday|生日” belongs to “time|时间” in general and is related to “day|日” which specifies the time of “ComeToWorld|问世”.

The primitives are then linked by a hierarchical tree to indicate the parent-child relationships as shown in the following example:



Note: Replace all the graphics above by simple text.

This hierarchical structure provides a way to link a concept with any other concept in HowNet, and the closeness of concepts can be represented by the distance between the two concepts.

### 3.2.2 Similarity Model Based on HowNet

Liu Qun [Liu 2002] defined word similarity as two words that can substitute for each other in the same context and still keep the sentence syntactically and semantically consistent. This is very close to our definition of synonyms. Thus, in this work, we will directly use the similarity function provided by Liu Qun, which is stated below.

A word in HowNet is defined as a set of concepts, and each concept is represented by primitives. We describe HowNet as a collection of *n* words, *W*:

$$W = \{w_1, w_2, \ldots w_n\}.$$

Each word $w_i$ is, in turn, described by a set of concepts *S*

$$w_i = \{S_{i1}, S_{i2}, \ldots S_{ix}\},$$

and, each concept $S_i$ is, in turn, described by a set of primitives:

$$S_i = \{p_{i1}, p_{i2}, \ldots p_{iy}\}.$$

For each word pair, $w_1$ and $w_2$, the similarity function is defined by

$$Sim(w_1, w_2) = \max_{i=1\cdots n, j=1\cdots m} Sim(S_{1i}, S_{2j}) \tag{10}$$

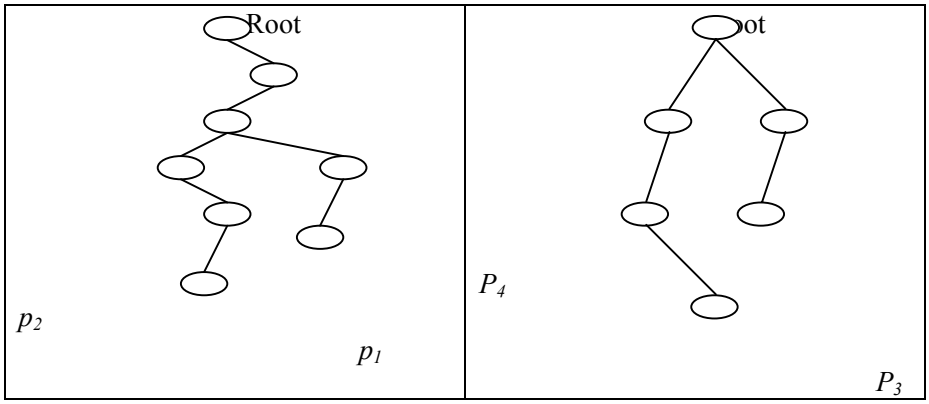
where $S_{1i}$ is the list of concepts associated with $w_1$ and $S_{2j}$ is the list of concepts associated with $w_2$.

As any concept, $S_i$ is represented by its primitives. The similarity of primitives for any $p_1$ and $p_2$ of the same type can be expressed by the following formula:

$$Sim(p_1, p_2) = \frac{\alpha}{Dis(p_1, p_2) + \alpha} \tag{11}$$

where α is an adjustable parameter with a value of 1.6 according to Liu [Liu 2002]. $Dis(p_1, p_2)$ is the path length between $p_1$ and $p_2$ based on the semantic tree structure. The above formula does not explicitly indicate that the depth of a pair of nodes in the tree affects their similarity. For two pairs of nodes ($p_1$, $p_2$) and ($p_3$, $p_4$) with the same distance, the deeper the depth, the more commonly shared ancestors they have, which means that they are semantically

closer to each other. In the following two tree structures, the pair of nodes ($p_1, p_2$) in the left tree should be more similar than ($p_3, p_4$) in the right tree:




To clarify this observation, α is modified as a function of the tree depths of the nodes using the formula $\alpha=min(d(p_1),d(p_2))$. Consequently, the formula (11) was rewritten as formula (11ª) below for our experiments.

$$Sim(p_1, p_2) = \frac{\min(d(p_1), d(p_2))}{Dis(p_1, p_2) + \min(d(p_1), d(p_2))} \tag{11ª}$$

where $d(p_i)$ is the depth of node $p_i$ in the tree. Calculating the word similarity by applying formulas (11) and (11ª) will be discussed in Section 4.4.

Based on the DEF descriptions in HowNet, different primitive types play different roles, and only some are directly related to semantics. To make use of both semantic and syntactic information, the similarity between two concepts should take into consideration all the primitive types with weighted considerations; and thus, the formula is

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(p_{1j}, p_{2j}) \tag{12}$$

where $\beta_i$ is a weighting factor given in [Liu 2002], where the sum of $\beta_1 + \beta_2 + \beta_3 + \beta_4$ is 1 and $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$. The distribution of the weighting factors is given for each concept a priori in HowNet to indicate the importance of primitive $p_i$ for the corresponding concept $S$. The similarity model given here is the basis for building a synonyms set where $\beta_1$ *and* $\beta_2$ represent the semantic information, and $\beta_3$ *and* $\beta_4$ represent the syntactic relation.

### 3.2.3 The Set of Synonyms Headwords

For each given headword $w_h$, we apply the similarity formula in Equation (10) to generate its synonym set, $W_{syn,}$ which is defined as

$$W_{syn} = \{w_s : Sim(w_h, w_s) > \theta\} \quad (13)$$

where $0 < \theta < 1$ is an algorithm parameter which is adjusted based on experience. We set it to 0.85 based on our experiment because we wanted to balance the strength of the synonym relationship and the coverage of the synonym set. Setting the parameter $\theta < 0.85$ will weaken the similarity strength of the extracted synonyms. For example, a given collocation "改善关系" is unlikely to include the candidates "改善护照" and ，"改善契据". On the other hand, setting the parameter $\theta > 0.85$ will limit the coverage of the synonym set, thus valuable synonyms will be lost. For example, for a given bi-gram "重大贡献", we hope to include candidate synonymous collocations such as "重大成果", "重大成绩", and "重大成就". We will discuss the test on $\theta$ in section 5.2.

## 3.3 Synonyms Collocation

H. Wu [Wu 2003] defined a synonymous collocation pair as two collocations that are similar in meaning, but not identical in wording. Actually, in natural language, there exist many synonym collocations. For example, "switch on light" and "turn on light", "财务问题" and "财政问题". However, the sparse appearance of word combinations in a training corpus due to the limitation on the corpus size itself, some synonym collocations may not be extracted by the statistical method because of their lower co-occurrence frequencies. Based on this observation, we perform a further step. Our basic idea is to use a bi-gram collocation ($w_h$, $w_c$, $d$) to further obtain the synonym set $W_{syn}$ of $w_h$, quantified by the similarity function. Then, for each $w_s$ in $W_{syn}$, we consider ($w_s$, $w_c$, $d$) as a collocation if it indeed appears in the corpus at least a given number of times.

Our definition of a synonym collocation as follows. For a given collocation $(w_s, w_c, d)$, if $w_s \in W_{syn}$, then we deem the triple $(w_s, w_c, d)$ to be a synonyms collocation with respect to the collocation $(w_h, w_c, d)$ if $(w_s, w_c, d)$ appears in the corpus $N$ times, where $N$ is a threshold value which we set to 2 in our experiment. Therefore, we define the collection of synonym collocations $C_{syn}$ as

$$C_{syn} = \{(w_s, w_c, d) : Freq(w_s, w_c, d) >= N\} \quad (14)$$

where $w_s \in W_{syn}$.

Our experimental results show that the precision rate of synonym collocation extraction is around 80% when we use the knowledge of HowNet. Some pseudo collocations can be automatically excluded because of the fact that they do not appear in the corpus. For example, for the headword "增长" in the collocation "增长见识", the synonym set extracted from our system contains {"增加", "增高", "增多"}, so the pseudo-collocations "增高见识", "增加见识", and "增多见识" will be excluded because they are not being used customarily used and,

thus, do not appear in the corpus. We checked them using Google and found that they did not appear either. On the other hand, for the collocated word “见识”, our system extracts the synonyms set {“眼光”, “眼界”}, and the word combination “增长眼界” appears twice in our corpus, thus according to our definition, it is a collocation. Therefore, the collocations “增长见识” and “增长眼界” are synonym collocations, and we can successfully extract “增长眼界” even though its frequency is very low (below 10 in our system).

## 4. Data Set and Evaluation Method

We modified Liu Qun’s similarity model based on HowNet to obtain the synonyms of specified words. HowNet is a Chinese-English Bilingual Knowledge Dictionary. It includes both word entries and concept entries. There are more than 60 thousand Chinese concept entries and around 70 thousand English concept entries in HowNet. Both Chinese and English word entries are more than 50 thousand.

The corpus we used contains over 60MB of tagged sentences. Our experiment was conducted using tagged corpus of 11 million words collected six months from the People’s Daily. For word bi-gram extraction, we considered only content words, thus, headwords were nouns, verbs or adjectives only.

In order to illustrate the effect of our algorithm, we used the statistically based system discussed in Section 3.1 as our baseline systems where the output data is called Set A. Using the output of the baseline system, we could further apply our algorithm to produce a data set called Set B.

The collocation performance is normally evaluated based on precision and recall as defined below:

$$precision = \frac{number\ of\ correct\ Extracted\ Collocations}{total\ number\ of\ extracted\ Collocations}, \tag{15}$$

$$recall = \frac{number\ of\ correct\ Extracted\ Collocations}{total\ number\ of\ actual\ Collocations}. \tag{16}$$

However, in collocation extraction, the absolute recall rate is rarely used because there are no benchmark “standard answers”. Alternatively, we can use recall improvement to evaluate our system as defined below.

$$recall = \frac{(N_{none\_syn} + N_{syn})/X - N_{none\_syn}/X}{N_{syn}/X}, \tag{17}$$

where $N_{none\text{-}syn}$ stands for the number of non-synonyms collocations extracted by a statistical model, $N_{syn}$ stands for the number of synonym collocations extracted based on synonym

relationships, and *X* stands for the total number of collocations in the corpus with respect to the given headwords.

Because there are no readily available "standard answers" for collocations, our results were checked manually to verify whether each candidate bi-gram was a true collocation or not. Since the output from the baseline system obtained using 60MB of tagged data consisted of over 200,000 collocations, we had to use the random sampling method to conduct an evaluation. In order to perform a fair evaluation, we tried to avoid subjective selection of words. Therefore, we randomly selected 5 words for each of the three types of words, namely, 5 nouns, 5 verbs, and 5 adjectives. Because headwords we chose were completely random and we did not target any particular words, our results should be statistically sound. Following is a list of the 15 randomly selected words used for the purpose evaluation:

nouns: 基础, 思想, 研究, 条件, 评选;
verbs: 改善, 加大, 增长, 提起, 颁发;
adjectives: 明显, 全面, 重要, 优秀, 大好

Table 1 shows samples of word bi-grams extracted using our algorithm that are considered collocations of the headwords "重大", "改善" and "加大". Table 2 shows bi-grams extracted by our algorithm that are not considered true collocations.

***Table 1. Sample table for true collocations of the headwords "重大","改善","加大"***

| F_5 | F_4 | F_3 | F_2 | F_1 | Headword | F1 | F2 | F3 | F4 |
|---|---|---|---|---|---|---|---|---|---|
| * | * | * | * | * | 重大 | 意义 | * | * | * |
| * | * | * | * | * | 重大 | 影响 | * | * | * |
| * | * | * | * | * | 重大 | 作用 | * | * | * |
| * | * | * | * | * | 改善 | 关系 | * | * | * |
| * | * | * | * | * | 改善 | * | 环境 | * | * |
| * | * | * | * | * | 改善 | * | 交通 | * | * |
| * | * | * | * | * | 改善 | * | 结构 | * | * |
| * | * | * | * | 进一步 | 改善 | * | * | * | * |
| * | * | * | * | 明显 | 改善 | * | * | * | * |
| * | * | * | * | * | 改善 | * | 条件 | * | * |
| * | * | * | * | * | 改善 | * | 状况 | * | * |
| * | * | * | * | 进一步 | 加大 | * | * | * | * |
| * | * | * | * | * | 加大 | * | 力度 | * | * |
| * | * | * | * | * | 提起 | 公诉 | * | * | * |
| * | * | * | * | * | 提起 | 诉讼 | * | * | * |
| * | * | * | * | * | 增加 | * | 负担 | * | * |

***Table 2. Sample table of bi-grams that are not true collocations***

| F_4 | F_3 | F_2 | F_1 | Headword | F1 | F2 | F3 | F4 | F5 |
|---|---|---|---|---|---|---|---|---|---|
| * | * | * | * | 重大 | 政治 | * | * | * | * |
| * | 中 | * | * | 重大 | * | * | * | * | * |
| * | * | * | 着 | 重大 | * | * | * | * | * |
| * | * | * | 作出 | 重大 | * | * | * | * | * |
| * | * | * | * | 改善 | * | * | 關係 | * | * |
| * | * | 要 | * | 改善 | * | * | * | * | * |
| * | * | 将 | * | 改善 | * | * | * | * | * |
| * | * | * | * | 改善 | 金融 | * | * | * | * |
| * | * | * | * | 改善 | 农村 | * | * | * | * |
| * | * | * | 将 | 加大 | * | * | * | * | * |
| * | * | * | * | 加大 | 科技 | * | * | * | * |
| * | * | * | * | 加大 | 农业 | * | * | * | * |
| * | * | * | * | 加大 | * | 企業 | * | * | * |
| * | * | * | * | 加大 | 投入 | * | * | * | * |
| * | * | * | 要 | 加大 | * | * | * | * | * |
| * | * | * | | 加大 | * | * | 企業 | | * |

## 5. Evaluation and Analysis

### 5.1 Improvement in precision and recall rates

In Step 1 of the algorithm, 15 headwords were used to extract bi-gram collocations from the corpus, and 703 pairs of collocations were extracted. Evaluation by hand identified 232 true collocations in the set A test set. The overall precision rate was 31.7% (see Table 3).

***Table 3. Statistics of the test set for set A***

| | n. + v. + a. |
|---|---|
| Headwords | 15 |
| Extracted Bi-grams | 703 |
| True collocations obtains using lexical statistics only | 232 |
| Precision rate | 31.7 % |

In Step 2 of our algorithm, where $\theta = 0.85$ was used, we obtained 94 synonym headwords (including the original 15 headwords). Out of these 94 synonym headwords, 841 bi-gram pairs were then extracted from the baseline system, and 243 were considered true collocations. Then, in Step 3 of our algorithm, we extracted an additional 311 bi-gram pairs; among them, 261 were considered true collocations. Because the synonym collocation extraction algorithm has

achieved a high precision rate of around 84% (261/311 = 83.9%) according to our experimental result as shown in Table 4.

***Table 4. Statistics of the test set for mode B***

| | n. + v. + a. |
|---|---|
| Synonym headwords | 94 |
| Bi-grams (lexical statistics) | 841 |
| Non-synonym collocations (lexical statistics only) | 243 |
| Synonym collocations extracted in Step 3 | 311 |
| True synonym collocations obtained in Step 3 | 261 |
| Overall precision rate | 83.9% |

Since the data for Set B consisted of the additional extracted collocations. When we employed both Set A and Set B together as an overall system, the precision increased to 44 % ((243+261)/(841+311) = 43.7%), an improvement of almost 33% (43.7%-32.9%)/32.9% = 32.8%) comparing with the precision rate of the baseline system as shown in Table 5. As stated earlier, we are not able to evaluate the recall rate. However, compared with the statistical method indicated by Set A, an additional 261 collocations were recalled. Thus, we can record the recall the improvement which is ((243+261) – 243) /243= 107.4% as shown in Table 5.

***Table 5. Comparison of sets A and B***

| Precision Rate of the Statistic Model (Set A) | Precision Rate if the Synonyms Model(Set B) | Overall Precision Rate | Overall Improvement in Recall |
|---|---|---|---|
| 32% | 84% | 44% | 107.4% |

## 5.2 A analysis of the loss / gain in recall

To test the average recall improvement achieved with synonym collocation extraction, we experimented on three set tests with 9, 15, and 21 distinct headwords respectively. The results are shown in Table 6.

***Table 6. Statistics of three test sets***

| Test 1 | | | Test 2 | | | Test 3 | | |
|---|---|---|---|---|---|---|---|---|
| Set A | Headwords | 9 | Set A | Headwords | 15 | Set A | Headwords | 21 |
| | Bi-grams | 253 | | Bi-grams | 703 | | Bi-grams | 153 |
| | Collocations | 77 | | Collocations | 232 | | Collocations | 445 |
| Set B | Synonym Headwords | 55 | Set B | Synonym Headwords | 94 | Set B | Synonym Headwords | 121 |
| | Bi-grams | 614 | | Bi-grams | 841 | | Bi-grams | 203 |
| | Non-synonym Collocations | 179 | | Non-synonym Collocations | 243 | | Non-synonym Collocations | 576 |
| | Extracted Synonym Collocations | 201 | | Extracted Synonym Collocations | 311 | | Extracted Synonym Collocations | 554 |
| | Synonym Collocations | 178 | | Synonym Collocations | 261 | | Synonym Collocations | 476 |
| Recall improvement: 99.49% | | | Recall improvement: 107.4% | | | Recall improvement: 82.6% | | |
| Average improvement in recall: 96.5% | | | | | | | | |

The above table shows that the average recall improvement was close to 100% when using the synonyms relationships were used in the collocation extraction. With different choices of headwords, the improvement averaged about 100% with a standard deviation of 0.106, which indicates that our sampling approach to evaluation is reasonable.

## 5.3 The choice of θ

We also conducted a set of experiments to choose the best value for the similarity function's threshold $\theta$. We tested the best value of $\theta$ based on both the precision rate and the estimated recall rate using so-called remainder bi-grams. The remainder bi-grams are all the bi-grams extracted by the algorithm. When the precision goes up, the size of the result is smaller, indicating a decreasing of recalled collocations. Figure 1 shows the precision rate and estimated recall rate recorded when we tested the value of $\theta$.

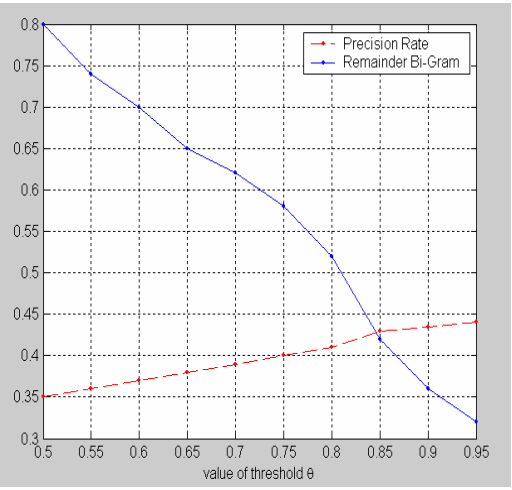



***Figure 1. Precision rate vs. the value of θ***

From Figure 1, it is obvious that at $\theta=0.85$, the recall rate starts to drop more drastically without much improvement in precision.

## 5.4 The test of ($K_0$, $K_1$, $U_0$)

The original threshold for CXtract is *(1.2, 1.2, 12)* for the parameters $(K_0, K_1, U_0)$**.** However, with respect to synonym collocations, we also conducted some experiments to see whether the parameters should be adjusted. Table 7 shows the statistics used to test the value of $(K_0, K_1, U_0)$. The similarity threshold $\theta$ was fixed at 0.85 throughout the experiments.

***Table 7.Values of (K0, K1, U0)***

| | Bi-grams extracted using lexical statistics | Synonym collocations extracted in Step2 |
|---|---|---|
| (1.2,1.4,12) | 465 | 328 |
| (1.4,1.4,12) | 457 | 304 |
| (1.4,1.6,12) | 394 | 288 |
| (1.2,1.2,12) | 513 | 382 |
| (1.2,1.2,14) | 503 | 407 |
| (1.2,1.2,16) | 481 | 413 |

The experimental results show that varying the value of $(K_0, K_1)$ does not benefit our algorithm. However, increasing the value of $U_0$ does improve the extraction of synonymous collocations. Figure 2 shows that $U_0$=14 provides a good trade-off between the precision rate and the remainder Bi-grams. This result is reasonable. According to Smadja, $U_0$ as defined in equation (8) represents the co-occurrence distribution of the candidate collocation $(w_h, w_c)$ at the position $d$ ($-5 \le d \le 5$). For a true collocation ($w_h$, $w_c$, $d$), its co-occurrence frequency at the position $d$ is much higher than those at other positions, which leads to a peak in the co-occurrence distribution. Therefore, it is selected by the statistical algorithm based on equation (10). Based on the physical meaning, one way to improve the precision rate is to increase the value of the threshold $U_0$. A side effect of increasing the value of $U_0$ is a decreased recall rate because some true collocations do not meet the condition of co-occurrence frequency in the ten positions greater than $U_0$. Step 2 of the new algorithm regains some true collocations that are lost because of the higher value of $U_0$ in Step 1.

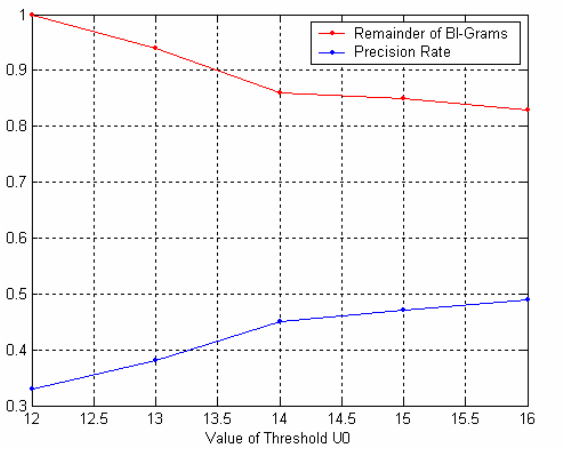



***Figure 2. Precision rate vs. the value of $U_0$***

## 5.5 A comparison of similarity calculation using equations (11) and (11ª)

Table 8 lists the similarity values calculated using equation (11), where α is a constant with a given value of 1.6, and equation (11ª), where α is replaced with a function of the depths of the nodes. Results show that (11ª) is finer tuned, and that it also reflects the nature of the data better. For example, 工人 and 农民 are more similar than 工人 and 运动员. 粉红 and 红 are similar but not the same.

***Table 8. Comparison of calculated similarity results***

| Word 1 | Word 2 | Formula(11) | Formula(11ª) |
|---|---|---|---|
| 男人 | 女人 | 0.86 | 0.95 |
| 男人 | 父亲 | 1.00 | 1.00 |
| 男人 | 和尚 | 0.86 | 0.95 |
| 男人 | 高兴 | 0.05 | 0.10 |
| 工人 | 农民 | 0.72 | 0.88 |
| 工人 | 运动员 | 0.72 | 0.88 |
| 中国 | 美国 | 0.94 | 0.92 |
| 粉红 | 红 | 1.00 | 0.92 |
| 粉红 | 红色 | 1.00 | 0.92 |
| 十分 | 非常 | 1.00 | 1.00 |
| 十分 | 特别 | 0.62 | 0.95 |
| 考虑 | 思想 | 0.70 | 1.00 |
| 思考 | 考虑 | 1.00 | 1.00 |

## 5.6 An Example

***Table 9. Substitution of headwords and collocated words for the collocation "迅速增长"***

| Substitution headword | Substitution collocated word | Freq. in corpus | Freq. in Google results | Substitution collocated word | Freq. in corpus | Freq. in Google results |
|---|---|---|---|---|---|---|
| 迅速增加 | | 15 | 17,000 | 迅捷增长 | 0 | 7 |
| 迅速增多 | | 2 | 14,900 | 迅速增长 | 20 | 224,000 |
| 迅速增高 | | 0 | 744 | 飞快增长 | 0 | 2,530 |
| | 快速增长 | 111 | 1,280,000 | 飞速增长 | 4 | 48,100 |
| | 急遽增长 | 4 | 64,100 | 高速增长 | 60 | 543,000 |
| | 急促增长 | 0 | 201 | 火速增长 | 2 | 211 |
| | 急速增长 | 2 | 19,700 | 全速增长 | 3 | 607 |
| | 急骤增长 | 0 | 1,020 | 神速增长 | 0 | 55 |
| | 迅猛增长 | 4 | 84,600 | 麻利增长 | 0 | 0 |
| | 迅疾增长 | 0 | 98 | 湍急增长 | 0 | 0 |

The above example shows for the collocation "迅速增长", how each word is substituted and the statistical data for the synonym collocations. Our system extracts twenty candidate synonym collocations. Seven of them are synonym collocations with frequencies below than 10. Four of them have frequencies above 10, which means that they can be extracted by using statistical models only. Another nine of them do not appear in our corpus, which including two pseudo collocations "麻利增长"and "湍急增长".

## 6. Conclusions and On-Going Work

In this paper, we have presented a method to extract bi-gram collocations using a lexical statistics model with synonym information. Our method achieved a precision rate of 44% for the tested data. Comparing with the precision of 32% obtained using lexical statistics only, our method results in an improvement of close to 33%. In addition, the recall improvement achieved reached 100% on average. The main contribution of our method is that we make use of synonym information to extract collocations which otherwise cannot be extracted using a lexical statistical method alone. Our method can supplement a lexical statistical method to increase the recall quite significantly.

Our work focuses on synonym collocation extraction. However, Manning [Manning 99] claimed that the lack of valid substitutions for synonyms is a characteristic of collocations in general [Manning and Schutze 1999]. Nevertheless, our method shows that synonym

collocations do exist and that they are not a minimal collection that can be ignored in collocation extraction.

To extend our work, we will further apply synonym information to identify collocations of different types. Our preliminary study has suggested that collocations can be classified into 4 types:

**Type 0 collocations**: These are fully fixed collocations which including some idioms, proverbs, and sayings, such as "缘木求鱼"，"釜底抽薪" and so on.

**Type 1 collocations**: These are fixed collocation in which the appearance of one word implies the co-occurrence of another one as in "历史包袱".

**Type 2 collocations**: These are strong collocation which allow very limited substitution of components, as in, "裁减职位", "减少职位", "缩减职位" and so on. These collocations are classified with type 3 collocations when substitution can occur at only one end, not both ends.

**Type 3 collocations**: These are loose collocations which allow more substitutions of components; however a limitation is still required to restrict the substitution as in "减少开支", "缩减开支", "压缩开支", "消减开支".

By using synonym information and defining substitutability, we can validate whether collocations are fixed collocations, strong collocations with very limited substitutions, or general collocations that can be substituted more freely. Based on this observation, we are currently working on a synonym substitution model for classifying the collocations into different types automatically.

### Acknowledgements and notes

Our great thanks go to Dr. Liu Qun of the Chinese Language Research Center of Peking University for letting us share their data structure in the Synonym Similarity Calculation. This work was partially supported by Hong Kong Polytechnic University (Project Code A-P203) and a CERG Grant (Project code 5087/01E). Ms. Wanyin Li is currently a lecturer in the department of Computer Science of Chu Hai College, Hong Kong.

## References

Benson, M., "Collocations and General Purpose Dictionaries," *International Journal of Lexicography,* 3(1), 1990, pp. 23-35.

Choueka, Y., "Looking for Needles in a Haystack or Locating Interesting Collocation Expressions in Large Textual Database," *Proceedings of RIAO Conference on User-oriented Content-based Text and Image Handling,* 1993, pp. 21-24, Cambridge.

Church, K. and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, 6(1), 1990, pp. 22-29.

Dagan, I., L. Lee and F. Pereira, "Similarity-based method for word sense disambiguation," *Proceedings of the 35th Annual Meeting of ACL,* 1997, pp. 56-63, Madrid, Spain.

Dong, Z. D. and Q. Dong, HowNet, http://www.keenage.com, 1999.

Lin, D. K., "Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity," *Proceedings of ACL/EACL-97*, 1997, pp. 64-71, Madrid, Spain

Lin, D. K., "Extracting collocations from text corpora*,"Proc. First Workshop on Computational Terminology*, 1998, Montreal, Canada.

Lin, D. K., "Using Collocation Statistics in Information Extraction," *Proceedings of the Seventh Message Understanding Conference* (MUC-7), 1998.

Liu, Q.,"The Word Similarity Calculation on <<HowNet>>," *Proceedings of 3rd Conference on Chinese lexicography,* 2002, TaiBei.

Lu, Q., Y. Li and R. F. Xu, "Improving Xtract for Chinese Collocation Extraction," *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 2003, Beijing.

Manning, C. D. and H. Schutze, "Foundations of Statistical Natural Language Processing," *The MIT Press*, 1999, Cambridge, Massachusetts.

Miller, G., WordNet, http://www.cogsci.princeton.edu/~wn/, 1998.

Miller, G and C. Fellbaum, "Semantic networks of English," *In Beth Levin & Steven Pinker (eds.), Lexical and conceptual semantics*, 1992, pp. 197-229.

Pearce, D., "Synonymy in Collocation Extraction," *Proceedings of NAACL'01 Workshop on Wordnet and Other Lexical Resources: Applications, Extensions and Customizations*, 2001.

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

Sun, M. S., C. N. Huang and J. Fang, "Preliminary Study on Quantitative Study on Chinese Collocations*," ZhongGuoYuWen*, No.1, 1997, pp. 29-38, (in Chinese).

Wu, H. and M. Zhou, "Synonymous Collocation Extraction Using Translation Information," *Proceeding of the 41st Annual Meeting of ACL*, 2003.

Yang, E., G. Zhang and Y. Zhang, "The Research of Word Sense Disambiguation Method Based on Co-occurrence Frequency of HowNet," *Communication of COLIPS*, 8(2) 1999, pp. 129-136.

Yao, T., W. Ding and G. Erbach, "CHINERS: A Chinese Name Entity Recognition System for the Sports Domain," *Second SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 55-62.