International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

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

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

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

Vol.16 No.1-2 March/June 2011 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

The editing of this journal is subsidized by Center for Humanities Research National Science Council in 2011.

Aims and Scope

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

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

Membership & Subscriptions

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

Copyright

Cover

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

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 16, no. 1-2, March/June 2011

Contents

Papers

The Association for Computational Linguistics and Chinese Language Processing

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 1-2, March/June 2011, pp. 1-14

Performance Evaluation of Speaker-Identification Systems for Singing Voice Data

Wei-Ho Tsai[*] and Hsin-Chieh Lee[*]

Abstract

Automatic speaker-identification (SID) has long been an important research topic. It is aimed at identifying who among a set of enrolled persons spoke a given utterance. This study extends the conventional SID problem to examining if an SID system trained using speech data can identify the singing voices of the enrolled persons. Our experiment found that a standard SID system fails to identify most singing data, due to the significant differences between singing and speaking for a majority of people. In order for an SID system to handle both speech and singing data, we examine the feasibility of using model-adaptation strategy to enhance the generalization of a standard SID. Our experiments show that a majority of the singing clips can be correctly identified after adapting speech-derived voice models with some singing data.

Keywords: Model Adaptation, Singing, Speaker Identification.

1. Introduction

As an independent capability in biometric applications or as part of speech-recognition systems, automatic speaker-identification (SID) (Rosenberg, 1976; Reynolds & Rose, 1995; Reynolds, 1995; Campbell, 1997; Reynolds *et al*., 2000; Bimbot *et al*., 2004; Nakagawa *et al*., 2004, 2006; Murty & Yegnanarayana, 2006; Matusi & Tanabe, 2006; Beigi, 2011) has been an attractive research topic for more than three decades. It is aimed at identifying who among a set of enrolled persons spoke a given utterance. Currently, existing SID systems operate in two phases, training and testing, where the former models each person's voice characteristics using his/her spoken data and the latter determines unknown speech utterances based on some comparisons between models and utterances. As the purpose of SID is distinguishing one

[*] Department of Electronic Engineering & Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan
Tel: +886-2-27712171 ext. 2257   Fax: +886-2-27317120
E-mail: whtsai@ntut.edu.tw
The author for correspondence is Wei-Ho Tsai.

person's voice from another's, it is worth investigating if an SID system can not only identify speech voices but also singing voices.

There are a number of real applications where an SID system may need to deal with singing voices. For example, if we record the sounds from TV, it is very likely that the recording contains performers speaking then singing or singing then speaking. In such a case, an SID system capable of handling both speech and singing voices would be very useful to index the recording. Another example is when people gather to sing at a Karaoke. It would be helpful to record everyone's performance onto CDs or DVDs to capture memories of the pleasant time. For the audio in CDs or DVDs to be searchable, audio data would preferably be written in separate tracks, each labeled with the respective person. In this case, an SID system capable of identifying both speech and singing voices will be helpful to automate the labeling process.

To the best of our knowledge, there is no prior literature devoted to the problem of using an SID system to identify singing voices. Most related work (Rosenau, 1999; Gerhard, 2004, 2003) has investigated the differences between singing and speech. Some studies have developed methods for singing voice synthesis (Bonada & Serra, 2007; Kenmochi & Ohshita, 2007; Saino *et al.*, 2006; Saitou *et al.*, 2005), and some have discussed how to convert speech into singing (Saitou *et al.*, 2007) according to the specified melody. In this paper, we begin our investigation by evaluating the performance of an SID system trained using speech voices when the testing samples are changed from speech to singing voices. Then, a well-studied model-adaptation strategy is applied to improve the system's capability in handling singing voices. Our final experiments show that a majority of the singing clips can be correctly identified after adapting speech-derived voice models with some singing data.

The rest of this paper is organized as follows. Section 2 reviews a prevalent SID system. Section 3 describes an improved SID system using some singing data to adapt speech-derived voice models. Then, Section 4 discusses our experiment results. In Section 5, we present our concluding remarks.

2. A Popular Speaker-Identification (SID) System

Figure 1 shows the most prevalent SID system currently, stemming from (Reynolds & Rose, 1995). The system operates in two phases: training and testing. During training, a group of N persons is represented by N Gaussian mixture models (GMMs), λ_1, λ_2, …, λ_N. It is found that GMMs provide good approximations of arbitrarily shaped densities of a spectrum over a long span of time (Murty & Yegnanarayana, 2006); hence, they can reflect the vocal tract configurations of individual persons. The parameters of GMM λ_i, composed of means, covariances, and mixture weights, are estimated using the speech utterances of the i-th person. The estimation consists of k-means initialization and Expectation-Maximization (EM)

(Dempster *et al.*, 1977).

Prior to Gaussian mixture modeling, audio waveforms are converted, frame-by-frame, into Mel-scale frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980). The merit of MFCCs lies in the auditory modeling, which has been shown to be superior to other speech-production-based features in numerous studies. Given a test voice sample, the system computes its MFCCs $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2,..., \mathbf{y}_T\}$ and the likelihood probability $\Pr(\mathbf{Y}|\lambda_i)$ for each model λ_i:

$$\Pr(\mathbf{Y} \mid \lambda_i) = \prod_{t=1}^{T} \sum_{k=1}^{K} w_i^{(k)} \cdot \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_i^{(k)}, \mathbf{C}_i^{(k)}) \,, \tag{1}$$

$$\mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_i^{(k)}, \mathbf{C}_i^{(k)}) = \frac{1}{\pi^N \, | \mathbf{C}_i^{(k)} |} \exp\left\{ -\left(\mathbf{y}_t - \boldsymbol{\mu}_i^{(k)}\right)' \mathbf{C}_i^{(k)-1} \left(\mathbf{y}_t - \boldsymbol{\mu}_i^{(k)}\right) \right\} \tag{2}$$

where K is the number of mixture Gaussian components; $w_i^{(k)}, \boldsymbol{\mu}_i^{(k)}$, and $\mathbf{C}_i^{(k)}$ are the k-th mixture weight, mean, and covariance of model λ_i, respectively; and prime (′) denotes the vector transpose. According to the maximum likelihood (ML) decision rule, the system decides in favor of person I^* when the condition in Eq. (3) is satisfied:

$$I^* = \arg\max_{1 \le i \le N} \Pr(\mathbf{Y} \mid \lambda_i) \,. \tag{3}$$

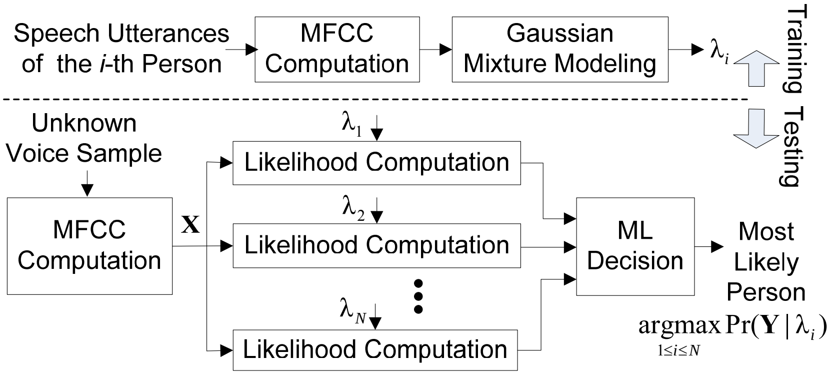


Figure 1. The most prevalent SID system.

3. An SID System Based on Model Adaptation for Singing Voices

Our experiments, discussed in detail in Section 4, find that the above-described SID system performs rather poorly in identifying singing voices of enrolled persons, since a person's singing voice can be significantly different from his/her speech voice. To see if the system can be improved, we apply a well-studied model-adaptation strategy to adapt each person's GMM using some of his/her singing voice data. The adaptation is based on the Maximum A Posterior (MAP) estimation of GMM parameters (Reynolds *et al.*, 2000). We assume that the amount of available singing data for adaptation is very limited; hence, only the mean vectors of GMMs are adapted. For the *i*-th person's GMM, the mean vector of the *k*-th mixture is updated using

$$\hat{\boldsymbol{\mu}}_i^{(k)} = \frac{\tau_i^{(k)}}{\tau_i^{(k)} + \gamma} \bar{\boldsymbol{\mu}}_i^{(k)} + \frac{\gamma}{\tau_i^{(k)} + \gamma} \boldsymbol{\mu}_i^{(k)} , \tag{4}$$

$$\tau_i^{(k)} = \sum_{\ell=1}^{L} \Pr(k \mid \mathbf{x}_\ell, \lambda_i) , \tag{5}$$

$$\bar{\boldsymbol{\mu}}_i^{(k)} = \frac{1}{\tau_i^{(k)}} \sum_{\ell=1}^{L} \Pr(k \mid \mathbf{x}_\ell, \lambda_i) \mathbf{x}_\ell , \tag{6}$$

$$\Pr(k \mid \mathbf{x}_\ell, \lambda_i) = \frac{w_i^{(k)} \mathcal{N}(\mathbf{x}_\ell; \boldsymbol{\mu}_i^{(k)}, \mathbf{C}_i^{(k)})}{\sum_{n=1}^{K} w_i^{(n)} \mathcal{N}(\mathbf{x}_\ell; \boldsymbol{\mu}_i^{(n)}, \mathbf{C}_i^{(n)})} , \tag{7}$$

where $\mathbf{x}_\ell$, $1 \le \ell \le L$, are the MFCCs of the available adaptation (singing) data, $\hat{\boldsymbol{\mu}}_i^{(k)}$ is the resulting mean vector after the adaptation, $\mathcal{N}(\cdot)$ is a multivariate Gaussian density function, and γ is a weighting factor of the *a priori* knowledge to the adaptation data. The block diagram of the system based on MAP adaptation is shown in Figure 2.

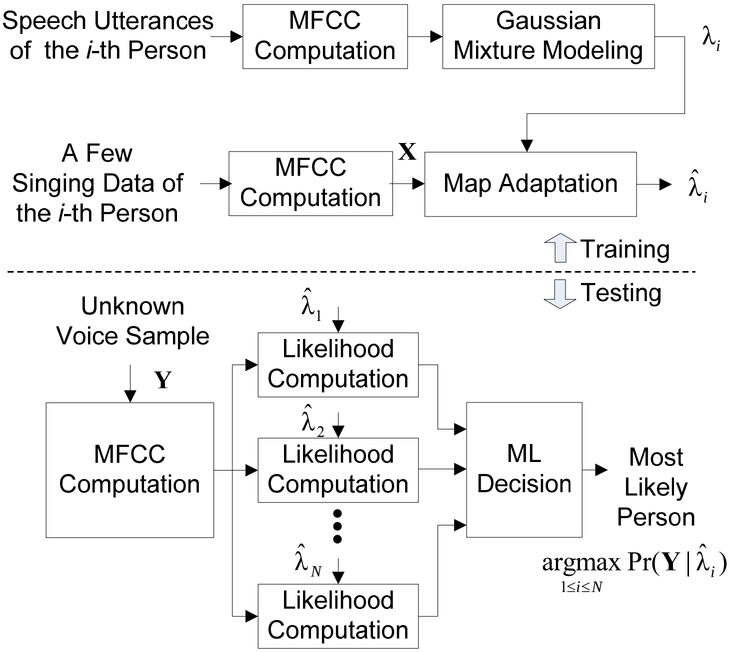


Figure 2. An SID system based on MAP adaptation of a speaker GMM to a singer GMM.

4. Experiments

4.1 Voice Data

We created a database of test recordings ourselves, since no public corpus of voice data currently meets the specific criteria we set up for this study. The database contains vocal recordings by twenty male participants between the ages of 20 and 39. We asked each person to perform 30 passages of Mandarin pop songs using a karaoke machine in a quiet room. All of the passages were recorded at 22.05 kHz, 16 bits, in mono PCM wave. The karaoke accompaniments were output to a headset and were not captured in the recordings. The duration of each passage ranges from 17 to 26 seconds. We denoted the resulting 600 recordings by DB-Singing. Next, we asked each person to read the lyrics of the 30 song passages at a normal speed. All of the read utterances were recorded using the same conditions as those in DB-Singing. The resulting 600 utterances were denoted as DB-Speech.

For ease of discussion in the following sections, we use a term “parallel” to represent the association between a speech utterance and singing recording that are based on the same texts. For example, when the texts are in turn spoken and sung by a person, the speech utterance is referred to as the “parallel” speech utterance of the resulting singing recording, and *vice-versa*. In addition, for use in different purposes, we divided DB-Singing into two subsets, DB-Singing-1 and DB-Singing-2, where the former contains the first 15 recordings per person and the latter contains the last 15 recordings per person. Similarly, DB-Speech was divided into subsets DB-Speech-1 and DB-Speech-2, where the former contains the first 15 speech utterances per person and the latter contains the last 15 speech utterances per person.

4.2 Experiment Results

We used the 15 speech utterances per person in DB-Speech-1 to train each person-specific GMM, and tested the singing recordings in DB-Singing-2. To obtain a statistically-significant experimental result, we repeated the experiment using the 15 speech utterances in DB-Speech-2 to train each person-specific GMM and tested the singing recordings in DB-Singing-1. The number of Gaussian components used in each GMM was tuned to optimum according to the amount of training data. The SID performance was assessed with the accuracy:

$$\text{SID Accuracy (in \%)} = \frac{\text{\#correctly-identified recordings}}{\text{\# testing recordings}} \times 100\% \,.$$

In addition, to make sure if the system could work well for the conventional SID task, we also evaluated the SID performance using DB-Speech-1 to train each person-specific GMM and tested the speech utterances in DB-Speech 2. Also, in order for the result to be statistically significant, the experiments were repeated using DB-Speech-2 to train each person-specific GMM before testing the speech utterances in DB-Speech-1. Table 1 shows the SID results. We can see from Table 1 (a) and (b) that the system trained using a set of speech data can perfectly identify the speakers of another set of speech data. Nevertheless, the system fails to identify most persons' voices in DB-Singing-1 and DB-Singing-2. Such poor results indicate the significant differences between most people’s speaking and singing voices.

Table 1. Accuracies of the SID systems trained using speech data

(a) System trained using DB-Speeech-1

Testing Data	SID Accuracy (%)
DB-Speech-2	100.0
DB-Singing-2	17.7

(b) System trained using DB-Speeech-2

Testing Data	SID Accuracy (%)
DB-Speech-1	100.0
DB-Singing-1	16.3

Table 2 shows the confusion matrix of the SID results in Table 1. The columns of the matrix correspond to the ground-truth of the singing recording, while the rows indicate the hypotheses. It can be seen from Table 2 that there are a large number of persons whose voice recordings were completely mis-identified. There were only a few people, *e.g.*, #4 and #9, whose singing recordings mostly could be identified well. Further analysis found that persons #4 and #9 are not good at singing, and often cannot follow the tune. They cannot modify their voices properly to make the singing melodious either. Perhaps due to a lack of singing practice, persons #4 and #9 do not change their normal speech voices too much during singing; hence, the system trained using their speaking voices can identify their singing voices well.

To gain insight into the SID errors with respect to different persons, we analyzed the spectrograms of the singing recordings and their parallel speech utterances produced by persons #9 and #10. The waveforms were divided into segments of 512 samples with 50% overlap for the computation of short-term Fourier transform. We can see from Figure 3 (a) and (b) that the formant structure of #9's singing recording is relatively similar to that of his speech utterance, compared with the case of #10, shown in Figure 3 (c) and (d). There is almost no vibrato in #9's singing voice. This is consistent with the observation that #9’s voice does not differ too much from speech to singing; thus, it can be handled with speech-derived GMM.

Table 2. Confusion matrix of the SID results in Table 1.

Actual Person Index	Hypothesized Person Index																				Accuracy (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	**3**	1	0	1	2	1	1	0	2	2	3	4	1	1	2	1	3	0	2	0	10.0
2	2	**1**	1	0	1	2	2	1	1	3	2	1	0	0	0	0	5	3	4	1	3.3
3	2	0	**2**	1	1	3	4	2	0	1	4	1	1	3	1	0	0	4	0	0	6.7
4	0	0	0	**25**	0	1	0	1	2	0	0	0	0	1	0	0	0	0	0	0	83.3
5	0	0	3	0	**3**	1	2	1	5	0	6	2	0	0	3	0	0	3	1	0	10.0
6	5	1	0	0	2	**3**	1	4	0	0	3	1	1	0	3	4	1	1	0	0	10.0
7	1	0	0	0	2	3	**2**	0	0	0	3	6	1	0	0	0	8	0	0	4	6.7
8	2	1	0	0	7	0	3	**4**	2	0	5	0	0	4	3	4	1	1	1	0	13.3
9	0	0	0	0	1	0	0	0	**28**	1	0	0	0	0	0	0	0	0	0	0	93.3
10	3	2	0	5	0	0	1	0	2	**2**	1	1	1	4	0	0	5	0	0	3	6.7
11	0	0	4	0	2	0	3	0	1	1	**3**	0	6	0	1	1	2	0	6	0	10.0
12	2	1	1	1	0	3	0	0	5	0	0	**2**	1	1	0	4	6	2	0	1	6.7
13	0	1	1	0	1	3	4	1	1	1	1	2	**3**	2	6	0	0	1	0	2	10.0
14	1	1	1	2	4	0	8	0	0	5	1	0	0	**5**	0	0	1	1	0	0	16.7
15	0	2	6	1	1	0	1	1	0	0	0	3	5	0	**2**	3	3	0	0	2	6.7
16	0	0	0	4	0	7	0	1	1	1	3	3	0	0	0	**3**	2	4	1	0	10.0
17	4	0	8	0	0	1	1	1	0	1	3	0	0	4	0	4	**2**	1	0	0	6.7
18	2	1	1	0	1	1	3	4	0	7	0	0	5	1	1	1	0	**2**	0	0	6.7
19	0	0	0	0	5	1	1	0	1	5	4	0	0	0	2	2	2	3	**4**	0	13.3
20	1	0	8	0	0	2	3	2	1	1	1	1	0	0	0	0	4	0	1	**5**	16.7

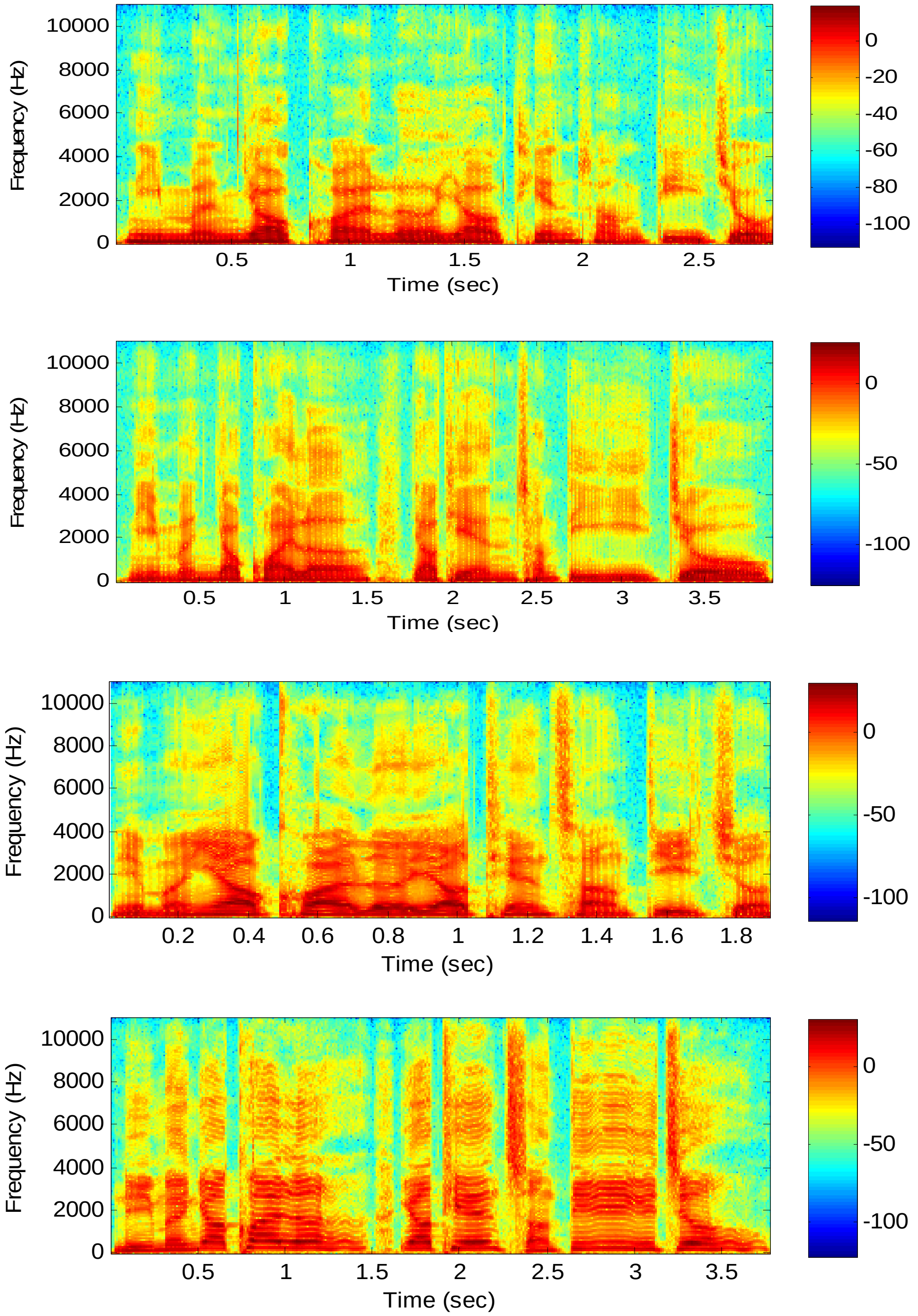
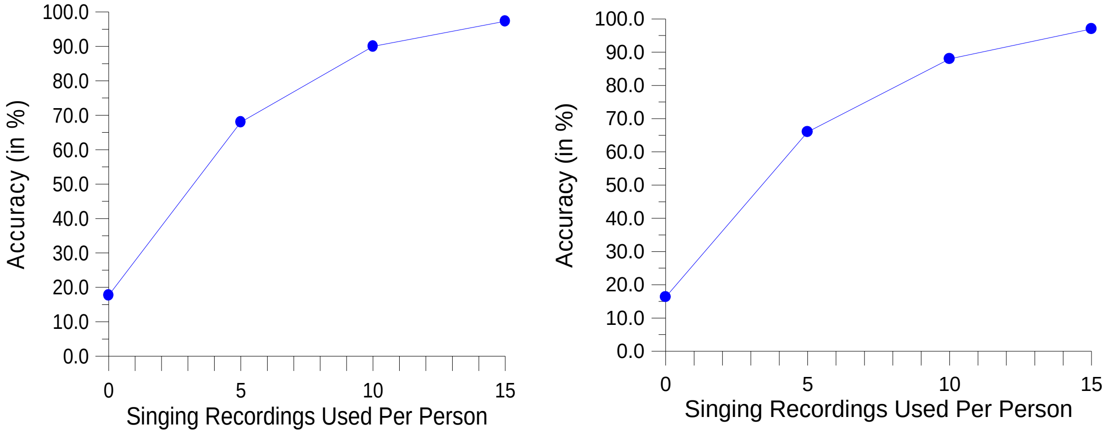


Figure 3. (a) spectrogram of a speech utterance produced by person #9, (b) spectrogram of a singing recording produced by person #9, (c) spectrogram of a speech utterance produced by person #10, and (d) spectrogram of a singing recording produced by person #10, where all the singing recordings and speech utterances are based on the same lyrics: "/ni/ /man/ /iau/ /kuai/ /le/ /iau/ /tian/ /chang/ /di/ /jiou/".

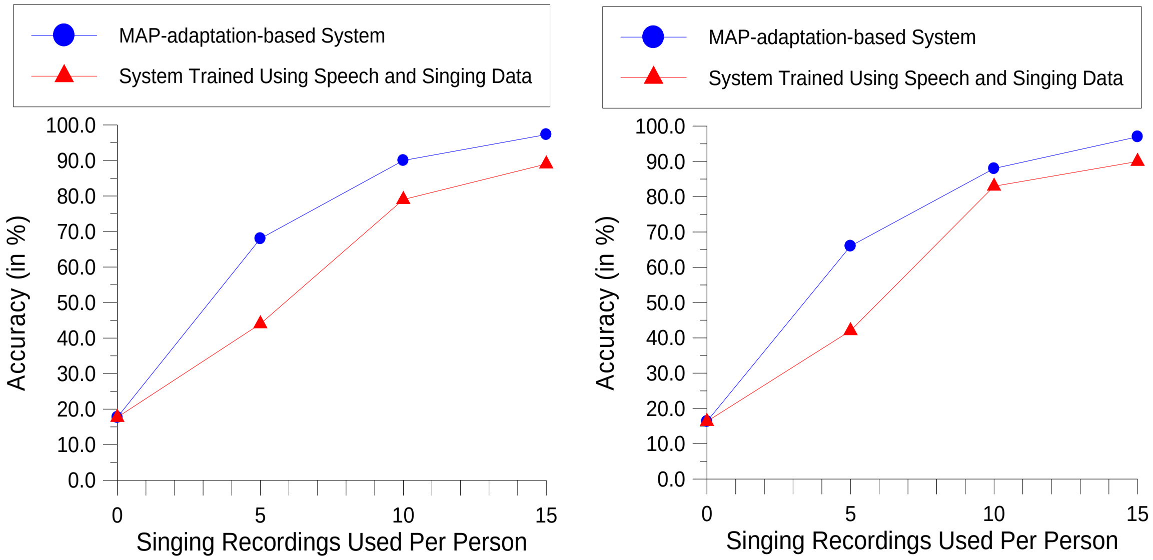
Next, the SID performance of the "MAP-adaptation-based system" described in Sec. 3 was evaluated. We used the 15 speech utterances per person in DB-Speech-1 to train the person-specific GMMs. Each GMM then was adapted using *J* randomly-selected singing recordings per person in DB-Singing-1, where *J* = 5, 10, and 15. Based on the adapted GMMs, the system identified the persons of the singing recordings in DB-Singing-2. In addition, to obtain statistically-significant experiment results, we repeated the experiment by using DB-Speech-2 as the training data, DB-Singing-2 as the adaptation data, and DB-Singing-1 as the testing data. The identification accuracy then was computed as the percentage of the correctly-identified recordings. Figure 4 shows the SID accuracies obtained with the MAP-adaptation-based system. It can be seen from Figure 4 that, as expected, the SID accuracies increase with the increase in the amount of singing data used.



(a)Testing DB-Singing-1 ***(b) Testing DB-Singing-2***

Figure 4. SID accuracies obtained with the MAP-adaptation-based System.

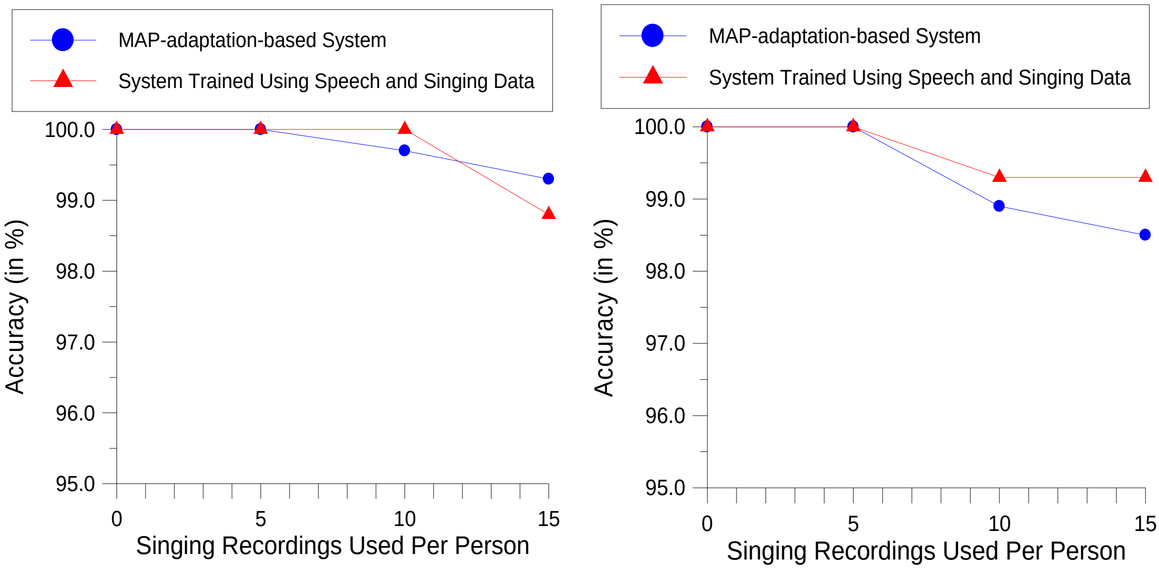
As the MAP-adaptation-based system uses more voice data than the system using speech data only, it is worth comparing the SID performance of the MAP-adaptation-based system with that of the system trained using both speech data and singing data. We thus generated an SID system using 15 utterances plus *J* singing recordings per person in Gaussian mixture modeling. Figure 5 shows our experiment results. We can see from Figure 5 that the system trained using both speech data and singing data cannot achieve comparable performance to the MAP-adaptation-based system, especially when the amount of singing data is small. This may be because a GMM trained using a mix of speech and singing data tends to model the common voice characteristics of speech and singing, but overlooks their individual differences.



(a)Testing DB-Singing-1 ***(b) Testing DB-Singing-2***

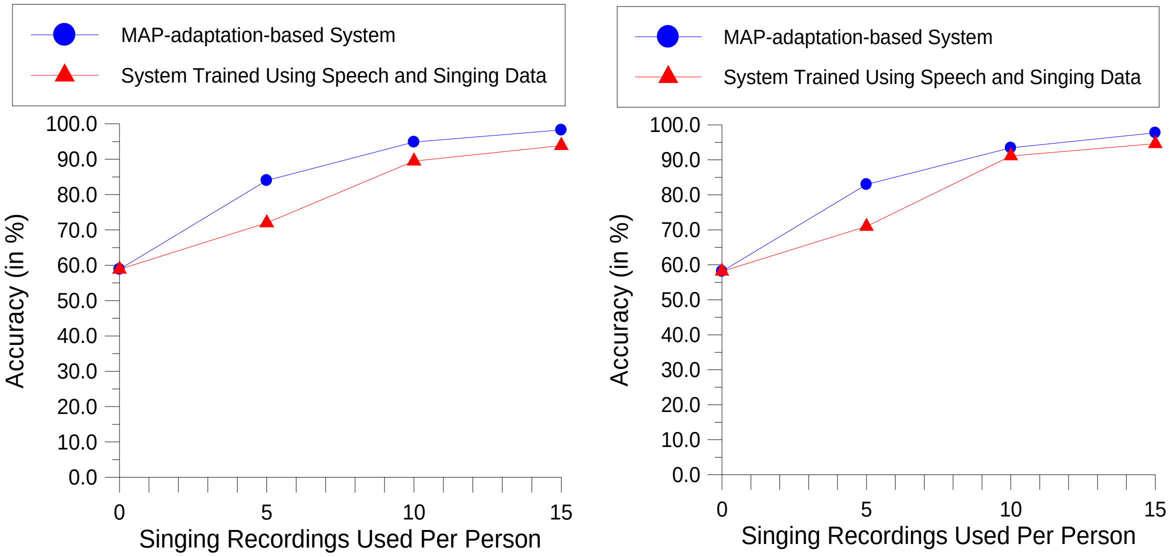
Figure 5. Comparison of the SID performance of the MAP-adaptation-based system with that of the system trained using both speech data and singing data.

In addition, it is worth examining if the MAP-adaptation-based system is still capable of identifying speech data, since its models have been adapted to handle singing data. Figure 6 shows the SID accuracies of testing speech utterances using the MAP-adaptation-based system. For the purpose of comparison, we also evaluated the SID accuracies obtained with the system trained using both speech and singing data. It can be seen from Figure 6 that both of the systems work well in identifying speech utterances. This indicates that the GMMs in the MAP-adaptation-based system do not lose the essence of covering the speaking voice characteristics after they are adapted to cover the singing voice characteristics. Figure 7 presents the accuracies of identifying all of the speech utterances and singing recordings in our database. We can see from Figure 7 that the MAP-adaptation-based system performs better overall than the system trained using both speech and singing data.



(a) Testing DB-Speech-1 ***(b) Testing DB-Speech-2***

Figure 6. Accuracies of identifying speech utterances



(a) Testing DB-Speech-1 and DB-Singing-1 (b) Testing DB-Speech-2 and DB-Singing-2

Figure 7. Accuracies of identifying both speech utterances and singing recordings.

5. Conclusion

In this study, the problem of speaker identification has been extended from identifying a person's speech utterances to identifying a person's singing recordings. Our experiment found that a standard SID system trained using speech utterances fails to identify most singing data, due to the significant differences between singing and speaking for a majority of people. In order for an SID system to handle both speech and singing data, we examine the feasibility of applying a well-known model-adaptation strategy to enhance the generalization of a standard SID. The basic strategy is to use a small sample of the singing voice to adapt each speech-derived GMM based on MAP estimation. The experiments show that, after the model adaptation, the system can identify a majority of the singing clips, while retaining the capability of identifying speech utterances.

Although this study shows that a speech-derived SID system can be improved significantly through the use of a model-adaptation strategy, the system pays the cost of acquiring the singing voice data from each person. In realistic applications, acquiring singing voice data in the training phase may not be feasible. As a result, further investigation on robust audio features invariant to speech and singing would be needed. Our future work will focus on this topic and extend our voice database to a larger scale.

Acknowledgement

This research was partially supported by the National Science Council, Taiwan, ROC, under Grant NSC 98-2622-E-027-035-CC3.

References

Beigi, H. (2011). *Fundamentals of Speaker Recognition*. New York: Springer. ISBN 978-0-387-77591-3, 2011.

Bimbot, F. J., Bonastre, F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., & Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP J. Appl. Signal Process.*, 430-451.

Bonada, J., & Serra, X. (2007). Synthesis of the singing voice by performance sampling and spectral models. *IEEE Signal Processing Magazine*, 24(2), 67-79.

Campbell, J. P. (1997). Speaker recognition: a tutorial. *Proc. IEEE*, 85(9), 1437-1462.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Sspeech, Signal Process.*, 28, 357-366.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.*, 39, 1-38.

Gerhard, D. (2002). Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing. *Journal of the Canadian Acoustical Association*, 30(3), 152-153.

Gerhard, D. (2003). *Computationally measurable differences between speech and song*. Ph.D. dissertation, Simon Fraser University.

Kenmochi, H., & Ohshita, H. (2007). VOCALO-ID – commercial singing synthesizer based on sample concatenation. In *Proc. Interspeech*, 4011-4010.

Matusi, T., & Tanabe, K. (2006). Comparative study of speaker identification methods: DPLRM, SVM and GMM. *IEICE Trans. on Information and Systems*, E89-D(3), 1066-1073.

Murty, K. S. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker verification. *IEEE Signal Process. Lett.*, 13(1), 52-55.

Nakagawa, S., Zhang, W., & Takahashi, M. (2004). Text-independent speaker recognition by combining speaker specific GMM with speaker adapted syllable-based HMM. In *Proc. ICASSP*, I, 81-84.

Nakagawa, S., Zhang, W., & Takahashi, M. (2006). Text-independnt/text-prompted speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM. *IEICE Trans. on Information and Systems*, E89-D(3), 1058-1064.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Commun.*, 17(1-2), 91-108.

Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.*, 3(1), 72-83.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Dig. Signal Process.*, 10(1-3), 19-41.

Rosenau, S. (1999). An analysis of phonetic differences between German singing and speaking voices. In *Proc. 14th Int. Congress of Phonetic Sciences* (ICPhS).

Rosenberg, A. E. (1976). Automatic speaker verification: A review. In *Proc. IEEE*, 64(4), 475-487.

Saitou, T., Goto, M., Unoki, M., & Akagi, M. (2007). Speech-to-singing synthesis: vocal conversion from speaking voices to singing voices by controlling acoustic features unique to singing voices. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2007)*, 215-218.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Comm.*, 46, 405-417.

Saino, K., Zen, H., Nankaku, Y., Lee, A., & Tokuda, K. (2006). HMM-based singing voice synthesis system. In *Proc. Int. Conf. Spoken Lang. Process.* (*ICSLP*), 1141-1144.

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 1-2, March/June 2011, pp. 15-26

Resolving Abstract Definite Anaphora in Chinese Texts

Tyne Liang[*] and Jyun-Hua Cheng[*]

Abstract

Anaphora is a rhetorical device commonly used in written texts. It denotes the use of terms referring to previously-mentioned entities, concepts, or events. In this paper, the definite anaphora in Chinese texts is addressed and empirical approaches to tackle abstract anaphors are presented. The resolution is built on the association between target anaphors and the corresponding referents in their multiple-type features extracted from different levels of discourse units. Experimental results show that features extracted from clauses are more useful than those extracted from sentences in referent identification. Besides, the presented salience-based model outperforms the SVM-based model no matter whether the best set of extracted features is employed or not.

Keywords: Anaphora Resolution, Chinese Text, Definite Anaphora, Feature Extraction

1. Introduction

1.1 Motivation

Anaphora is an instance of an expression referring to the preceding utterances. Effective anaphora resolution enhances understanding of a text and facilitates many applications of natural language processing. The resolution involves anaphor recognition and referent recognition. In Chinese texts, anaphors can be missing or be present as pronouns, demonstratives and definite descriptions. Common pronouns are like “他” (“he, him”), “她” (“she, her”), “它” (“it”), “我們” (“we, us”), “他們” (“they, them”); demonstratives are “這” (“this”), ”那” (“that”) and definite description are like the pattern “這+[quantifier]+noun phrase.” Without concerning zero anaphora, about 54% of the explicit anaphors are pronouns, 40% are definite descriptions, and 6% are demonstratives in a corpus containing 20 news articles.

Essentially, the challenges involved with Chinese anaphora resolution are attributed to the complexities of Chinese sentence structures. It is known that although a Chinese sentence

[*] College of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.
E-mail: tliang@cs.nctu.edu.tw; sunrise0406.iit97g@g2.nctu.edu.tw

features the subject-verb-object order, the sentence may be formed by a series of verbs or by pronoun or subject dropping, thus making sentence parsing difficult. Moreover, there is no blank space between adjacent words in Chinese sentences, making word or noun phrase identification difficult. Unlike most previous research projects focusing on non-abstract anaphora resolution, this paper addresses the definite anaphora in Chinese written texts and presents empirical parser-free approaches to resolve abstract anaphors, like ”這項方案” (“this plan”). The resolution is based on the linking between anaphors and their referents in multiple aspects of contextual, semantic and surface features. Among them, semantic features are extracted with the help of three outer resources, namely, Tongyici Cilin[1] (TYCC for short), CKIP Lexicon[2], and Google search results[3]. Additionally, the features extracted from different discourse units are investigated and the best feature set is verified at referent identification. In the experiments, both SVM-based and salience models are implemented for model comparison. Experimental results show that the features extracted from clauses are more useful than those extracted from sentences for anaphora resolution. Besides, the presented salience-based model outperforms the SVM-based model regardless of whether the best set of extracted features is employed or not.

1.2 Abstract Definite Anaphora

In Chinese texts, a definite anaphor contains a demonstrative (tagged as “Nep” by CKIP Chinese word segmentation system[4] (CKIP tagger for short)) followed with an optional quantifier (tagged as “Nf”) and a noun phrase. Lexicons with Nep-tag are like ”這, 此, 其, 那, 什麼, 其中, 個中, 甚, 啥, 哪, 斯, 甚麼”. Such anaphora is similar to the definite description anaphora in English texts in which the anaphors are composed of the definite article “the” followed by a noun phrase. In fact, there is no definite article in Chinese, so we may treat the definite noun “the+noun phrase” and demonstrative noun “this or that+noun phrase” to be the same in Chinese texts. In this paper, we focus on the “這+[quantifier]+ abstract-type noun phrase” anaphor since it is frequently expressed in Chinese texts. The abstract noun phrases are defined and categorized according to CKIP Lexicon. Table 1 shows some target anaphor instances we identified from our corpus.

Abstract definite anaphora can be expressed in two ways. One is direct anaphora in which both the referent and the anaphor contain the same head noun. For example, both anaphor ”這項方案” (this plan”) and its referent ”學生停車方案” (“student parking plan”)

[1] TongyiciCilin extended version: http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

[2] CKIP (Chinese Knowledge Information Processing Group) Lexicon: http://www.aclclp.org.tw/use_ckip_c.php

[3] Google: http://www.google.com.tw

[4] CKIP Chinese word segmentation system: http://ckipsvr.iis.sinica.edu.tw/

contain the head noun “方案”. The other is indirect anaphora in which the anaphor “這項方案” and its referent “課後輔導” (“after-school assistance”) do not contain the same head noun and their resolution has to be done by considering their linking in contextual, syntactic and semantic structures. More challenges associated with indirect anaphora resolution come from the boundary identification for those referents crossing multiple clauses or sentences. For example, “這項方案” refers to “學校計劃將強制要求所有過重學生放學後都要留下來，參加二小時的體能訓練，直到學期結束。” (“The school will require all overweight students to remain after school to participate in two hours of physical training until the end of the term”). Besides, it is observed that Chinese texts are usually not written with accurate usage of punctuation marks; thus identifying such types of referents in Chinese texts is harder than in English texts.

Table 1. Some abstract instances and their CKIP Lexicon categories

Category	Example
特徵 (Characteristics)	想法 thought, 行爲 behavior
文明 (Enlightenment)	問題 problem, 決議 decision
法則 (Principles)	方式 way, 制度 system
社會活動 (Social_activities)	比賽 competition, 會議 meeting
法人 (Corporation)	社會 society, 學校 school
名稱 (Nomenclature)	職位 Job, 名字 name
狀況 (Situations)	情況 situation, 現象 phenomenome
社會關係 (Social_relation)	關係 relation, 情誼 frienship
財務關係 (Monetary_relation)	經費 funding, 收入 income
權力 (Authority)	政權 regime, 主權 sovereignty
疾病 (Illness)	病 disease, 病變 lesion
時間 (Temporal_relation)	期間 period, 階段 stage
事件 (Events)	行動 action, 過程 process
動名詞 (nominal verb)	調查 investigation, 會談 meeting

2. Related Work

In general, abstract anaphora can be resolved using pattern rules, statistical or hybrid approaches. For example, Byron (2002) presented PHORA to resolve pronouns referring to abstract entities in a dialogue corpus. The resolution is based on the semantic constraints imposed on verbs, predicate noun phrases and predicative adjectives in sentences. Later

Navarretta (2004) extended these semantic constraints with dialogue structures to resolve inter-sentential pronominal anaphors in Danish texts. Beside the rule-based approaches, Strube and Muller (2003) presented a decision tree resolution to identify both NP-type and non-NP antecedents with the employment of 23 features including NP and coreference features. The supervised learning method is also found in (Yang *et al.*, 2006) for pronoun resolution by taking into account the coreferential information of a candidate. In addition, some researches have tried to resolve indirect nominal anaphora via web search (Bunescu, 2003), WordNet (Poesio *et al.*, 2002), or statistical models like multi-layer perceptrons (Poesio *et al.*, 2004).

In contrast to the prevalent discussion on English anaphora, effective approaches to tackle Chinese abstract anaphora have not been widely discussed. Either parsing-tree based or machine based approaches have been primarily presented to resolve noun-phrase type references (Yeh & Chen, 2004; Zhao & Ng, 2007; Wu & Liang, 2008, 2009, 2011; Kong & Zhou, 2010). Nevertheless, it is observed that the average length of the referents includes more than one clause in a corpus like the news reports we extracted from Academia Sinica Balanced Corpus[5] (ASBC for short). Hence, this paper is motivated to present some feasible methods to facilitate such type referent identification.

3. Corpus Preparation

3.1 Corpus Tagging

The corpus we used for developing the resolution approach is extracted from ASBC, a corpus used for modern Chinese text processing research. For each extracted text, we manually tagged the target anaphors and filter out those “這” without a following noun phrase. For example, we would not use“這是” (“this is”), “這可能是” (“this might be”),...etc. We did not tag those ”這” if that functioned as discourse-new or cataphor.

The corpus contained 885 texts and out of which there were 24062 sentences and 82783 clauses identified by any of punctuation marks (“。？！；”) and (”。？！；，”) respectively. Each clause was tagged with a sentence number s_i and a clause serial number c_j. Each clause was also manually tagged with < ana$_i$> or < ref$_i$> if an anaphor *i* or a referent for anaphor *i* was found in that clause. There were total 1538 definite abstract anaphor instances occurring in the corpus. Followings are three tagged examples in which referents were shown in italic form and anaphors were shown with underlines.

Example a：淡大自本學期開始<s_1, c_1>，實施*學生收費停車方案*<s_1, c_2><ref$_1$>。這項收費停車方案<s_2, c_3> <ana$_1$>，…。

Example b：*天然氣*這項乾淨的能源<s_1, c_1><ana$_1$ ><ref$_1$> …。

[5] Academia Sinica Balanced Corpus: http://www.sinica.edu.tw/SinicaCorpus/

Example c：*學校計劃將強制要求所有過重學生放學後都要留下來*<s$_1$, c$_1$><ref$_1$>，*參加二小時的體能訓練*<s$_1$, c$_2$><ref$_1$>，*直到學期結束*<s$_1$, c$_3$><ref$_1$>。這項方案已經校務會議通過<s$_2$, c$_4$> <ana$_1$>, 將在九十年學年開始實施。

Table 1 lists the 14 categories defined by CKIP Lexicon and some instances identified in our corpus. There are some observations from our tagged corpus. First, there might be multiple referents for a tagged abstract anaphor. In our corpus, 25% of the anaphor instances referred to more than one referent. 59% of the tagged referents contained more than one clause. 52% of the referents occurred in three clauses away from their anaphors and more referents occurred in the preceding sentences than the ones in the same sentences. Besides, 90% of the addressed anaphors and their referents were far away in three sentences. In this paper, the tagged referents were should be in consecutive clauses if they are referred to the same anaphor.

3.2 The Target Referent

Table 2 lists the statistical data of tagged anaphors and their referents. It is found that almost one-third of the tagged anaphors are either characteristics-type or enlightenment-type. The referents for situation-type anaphors contain more clauses than the ones for other types.

Table 2. Anaphors counts and referents lengths

Categories	Anaphors	Average referents length (in clause)
Characteristics	350	2.23
Enlightenment	205	2.79
Principles	231	2.44
Social_activities	106	1.21
Corporation	48	1.46
Nomenclature	12	1.09
Situations	130	3.34
Social_relation	9	2.7
Monetary_relation	28	2.71
Authority	5	1.4
Illness	11	1.73
Temporal_relation	65	2.25
Events	155	2.79
Nominal verb ([+nom])	148	2.07

4. The Proposed Resolution

Figure 1 is the presented resolution which involves POS tagging, anaphor recognition, feature extraction, and referent identification with the employment of three outer resources and processing tools, like a CKIP tagger and a well-designed NP-chunker. The three resources are CKIP Lexicon, TYCC and web search result which is a set of words extracted from Google's snippets. All of these resources will be used for semantic computation for identifying both anaphors and referent candidates.

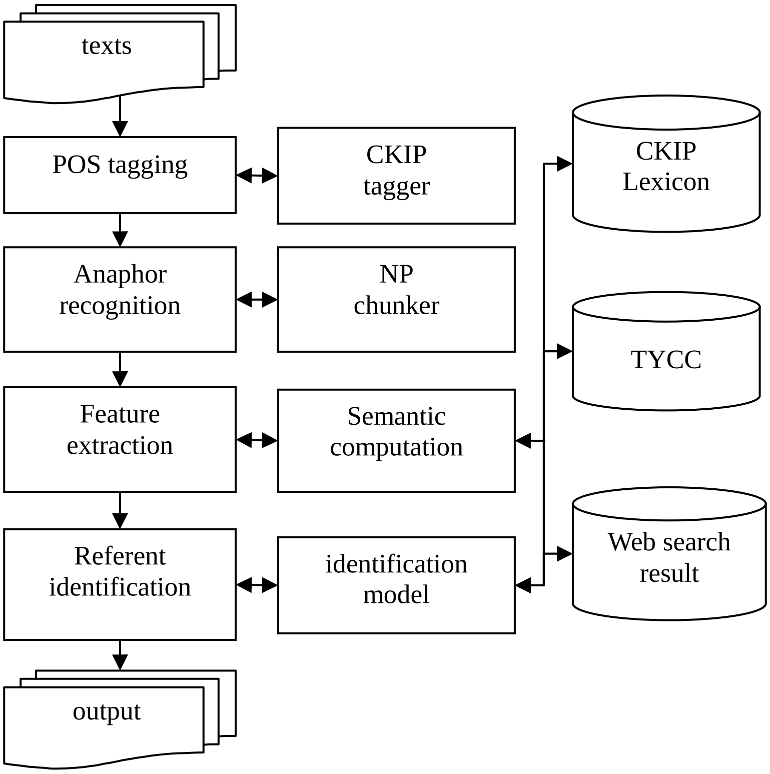


Figure 1. System architecture

4.1 Anaphor Recognition

The anaphor recognition is implemented by a finite-state-machine based NP-chunker (Yu, 2000). Following are some identified anaphors:

(1) 這/this(Nep) 個(Nf) 國際/international(Nc) 金融/finance(Na) 中心/center(Nc)

(2) 這/this(Nep) 種(Nf) 方式/way(Na) 進行/processing(A) 的(DE) 消費/consuming(Na) 商品/product(Na) 交易/trade(Na)

(3) 這/this(Nep) 項(Nf) 國小/elementary school(Nc) 教師/teacher(Na) 心得/report(Na) 公開/publication(A) 發表會/meeting(Na)

(4) 這/this(Nep) 個(Nf) 妥協/compromising(VA)

Afterwards, the last word of an anaphor will be extracted as the head noun and will be checked as to whether it is listed or not as an abstract object in CKIP Lexicon. The experimental results on 1538 anaphor instances show that the presented anaphor recognizer can yield 89.99% accuracy. The failures are summarized into three types as follows:

a. Verbal nominalization: our chunker extracts the anaphor "義工" only, rather than "義工培訓" out of "參加/participate(VC) 這(Nep) 次(Nf) 義工/volunteer(Na) 培訓/train(VC)[+nom]“

b. Complex sentence structure: for example, the correct anaphor is "超級省油車比賽", not "國際自動機工程" in the clause "這/this(Nep) 項(Nf)由/from(P)國際/international(Nc)自動機/automobile(Na)工程/engineering(Na)學會/academic society(Nc)中華民國/ROC(Nc)分會/sub- academic society (Nc)舉辦/hold(VC)的(DE)超級/super(A)省/economic(VJ)油/oil(Na)車/car(Na)比賽/competition(Na)"

c. Inverted sentence: the correct anaphor should be "事" rather than "事研究院方面" in a sentence like "這(Nep) 事/thing(Na) 研究院/research institute(Nc) 方面(Na) 也/too(D) 漫無頭緒(VH)/with no idea about".

4.2 Feature Extraction

It is found that 90% of the referents in our training corpus are within the distances of three sentences away from their corresponding anaphors. So the clauses within this distance are considered as candidate referents. Candidates also include the clauses like "中東這個地區" (Middle-east this area). Here referent ”中東” (Middle-east) is in the same clause of its anaphor "這個地區".

Table 3 lists the four types of the 10 features used in referent recognition. Among them, the thresholds for the distance and similarity features are measured by chi-square test so that each feature value is either one or zero. *Dice_Coefficient* is used to compute the semantic similarity between the words in candidate clause C and the words in anaphor A by measuring how many common nouns, proper nouns, location names, temporal lexicons and verbs are in common.

Table 3. The extracted features

Feature type	Feature description C: candidate clause, A: anaphor clause
Location	C and A are in the same sentence
	C and A are in the same clause
Distance	C and A are within the threshold distance in the terms of sentences
	C and A are within the threshold distance in the terms of clauses
Lexicon	C contains all words of A
	C contains some words of A
	C contains verbs occurring in A
Semantic	C and A are similar enough by computing *Dice_Coefficient*
	C and A contain the same sentential topic word
	C contains the words frequently occurring in text

$$\text{Dice\_Coefficient} = \frac{2|C \cap A|}{|C| + |A|}$$

$$|C \cap A| = \sum \text{Related}(c_i, a_j)$$

$$\text{Related}(c_i, a_j) = \begin{cases} 1, \text{ if } c_i = a_j \\ \quad \text{or CKIP}(c_i) = \text{CKIP}(a_j) \\ \quad \text{or TONYI}(c_i) = \text{TONYI}(a_j) \\ \quad \text{or } c_i \text{ in web(A)} \\ 0, \text{otherwise} \end{cases}$$

C：set of words in candidate clause

A：set of words in anaphor

CKIP(x)：CKIP label for word x

TYCC (x)：TYCC label for word x

Web (C)：search result words of C

It is noticed that the computation is based on word expansion using the mentioned TYCC, CKIP Lexicon and the words extracted from web search results. TYCC contains 77270 words, each of them being represented with one code and tagged with five labels, representing five levels of word categorization. We use the chi-square test to select an appropriate category level of words for word expansion. CKIP Lexicon contains 14935 words as abstract-type words and the words of the same type are treated as related words. For those words neither in TYCC nor in CKIP Lexicon, they will be expanded using search results. The expansion is built with the employment of anaphors and their sentential-topic words as queries to the search engine, Google. From the retrieved 100 snippets with respect to each query, we used chi-square test to find those words frequently co-occurring with the queries.

The topic word feature is employed by assuming that an anaphor and its referent may address the same topic in neighboring sentences. The sentential topic words are identified using the centering-theory based method (Pan, 2008) with which 76.59% F-score was yielded on an experiment containing 88 sentences. The employment of the feature of frequent words is based on the assumption that main concepts in an article may be mentioned repeatedly. In this paper, the main concept words are selected by evaluating the occurrence frequencies of those nouns and verbs in an article. The number of frequent words is also decided by chi-square test.

4.3 Referent Identification

We randomly selected 708 articles containing 1226 target anaphors as our training corpus and use the remaining 171 articles (containing 241 target anaphors) as the testing corpus. The candidate referents are those clauses in the distance of three sentences ahead of the anaphors. The referent identifier is implemented with a statistical model and a salience model for model comparison. All referents are searched backward from the target anaphors. For identification comparison, we implemented both SVM-based (LIBSVM[6]) and salience-based approaches on different discourse units, namely, single-clause, bi-clauses, and single sentence. Both models are incorporated with feature extraction which yields an optimal set of features and feature weights by running PyGene[7], a genetic algorithm tool in Python. The performance is measured in terms of accuracy which is the ratio of the number of referents tagged correctly at their sentential boundaries by the presented model to the number of referents tagged manually.

Table 4 lists the results of different identification models. It is observed that the salience-based model outperforms the statistical model in terms of higher accuracy. This is because the salience-based model is aimed at selecting the candidate that is the most relevant to the corresponding anaphor while the statistical model picks the relevant one only. On the other hand, clause-level approaches turn out to yield higher accuracy than sentence-level approaches. This is because there might be more irrelevant information acquired from larger discourse-units, like bi-clauses or sentences, thus affecting the selection of the right candidates.

[6] The LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/.

[7] PyGene: http://www.freenet.org.nz/python/pygene

Table 4. Results of different identification models

Model	SVM-based			Salience-based	
Feature extracted	Single clause	Bi-clause	Single sentence	Single clause	Single sentence
Same sentence	1	1	1	0.1	0.1
Same clause	0	0	0	0	0
Sentence distance	0	0	0	0.1	0.1
Clause distance	0	0	0	0	0
Full lexicon agreement	1	0	0	0.1	0.1
Partial lexicon agreement	1	1	1	0.2	0.2
Same verb	0	0	0	0.1	0.1
Clause similarity	1	1	1	0.1	0.1
Same topic word	0	0	0	0	0
Frequent words	1	0	1	0.2	0.2
accuracy	68.46	65.14	53.65	70.54	60.34

Some failures in identification are attributed to the errors in noun phrase chunking. For example, in the text "『天才是九十九分的努力加一分的才氣』這種話銘記在心…" ("Genius is ninety-nine-point hardship plus one-point talent" such saying should be memorized in mind")，the anaphor "這種話" ("such saying") refers the idiom "天才是九十九分的努力加一分的才氣". Such failure may be resolved by taking into account the punctuation marks as one useful feature at referent identification. One the other hand, the present resolution is unable to identify the semantic association between "大清帝國" ("Qing Empire") and "這個時代" (this era) in resolving the anaphor in the text like "在大清帝國這個時代中…" ("In this era of Qing Empire…"). How to improve the presented semantic computation should be concerned as the future work.

5. Conclusion and Future Work

In this paper, we describe definite anaphora in Chinese texts and present empirical methods to resolve the target abstract anaphors which are not widely addressed in previous research projects. In addition, we consider the factors of discourse levels from which the feature extraction is implemented. Without the help of a parser, our experimental results show that clause-level feature extraction is better than the sentence-level extraction in generating useful identification features. Besides, the salience-based approach yields higher accuracy than the SVM-based model whether the best set of extracted features is selected or not.

In the future, we will take into account the Chinese discourse structure and discourse markers in order to improve the boundary identification especially for those referents containing multiple clauses or sentences. Besides, improvement of the semantic computation model should be made so as to enhance the semantic linking between anaphors and their corresponding referents.

References

Bunescu, R. (2003). Associative Anaphora Resolution: A Web-Based Approach. In *Proceedings of EACL 2003 workshop on The Computational Treatment of Aanphora*, 47-52, Budapest.

Byron, D. K. (2002). Resolving Pronominal Reference to Abstract Entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 80-87.

Kong, F., & Zhou, G. (2010). A Tree Kernel-based Unified Framework for Chinese Zero Anaphora Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 882-891.

Navarretta, C. (2004). Resolving Individual and Abstract Anaphora in Texts and Dialogues. In *Proceedings of the 20th International Conference of Computational Linguistics (COLING)*, 233-239, Switzerland.

Pan, S.-C. (2008). *Sentence-based Topic Identification and Its Applications in Chinese Texts*, Master thesis, National Chiao Tung University, Taiwan.

Poesio, M., Ishikawa, T., im Walde, S. S., & Vieira, R. (2002). Acquiring Lexical Knowledge for Anaphora Resolution. In *Proceedings of the 3rd Conference on Language Resource and Evaluation (LREC)*, Las Palmas.

Poesio, M., Mehta, R., Maroudas, A., & Hitzeman, J. (2004). Learning to Resolve Bridging References. In *Proceedings of Annual Conference for Association of Computational Linguistics*, 143-150.

Strube, M., & Müller, C. (2003). A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. In *Proceedings of the 41st Annual Meeting of Association for Computational Linguistics(ACL)*, 168-175.

Wu, D.-S., & Liang, T. (2008). Improving Chinese Pronominal Anaphora Resolution Using Lexical Knowledge and Entropy-based Weight. *Journal of the American Society for Information Science and Technology*, 59(13), 2138-2145.

Wu, D.-S., & Liang, T. (2009). Zero Anaphora Resolution by Case-based Reasoning and Pattern Conceptualization. *Expert Systems with Applications*, 36(4), 7544-7551.

Wu, D.-S., & Liang, T. (2011). Improving Definite Anaphora Resolution by Effective Weight Learning and Web-Based Knowledge Acquisition. *IEICE Transactions on Information and Systems*, E94-D(3), 535-541.

Yang, X., Su, J., & Tan, C. L. (2006). Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the ACL*, 41-48.

Yeh, C.-L., & Chen, Y.-C. (2004). Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing*, 17(1), 41-56.

Yu. C. H. (2000). *A study of Chinese information extraction construction and coreference,* Master thesis, National Taiwan University, Taiwan.

Zhao, S., & Ng, H. T. (2007). Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Joint Conference on empirical methods in natural language processing and computational natural language learning,* 541-550.

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 1-2, March/June 2011, pp. 27-46

Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese

Chao-Lin Liu[*], Guantao Jin[+], Qingfeng Liu[#],

Wei-Yun Chiu[‡], and Yih-Soong Yu[§]

Abstract

We report applications of language technology to analyzing historical documents in the Database for the Study of Modern Chinese Thoughts and Literature (DSMCTL). We studied two historical issues with the reported techniques: the conceptualization of "huaren" (華人, Chinese people) and the attempt to institute constitutional monarchy in the late Qing dynasty. We also discuss research challenges for supporting sophisticated issues using our experience with DSMCTL, the Database of Government Officials of the Republic of China, and the *Dream of the Red Chamber*. Advanced techniques and tools for lexical, syntactic, semantic, and pragmatic processing of language information, along with more thorough data collection, are needed to strengthen the collaboration between historians and computer scientists.

Keywords: Temporal Analysis, Keyword Trends, Collocation, Chinese Historical Documents, Digital Humanities, Natural Language Processing, Chinese Text Analysis.

1. Introduction

Natural language processing (NLP) is a well-known research area in computer science and has been successfully applied to handle and analyze modern textual material in the past decades.

[*] Department of Computer Science, National Chengchi University, Taiwan
E-mail: chaolin@nccu.edu.tw

[+] National Chengchi University, Taiwan
E-mail: gtqf1908@gmail.com

[#] Institute of Chinese Studies, Chinese University of Hong Kong, Hong Kong
E-mail: gtqf1908@gmail.com

[‡] Department of Chinese Literature, National Chengchi University, Taiwan

[§] Department of History, National Chengchi University, Taiwan

Whether we can extend the applications of current NLP techniques to historical Chinese text and in the humanistic context (*e.g.*, Xiang & Unsworth, 2006; Hsiang, 2011a; Hsiang, 2011b; Yu, 2012) is a challenge. Word senses and grammar have changed over time, and people have assigned different meanings to the same symbols, phrases, and word patterns.

We explored the applications of NLP techniques to support the study of historical issues, based on the textual material from three data sources. These include the Database for the Study of Modern Chinese Thoughts and Literature (**DSMCTL**),[1] the Database of Government Officials of the Republic of China (**DGOROC**),[2] and the *Dream of the Red Chamber* (**DRC**).[3] DSMCTL is a very large database that contains more than 120 million Chinese characters about Chinese history between 1830 and 1930. DGOROC includes government announcements starting from 1912 to the present. DRC is a famous Chinese novel that was composed in the Qing dynasty. These data sources offer great chances for researchers to study Chinese history and literature, and, due to the huge amount of content, computing technology is expected to provide meaningful help.

In this paper, we report how we employed NLP techniques to support historical studies. Chinese text did not contain punctuation until modern days, so we had to face not only the well-known Chinese segmentation problem but also the problem of missing sentence boundaries. In recent attempts, we applied the PAT Tree method (Chien, 1999) to extract frequent Chinese strings from the corpora, and we discovered that the distribution over the frequencies of these strings conforms to Zipf's law (Zipf, 1949).

We investigated the issue of how the Qing government attempted to convert from an imperial monarchy to a constitutional monarchy between 1905 and 1911, using the emperor's memorials (奏摺, zou4 zhe2) [4] about the preparation of constitutional monarchy.[5] To this end, we selected the keywords from the frequent strings with human inspection, and we applied

[1] 中國近現代思想及文學史專業數據庫 (zhong1 guo2 jin4 xian4 dai4 si1 xiang3 ji2 wen2 xue2 shi3 zhuan1 ye4 shu4 ju4 ku4): http://dsmctl.nccu.edu.tw/d_about_e.html, a joint research project between the National Chengchi University (Taiwan) and the Chinese University of Hong Kong (Hong Kong), led by Guantao Jin and Qingfeng Liu

[2] 中華民國政府官職資料庫 (zhong1 hua2 min2 guo2 zheng4 fu3 guan1 zhi2 zi1 liao4 ku4): http://gpost.ssic.nccu.edu.tw/. The development of this database was led by Jyi-Shane Liu of the National Chengchi University.

[3] 紅樓夢 (hong1 lou2 meng4): http://en.wikipedia.org/wiki/Dream_of_the_Red_Chamber, a very famous Chinese novel that was composed in the eighteenth century

[4] Most Chinese words are followed by their Hanyu pinyin and tone the first time they appear in this paper.

[5] 清末籌備立憲檔案史料 (qing1 mo4 chou2 bei4 li4 xian4 dang3 an4 shi3 liao4) : http://baike.baidu.com/view/3299810.htm

techniques of information retrieval to support the study.

We also studied the attitude of the Qing government towards the Chinese workers who worked in other countries between 1875 and 1911. We analyzed the co-occurrences, *i.e.*, collocations, of the keywords over the years of interest, using the documents recorded in the diplomatic documents of the late Qing dynasty.[6]

Detailed observations and discussions of this historical research are reported in two other papers (Jin *et al.*, 2011; Jin *et al.*, 2012) that will be presented in the Third Conference of Digital Archives and Digital Humanities.

While we have applied NLP techniques to support historical studies, we have also experienced some challenging problems at the lexical, syntactic, semantic, and pragmatic levels. For instance, what are the most appropriate computational functions that support a certain research need? Are the current databases good enough? We elaborate on these challenges based on our experience with the three data sources, *i.e.*, DSMCTL, DGOROC, and DRC.

No one may expect that NLP techniques will replace the major role of historians in historical studies, but the techniques should be able to work with historians to make their studies more efficient and more effective. Empirical experience reported in this paper and the literature have demonstrated the potential of NLP techniques. With the help of computing technology, historians can delegate some search work and basic analysis to computers and spend more time on higher-level philosophical issues than before.

2. Zipf's Law Applicability

The Database for the Study of Modern Chinese Thoughts and Literature contains six genres of text material that were published between 1830 and 1930. Except for the first category, most of them were collected from the late Qing dynasty: modern periodicals, personal publications of the literati, diplomatic documents, newspapers, official documents, and translated works by western commissioners. Currently, the database contains more than 120 million simplified Chinese characters.[7]

[6] 清季外交史料 (qing1 ji4 wai4 jiao1 shi3 liao4): http://zh.wikisource.org/zh-hant/清季外交史料選輯

[7] DSMCTL was first built in a project led by Guantao Jin and Qingfeng Liu while they were with the Chinese University of Hong Kong. Due to budget constraints, the historical documents were sent to China, where the simplified Chinese was used, to be scanned and entered into computers. Hence, the earliest version of DSMCTL was in simplified Chinese. A traditional Chinese version of DSMCTL is still under development.

Table 1. Statistics for five collections

Collection	Number of Different Pseudowords	Total Number of Characters	Number of Different Characters	Number of Documents
Constitution	3288	713131	4097	399
Diplomacy	29315	2875032	5225	5758
Min_Bow	7784	1450623	6230	325
Nations	2649	679410	4916	160
New_People	33378	5259590	6647	1524

For modern Chinese information processing with NLP techniques, researchers rely on good machine readable lexicons and good methods to segment Chinese strings into Chinese words. Both of these infrastructural facilities are missing for the processing of non-modern Chinese text. Hence, we bootstrapped our work by computing frequent Chinese strings with the PAT Tree technique in the documents, and we asked historians to select relevant words from the frequent strings.

Table 1 shows the statistics about five collections in the DSMCTL database: *Constitution* (清末立憲檔案), *Diplomacy* (清季外交史料), *Min_Bow* (民報),[8] *Nations* (海國圖志),[9] and *New_People* (新民叢報).[10] They contain about 11 million characters, about one tenth of the whole DSMCTL database. We refer to strings that occurred more than 10 times[11] in a collection as ***pseudowords***. Many of these pseudowords have specific meanings, but not all of them do.

We ranked the pseudowords based on their frequencies, *i.e*., the most and the second most frequent pseudowords were ranked first and second, respectively. Then, we computed the logarithmic values of the ranks and frequencies, resulting in the curves in Figure 1. The curves in Figure 1 indicate that the pseudowords in the Chinese historical documents, like documents written in modern English and Chinese languages (Ha *et al*., 2003; Xiao, 2008), conform to Zipf's law quite well (Zipf, 1949).

[8] 民報 (min2 bao4): http://zh.wikipedia.org/wiki/民報

[9] 海國圖志 (hai3 guo2 tu2 zhi4): http://zh.wikipedia.org/wiki/海國圖志

[10] 新民叢報 (sin1 min2 cong2 bao4): http://zh.wikipedia.org/wiki/新民叢報

[11] The selection of 10 as the threshold was by the historians. The choice was heuristic but arbitrary.

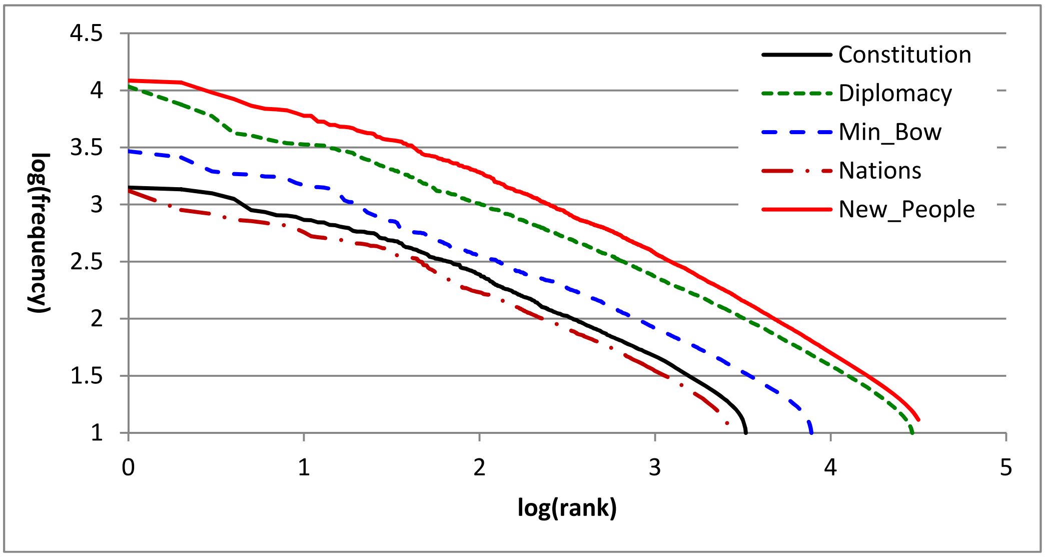


Figure 1. Pseudowords in the Chinese historical collections abide by Zipf's law

Let r and f denote the rank and frequency of a word in a collection of text, respectively, Zipf's law predicts that the product of r and f is a constant, c, as shown in Equation (1).

$$f = \frac{c}{r} \tag{1}$$

Hence, we will observe curves that are almost straight lines after we take the logarithm (usually abbreviated as "log") on both sides of Equation (1) to become Equation (2). In Figure 1, the log values of pseudoword frequencies are on the vertical axis, and the log values of the pseudoword ranks are on the horizontal axis.

$$\log(f) = \log(c) - \log{(r)} \tag{2}$$

Let N denote the total number of characters in a collection. We divided the word frequencies by the sizes of individual collections. In Figure 2, the vertical axis shows the log values of the pseudoword frequencies divided by N, namely, $\log\left(\frac{f}{N}\right)$. The curves for the distributions of the pseudowords almost overlap, suggesting that Zipf's law applied to the five collections quite uniformly, after we considered the influences of the sizes of collections.

The decision to divide term frequency, *f*, by the corpus size, *N*, was arbitrary, but it was very interesting to find that curves in Figure 2 almost overlap as a result. Evidently, sizes of corpora affected the shapes and positions of the Zipfian curves. Xiao (2009) attempted to study the influences of corpus size over the Zipfian curves. In one of the reported studies, Xiao sampled five small datasets of almost the same size from the General Contemporary Chinese Corpus, which contained approximately one billion Chinese characters. Zipfian

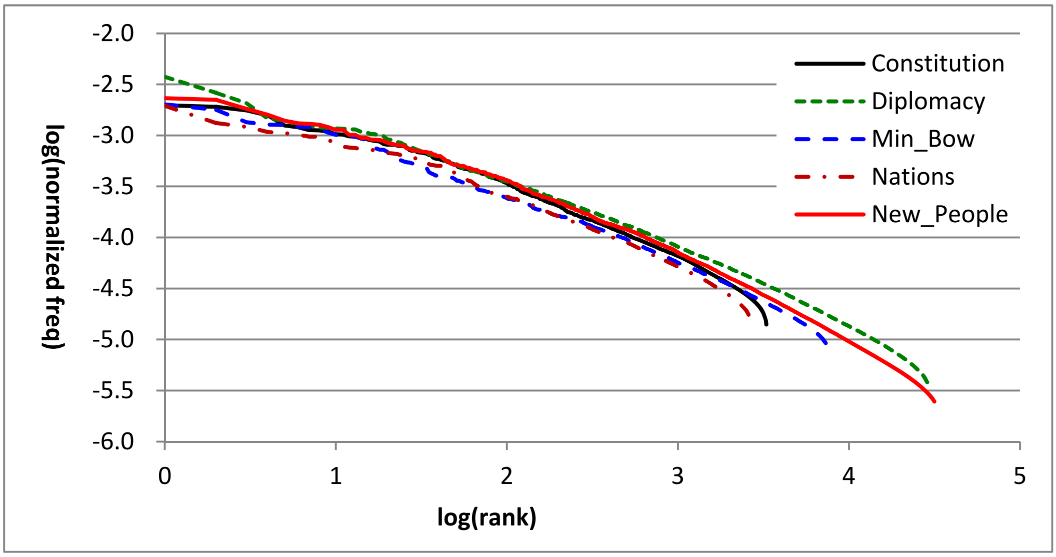


Figure 2. Reducing the influences of sizes of individual collections

curves drawn for these datasets overlapped almost perfectly.

3. Chronicle Trends of Multiple Keywords

We examined the pseudowords and selected those that are potentially relevant to historical issues as ***keywords***. We computed the annual and total frequencies of each of these keywords and computed the total number of keywords in each year.

The "Total" curve serves as the basis for the analysis of importance of keywords. Let t_{1905}, t_{1906}, t_{1907}, t_{1908}, t_{1909}, t_{1910}, and t_{1911} denote the total number of keywords appearing in 1905, 1906, 1907, 1908, 1909, 1910, and 1911, respectively. We could compute the total number of keywords in *Constitution*, *T*, using the following equation.

$$T = t_{1905} + t_{1906} + t_{1907} + t_{1908} + t_{1909} + t_{1910} + t_{1911} \quad (3)$$

Using the years on the horizontal axis and the ***annual percentage***, $\frac{t_i}{T}$, on the vertical axis, we analyzed the keywords in *Constitution* (*cf*. Table 1) to obtain the "Total" curve in Figure 3.

Analogously, let *K* denote the total number of a particular keyword, *e.g*., "官制," (guan1 zhi4, bureaucracy) that appeared in *Constitution* and k_n denote the number of instances the keyword appeared in a year *n*. We can draw a curve of annual percentage for a keyword. Figure 3 shows the curves of annual percentages of all words (Total) and six keywords over the years between 1905 and 1911 in *Constitution* (*cf*. Table 1).

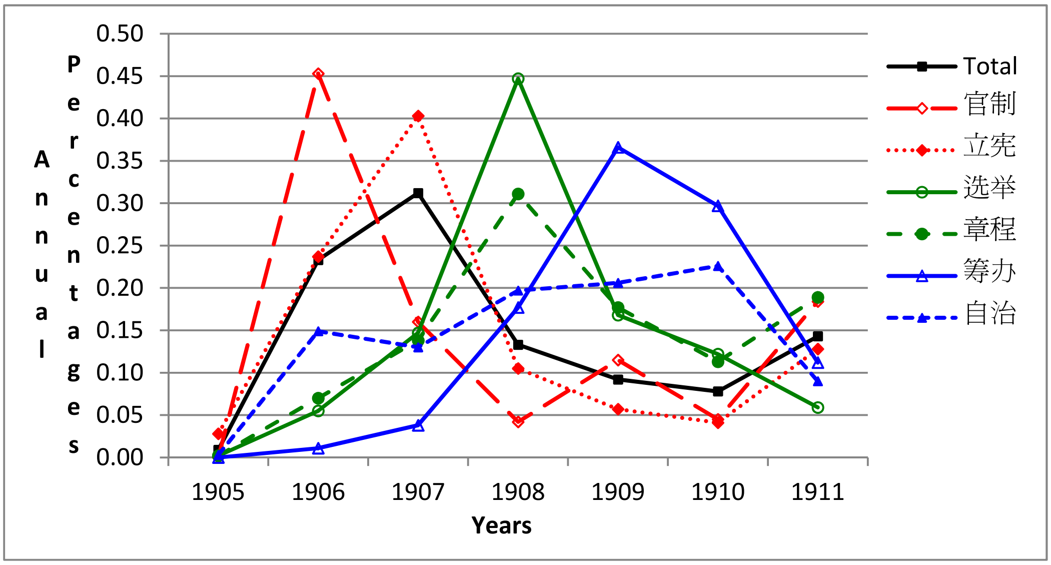


Figure 3. Some keywords appeared more frequently in particularly years

When the keywords appeared more frequently in a year, a historical event typically coincided with the increase in frequency (Jin *et al*., 2011). We considered the keyword to be special in the year *n*, if $\frac{k_n}{K} \geq \lambda \frac{t_n}{T}$. In Jin *et al*. (2011), we chose $\lambda = 1.1$ arbitrarily, but the selection of λ can be adjusted as needed in a computer-assisted document analysis environment.

For instance, in 1906, $\frac{k_{1906}}{K}$ for "官制" was about 0.45, and $\frac{t_{1906}}{T}$ was less than 0.25. In 1907, $\frac{k_{1907}}{K}$ for "立宪"[12] was about 0.40, and $\frac{t_{1907}}{T}$ was less than 0.33. Both "官制" and "立宪" qualified as special. In 1906 and 1907, the Qing government began to consider constitutional monarchy seriously, so government officials intensively discussed the issues of "bureaucracy" ("官制") and "constitutionality" ("立宪") for running the new form of government. Hence, keywords like "官制" and "立宪" appeared in the emperor's memorials more often than in other years.

In 1908, $\frac{k_{1908}}{K}$ for "选举"[13] was about 0.45, and $\frac{t_{1908}}{T}$ was less than 0.15. In fact, in 1908, the keywords about election ("选举" and "章程") were used more frequently in the emperor's memorials.

After years of discussion on the fundamental issues of a constitutional monarchy, the Qing government appeared to be prepared for the new form of government and was taking steps for its realization. In 1909 and 1910, words relating to self-governance ("筹办" and "自

[12] 立宪: li4 xian4, constitutionality

[13] 选举: xuan3 ju3, election; 章程: zhang1 cheng2, rules

治")[14] became relatively more important.

The temporal relationship between these six keywords' emerging importance further suggested the progression toward the establishment of a constitutional monarchy before the overthrow of the emperor in late 1911. Namely, the focus of discussion shifted from planning and preparation to realization and action.

Our approach is more appropriate for historical studies than the Google Trends[15] approach, although the difference is subtle and may appear minuscule. The analysis of occurrences of an individual keyword, like in Google Trends, is useful for studying the changing importance of a keyword over a period of time. Evaluating the chronicle change of importance of a keyword is certainly important, but we further compare the chronicle changes of multiple keywords, which allows us to visualize the trends more directly.

4. Temporal Analysis of Important Collocations

A ***collocation*** is formed by two keywords that appeared "close" to each other in a statement. A collocation carries more specific semantic information than an individual keyword. The occurrence of the keyword "Chinese labor" ("华工," hua2 gong1) could have referred to anything about Chinese labor, *e.g*., limiting ("限制," xian4 zhi4) or protecting ("保护," bao3 hu4) the Chinese labor, while a collocation "protect the Chinese labor" ("保护" and "华工") provides more specific meaning than the individual keywords.

Nevertheless, given that there were neither word boundaries nor sentence boundaries in pre-modern Chinese documents. We chose to define "close" based only on the "distance" between two keywords.

A keyword was considered to be collocated with another if the keywords were less than 30 characters apart. Our computer programs were flexible in setting the window size for "closeness". We defined the *collocation window* as the span of characters around a keyword that are considered "close". We ran experiments where the sizes of the collocation windows were set to 10, 20, and 30 characters. A collocation window of 30 characters will consider 30 characters on the left and on the right side of a keyword. The historians observed the computed collocations and preferred the size of 30.

We analyzed the statistics of collocations in the documents about the concept of "Chinese People" ("華人," hua2 ren2) in *Diplomacy* (*cf*. Table 1). We identified the keywords with the procedure that we applied to find individual keywords in *Constitution* that we explained in the previous section. Historians then chose the keywords of interest and we ran

[14] 筹备: chou2 bei4, preparation; 自治: zi4 zhi4, self-governance

[15] http://www.google.com/trends/

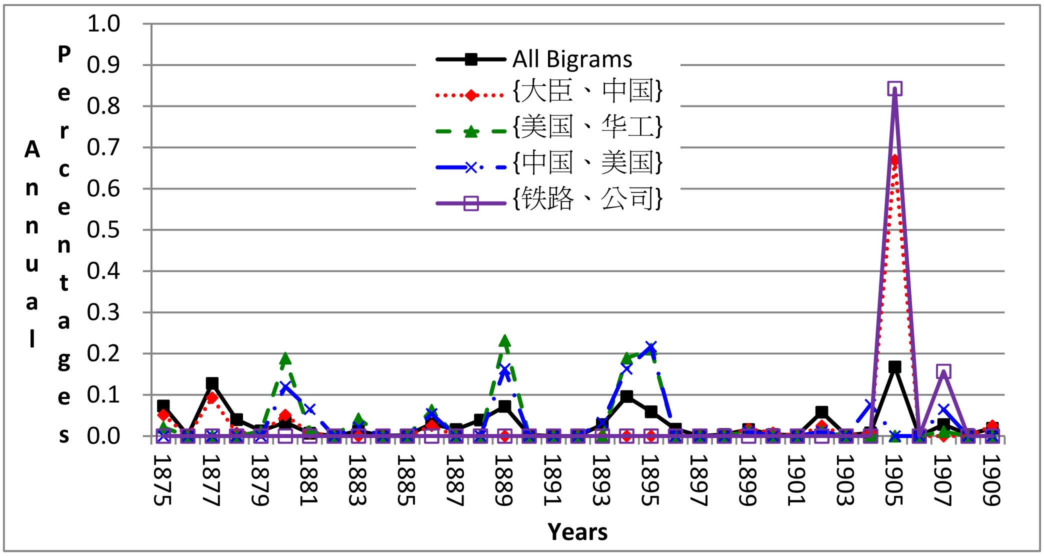


Figure 4. Importance of keyword collocations varied over the years

the computer programs to do the temporal analysis of the important collocations. This procedure is similar to the procedure that we used to obtain Figure 3; the only difference was whether the target of analysis was keywords or collocations.

Curves in Figure 4 show that the annual percentages of four collocations varied over the years between 1875 and 1909. Significantly large annual percentages again coincided with historical events of the given years. In 1894, the United States of America (USA) ("美国" (mei3 guo2) in the chart) and the Qing government signed a treaty to limit Chinese laborers ("华工") entering the USA.[16] In 1905, Chinese societies started to boycott American's products, mostly because the USA would extend the treaty signed in 1894.[17] Initially, the Qing dynasty was trying to protect only the Chinese laborers. Later, the protection was extended to *Chinese merchants* then extended to *Chinese people* (Jin *et al.*, 2012).

5. Ranking Individual Documents: An Application of Information Retrieval

As the statistics in Table 1 show, there can be thousands of documents containing millions of characters in a particular collection. Finding the most relevant documents or essays to read was not easy in the past. With our ability to identify the important keywords and collocations, we could rank the documents based on how documents included the important keywords and

[16] 《中美華工條約》、《限禁來美華工保護寓美華人條約》：
http://dict.zwbk.org/zh-tw/Word_Show/64744.aspx

[17] http://zh.wikipedia.org/wiki/抵制美貨運動；《*籌拒美國華工禁約公啓*》

Table 2. Three highly ranked emperor's memorials written in 1906 (in Constitution in Table 1) and their weights (i.e., relative importance), authors, and titles

1906		
Weights	**Authors**	**Document Title**
420	戴鸿慈	出使各国考察政治大臣戴鸿慈等奏请改定全国官制以为立宪预备折
312	杨晟	出使德国大臣杨晟条陈官制大纲折
122	殷济	内阁校签中书殷济为豫备立宪条陈筹经费建海军等二十四条呈

collocations. Table 2 shows a part of the table where we ranked the documents in *Constitution* (*cf*. Table 1). The weights in Table 2 were calculated based on the number of keywords that were used in a document. Larger weights imply that more keywords were used in the document, so the document might be more relevant to the research topic for which the researcher selected the keywords. The ranking function and other techniques for information retrieval and extraction could provide useful information for historians to study specific issues (Jin *et al.*, 2011; Jin *et al.*, 2012).

6. Discussion

In this section, we discuss some technical problems related to using computing techniques to support historical studies in Chinese.

6.1 Lexical Ambiguity, Pragmatics, and Term Identity

We have illustrated three possible applications of textual analysis for historical studies in previous sections. The applications were based on the frequencies of keywords in the collections. In NLP, we can refer to the frequencies of keywords as ***term frequencies***. In addition, we relied on the "time stamps" of the documents, where the "time stamps" are the recorded times of the documents. Based on our dependence on the terms frequencies and time stamps, we obtained and presented the figures that we discussed in Sections 3 and 4.

In these examples, we presume that the frequencies reflect the importance of the concepts that are represented by the terms and collocations, and the results of our work are quite convincing. Nevertheless, we have to watch for the problems of lexical ambiguity and pragmatics that are hidden under the term frequencies.

For instance, frequently cited events of the past may induce confusion about the significance of term frequencies. Tu *et al*. (2011) discovered that, although "張公藝" (zhang1 gong1 yi4) was a Chinese name that appeared frequently in collections in the Taiwan History Digital Library (THDL), "張公藝" referred to a person who actually lived in the Tang dynasty

(618AD-1907AD),[18] which is well before the time period of the documents in THDL. The documents in THDL referred to“張公藝” because of a story that was well-known in the Qing dynasty (1644AD-1912AD). That the term frequency of “張公藝” is high in THDL does not imply that “張公藝” himself was an important person in Taiwan in the Qing dynasty.

Lexical ambiguity may make the term frequencies less reliable. Yu (2012) accentuates this issue with “民主” (min2 zhu3). In modern Chinese, “民主” is the word for “democracy”. Nevertheless, it could represent the emperor (民之主, min2 zhi1 zhu3), the American president, and the Republic (in 民主國, min2 zhu3 guo2) in non-modern Chinese text.

The first author of this paper examined the DGOROC, and found that “陳建中” was a very common name in the database. Hence, finding the actual identities of names is an important issue, in addition to computing the term frequencies. Distinguishing persons of the same name in modern databases requires extraordinary sources of private information.

Although differentiating persons with the same names is not easy, identifying names in Chinese text is not an easy task for the research of Named Entity Recognition (often referred to as NER (Wu *et al.*, 2006)) in the first place. For instance, it may not be easy to extract names from Chinese text like “中央高層正醞釀安排令計畫接任中組部長”[19] (zhong1 yang1 gao1 ceng2 zheng4 yun4 niang4 an1 pai2 ling4 ji4 hua4 jie1 ren4 zhong1 zu3 bu4 zhang3) if we do not know “令計劃” (ling4 ji4 hua4) is a name.[20]

6.2 Word Segmentation and Sentence Division

In Section 6.1, we discuss the interpretation of a given term. In Chinese, however, we also have to define the concept of “term”. If we cannot define terms precisely, then we have no grounds for defining collocations. It is known that Chinese words are not separated by spaces like in alphabetic languages, and the task of separating Chinese words in Chinese text is generally called word segmentation (*e.g.*, Ma & Chen, 2005; Jiang *et al.*, 2006; Tseng *et al.*, 2005). It is less known, however, that pre-modern Chinese text does not have punctuation, and readers also have to figure out the divisions of sentences (Huang, 2008).

Clearly, if we could not divide sentences and segment words correctly, we would not be able to acquire correct term frequencies. This may happen when we process text like “五行者金主義木主仁水主智火主禮土主信” (wu3 xing2 zhe3 jin1 zhu3 yi4 mu4 zhu3 ren2 shui3 zhu3 zhi4 huo3 zhu3 li3 tu3 zhu3 xin4). We would have to add punctuation to divide this

[18] http://en.wikipedia.org/wiki/Tang_Dynasty

[19] Source: http://www.cbfcn.com/news_detail.aspx?strnew=1154

[20] In fact, “令路線” (ling4 lu4 xian4), “令政策” (ling4 zheng4 ce4), “令完成” (ling4 wan2 cheng2), and “令方針” (ling4 fang1 jhen1) are also Chinese names (although they are quite unusual) : http://zh.wikipedia.org/wiki/令計劃

string: “五行者，金主義，木主仁，水主智，火主禮，土主信”. After this step, we know that “智火” is not a term in the original string, although “智火” can be a meaningful term in modern Chinese.[21] Given the divided string, we still have to face the word segmentation problem. In this example, each character in “金主義” represents a specific meaning. We cannot interpret “主義” in “金主義” as we would interpret “主義” (-ism) in “帝國主義” (di4 guo2 zhu3 yi4; imperialism) or “資本主義” (zi1 ben3 zhu3 yi4; capitalism) in modern Chinese. Similarly, we have to know that “金主” is not a term in the original string, although “金主” (a wealthy person) is a meaningful term in modern Chinese.

An actual problem took place when we used the DSMCTL to investigate whether energy conservation was a concern in the Qing dynasty (Chou *et al.*, 2012). Without a Chinese segmenter for pre-modern Chinese, we found many occurrences of “能源” (neng2 yuan2) in the database, but, most of the time, “能源” was just a sub-string of “不能源源而來” (bu4 neng2 yuan2 yuan2 er2 lai2) when people talked about something that could not come indefinitely.[22]

6.3 Trends: Informative or Deceptive

In Sections 3 and 4, we briefly introduced applications of temporal trends of keywords (Figure 3) and trends of collocations (Figure 4) that were more thoroughly discussed in Jin *et al.* (2011) and Jin *et al.* (2012), respectively. Researchers in other fields also have found impressive applications of trends of keywords (*e.g.*, Caneior & Mylonakis, 2009). Despite these successful applications, caution is in need to interpret the observed trends.

Figure 5 shows temporal trends for the names of three main characters in a famous novel *Dream of the Red Chamber* (DRC). The horizontal axis shows the chapters of the DRC. The vertical axis shows the frequency of the keywords (persons’ names in this chart). The highs of the curves shows the times of being mentioned of a person in a particular chapter, so are indicative of the relatively importance of the persons. We discuss three main persons in DRC, “寶玉” (Bao3 Yu4), “黛玉” (Dai4 Yu4), and “寶釵” (Bao3 Tsai1), in the following.

[21] “智火” happens to be the name of a Chinese company: http://www.zhihuo.asia/.

[22] The Chinese segmentation service at Academia Sinica (http://ckipsvr.iis.sinica.edu.tw/) would return “不能,” “源源,” “而,” and “來” for “不能源源而來”. The online version of the Stanford parser (http://nlp.stanford.edu:8080/parser/index.jsp) would return “不,” “能,” and “源源而來”.

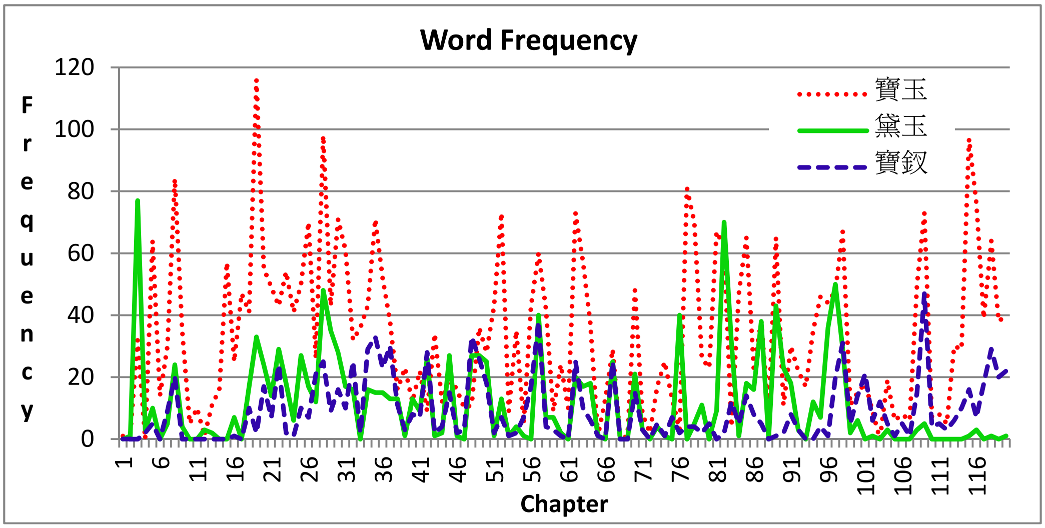


Figure 5. Frequencies of three main names in Dream of the Red Chamber

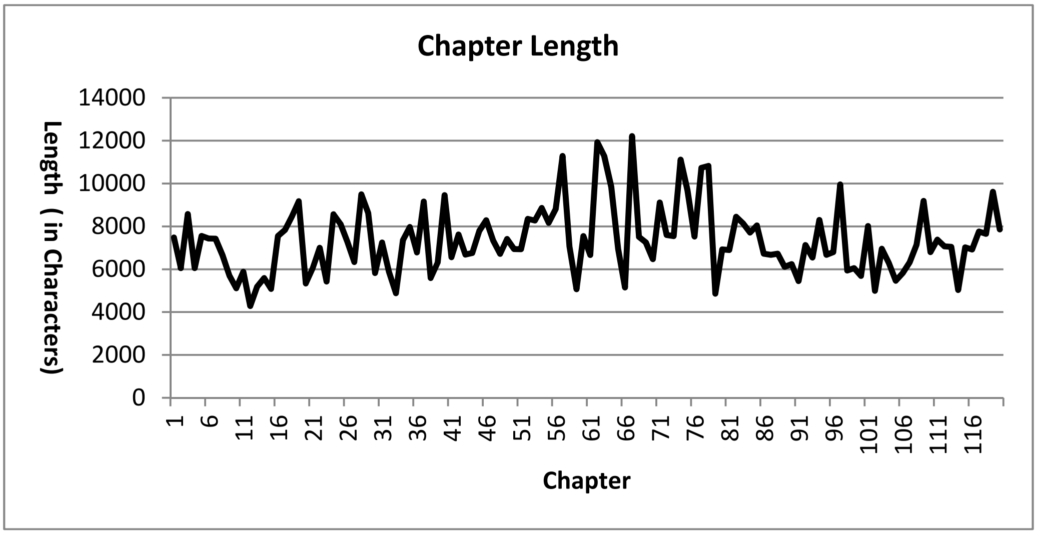


Figure 6. Lengths of chapters in Dream of the Red Chamber

Do the ups and downs of a particular curve show the changes of importance of a person? Intuitively, the answer may be yes. If the name of a person was mentioned more frequently, that particular person should be more involved in a chapter. This interpretation, however, is not flawless – a person being mentioned more times might be the result of a longer chapter. Being mentioned more times in a longer chapter might not be solid proof of the importance of the mentioned person.

Figure 6 shows the numbers of characters in each chapter in DRC. Evidently, some chapters are longer and some are shorter.

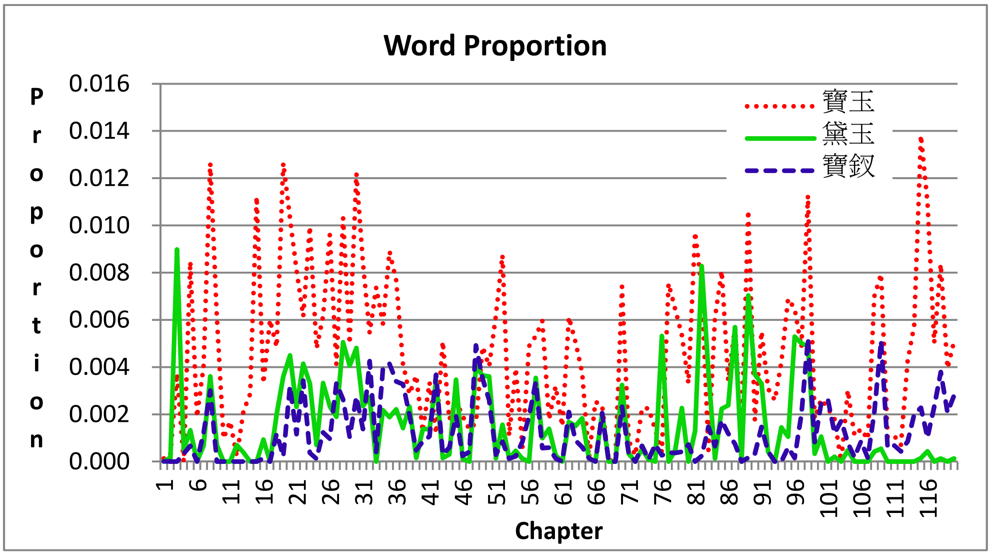


Figure 7. Proportions of three major names in individual chapters in Dream of the Red Chamber

Let f_t and l_t , respectively, denote the frequency of a keyword and the length of a chapter t in DRC. We divide f_t by l_t, for t=1, …, 120, for the three names in Figure 5. Figure 7 shows the resulting curves for the three persons.

We can observe some important changes in the curves. Take the curve for Bao-Yu ("寶玉") for example. Bao-Yu was mentioned 84, 116, and 98 times in Chapters 8, 19, and 28, respectively. These three instances formed the first three peaks above 80 in the curve for Bao-Yu in Figure 5. The frequencies may have suggested that Bao-Yu were more important in Chapter 19 than in Chapters 8 and 28. After we divided these frequencies by the chapter lengths, we observed that the proportions of Bao-Yu being mentioned in these chapters were almost the same in Figure 7. Hence, the trends illustrated in Figures 6 and 7 provide hints for different conclusions.

Consider another example. Assume that we want to know who among the three persons liked to "smile and say" most in DRC. Curves in Figure 8 show the frequencies of "寶玉笑道," "黛玉笑道," and "寶釵笑道" in DRC, where "笑道" (xiao4 dao4) is a way to say "smile and say". The curves suggest that, before Chapter 40, Bao-Yu was the person who liked to "smile and say" most.

Nevertheless, one may contend that the absolute frequency may not be a perfect indicator for how likely a person was to "smile and say". If a person was mentioned less frequently, then s/he would not be able to "smile and say" as frequently as another who was mentioned more frequently.

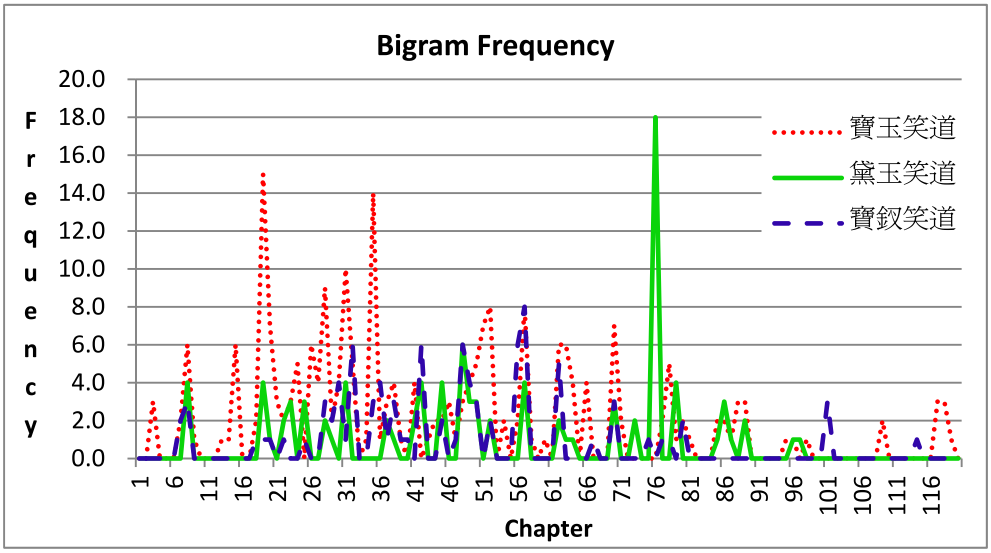


Figure 8. Word frequencies indicate that Bao-Yu laughed most in early chapters

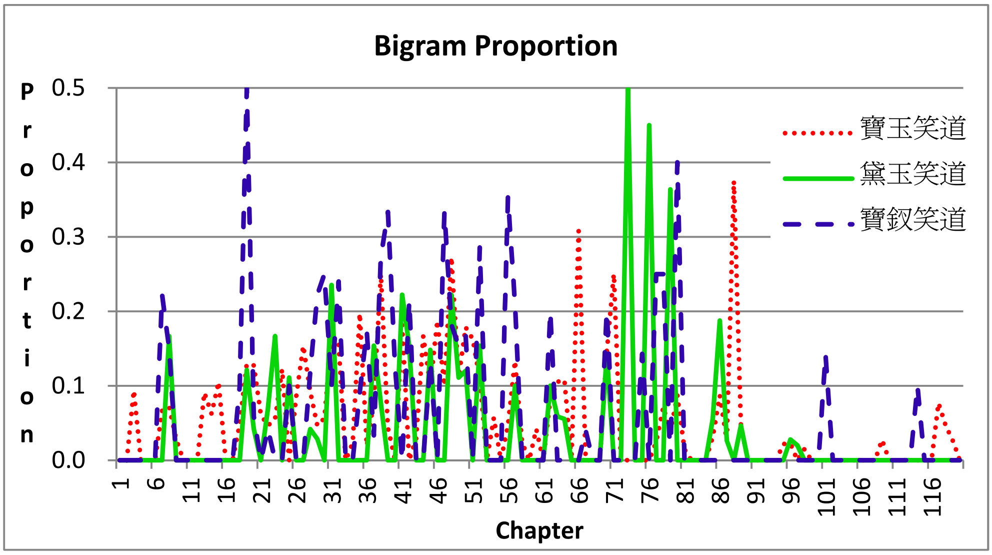


Figure 9. Bigram proportions show that Bao-Chai laughed most in early chapters

Let s_t and m_t, respectively, denote the frequency a person "smiled and said" and a person was mentioned in a chapter t. For the three persons in our current discussion, m_t was their individual term frequency f_t that we showed in Figure 5. We divided s_t by m_t for each person and came up with Figure 9.

Quite interestingly, the curve for Bao-Yu does not dominate the others anymore. Instead, Bao-Tsai ("寶釵") smiled and said something once every two appearances in Chapter 19, as did Dai-Yu (“黛玉”) in Chapter 73. In fact, never did Bao-Yu smile and say as often as 50% of the time he appeared in any chapter. The highest proportion of Bao-Yu’s “smile and say” took place in Chapter 88, where the proportion still fell short of 40%.

One researcher may be interested in the times a person smiled and said something, and another might be interested in the proportion a person smiled and said something when the person was mentioned in DRC. Take Bao-Yu for example. In the former case (Figure 8), the term frequency of Bao-Yu is the focus. In contrast, in the latter case (Figure 9), the conditional probability $\Pr(\text{Bao-Yu-Xiao-Dao} \mid \text{Bao-Yu})$ is of interest, and we have to compute the probability based on the observed frequencies. Different trends and analyses should be used for different purposes, and this is up to the researchers’ discretion. When designing tools for assisting historical studies, appropriate functionalities should be considered and explained to their users as clear as possible.

6.4 Transliteration and Translation

In addition to the ability to process normal pre-modern Chinese text, one may need to handle transliterated and translated words. Chinese people encountered western culture more directly and more frequently starting from late 1500s. Transliteration and translation are important ways for people to use Chinese words to convey and understand western concepts and entities.

To study the interactions between the Chinese and western cultures in pre-modern times, getting to know the Chinese transliterations and translations is an important step. For instance, “president” was transliterated into “伯理璽” (bo2 li3 si2), “伯理喜頓” (bo2 li3 si2 dun4), “伯理璽天德” (bo2 li3 si2 tian1 de2), and “伯力錫天德” (bo2 li3 si2 tian1 de2). Some Chinese characters were selected based on the pronunciations of “president”, and some were selected to show respect to the position of “president.” “Pacific Ocean” was translated into “大海” (da4 hai3), “大東洋” (da4 dong1 yang1), and “太洋海” (tai4 yang1 hai3), and transliterated into “卑西溢湖” (bei4 si1 yi1 hu2) and “比西非益海” (bi3 si1 fei1 yi4 hai3). The translations show people’s knowledge about the size and position of the Pacific Ocean. “Politics” was transliterated into “薄利第加” (bo2 li4 di4 jia1) and “波立特” (po2 li4 te4); both were transliterations.

Historians may spend their lifetimes identifying the translated and transliterated terms in historical documents. If one could provide researchers the Chinese terms for the western concepts and entities, the researchers would be able to investigate and understand how Chinese faced the West hundreds years ago.

Therefore, we imagine that it would be useful if computing technologies could help historical researchers identify transliterated and translated terms in historical documents. It may be not easy to use human experts to annotate a database that has 120 million characters, such as DSMCTL.

6.5 Advanced NLP Techniques: Trend Analysis

An anonymous reviewer of the manuscript of this paper pointed out that applications of advanced NLP techniques will strengthen the values of the collected statistics. For instance, one may classify the keywords into types and conduct temporal analysis on keywords of the same type. This may give us a trend analysis similar to the analysis of emotion trend reported in Yang *et al*. (2007). It is also possible to treat the network of keywords as a social network. The nodes can be verbs, nouns, and names, and links show strengths of associativity. Networks like this may shed light on historical events that were difficult to see by simply studying the historical documents.

6.6 Time Stamps, Missing Data, and Fundamental Changes in DGOROC

The DGOROC database[23] provides information about the appointments of government officials of the Republic of China in Taiwan. This database was constructed and verified with human labor. Information was copied from hard copies of official documents, entered into text files, and was verified for quality assurance. It contains more than 850 thousand records dated from 1912, and is useful for studying modern history and relevant applications about Taiwan and China (*e.g.*, Liu & Lai, 2011).

Since the data came from a real and changing government, there can be barriers that were difficult to overcome simply by computing technology. For instance, the current government in Taiwan was not in a really stable condition until she moved to Taiwan in 1949. Hence, the database is relatively more complete for records after circa 1949.

It should not be surprising that a government tries and evolves to serve the nation in the fast changing world. For instance, there was no “Ministry of Education” before 1928, although there must have been some government agents to handle national education policies before 1928. Hence, a researcher will have to know the names of the agents that were responsible for education to study the national education policies circa 1928. In this case, a simple keyword search service may not help very much.

Although the data collected after 1949 was more complete in DGOROC, the government may change the rules for whether or not to announce some types of assignments. For instance, there are departments in the Ministry of Education, and the department heads may change their appointments from a department to another, but this type of switch is not publicly

[23] The first author gained experience with DGOROC while serving as the project leader for maintaining DGOROC between February and August 2011. The comments about DGOROC in this section are of the first author.

announced in recent years.

The appointments of lower ranks of government officials may not be announced at real time. The announcement of such appointments may be delayed so that a larger group of appointments would be announced at the same time. If the time stamps of events for a certain study matter, then this kind of delay may be troublesome.

Despite these remaining challenges in DGOROC, we consider this database unique and important. By incorporating information available from other database maintenance agents of the central government and from national libraries, the database will offer researchers a great information source for studying modern history of Taiwan.

7. Concluding Remarks

We delineate our experience in using three sources of historical documents in Chinese: the database of Chinese historical documents that contain more than 120 million simplified Chinese characters, the Database of Government Officials of the Republic of China, and the Dream of the Red Chamber. Techniques for natural language processing were employed to analyze the contents of the documents to facilitate the studies of historical events. The exploration showed that NLP techniques are instrumental for the studies of non-modern Chinese historical documents. Our experience also suggested that advanced NLP techniques and more complete data collection are necessary for supporting research work in more precise ways.

Acknowledgments

The work was supported in part by the funding from the National Science Council in Taiwan under the contracts NSC-99-2221-E-004-007 and NSC-100-2221-E-004-014. The Database for the Study of Modern Chinese Thought and Literature (1830-1930) was funded by the sub-project 100H51 at the National Chengchi University, Aim for the Top Universities Project, Ministry of Education, Taiwan. The writing of this paper benefits tremendously from reviewers' comments on previous versions of this paper. Section 6.5, in particular, was prompted by the review comments of this journal version. The English editor, Mr. Sun, of this journal helped us a lot by providing suggestions for better presentation in addition to correcting language problems.

References

Carneior, H. A. & Mylonakis, E. (2009). Google trends: A Web-based tool for realtime surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10), 1557-1564.

Chien, L.-F.. (1999). PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. *Information Processing and Management*, 35(4), 501-521.

Chou, L.-F., Liu, C.-L., & Yu, Y.-S. (2012). 從歷史文獻關鍵字來看--清末民初百年間中國能源觀念演進初探, *Energy Monthly*, to appear, Taiwan Institute of Economic Research. (in Chinese)

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2003). Extension of Zipf's law to word and character n-grams for English and Chinese. *International Journal of Computational Linguistics and Chinese Language Processing*, 8(1), 77-102.

Hsiang, J. (Ed.) (2011a). *From Preservation to Knowledge Creation: The Way to Digital Humanities* (從保存到創造：開啓數位人文研究), *Series on Digital Humanities*, volume 1, National Taiwan University Press. (in Chinese)

Hsiang, J., (Ed.) (2011b). *New Eyes for Discovery: Foundations and Imaginations of Digital Humanities* (數位人文研究的新視野：基礎與想像), *Series on Digital Humanities,* volume 2, National Taiwan University Press. (in Chinese)

Huang, H.-H. (2008). *Classical Chinese Sentence Division by Sequence Labeling Approaches* (以序列標記方法解決古漢語斷句問題), Master's Thesis, National Chiao Tung University, Hsinchu, Taiwan. (in Chinese)

Jiang, W., Guan, Y., & Wang, X.-L. (2006). A pragmatic Chinese word segmentation approach based on mixing models, *International Journal of Computational Linguistics and Chinese Language Processing*, 11(4), 393-416.

Jin, G., Chiu, W.-Y., & Liu, C.-L. (2012). Frequency analysis and application of 'co-occurrence' phrases: the origin of the concept 'Hua-Ren' as an example (「共現」詞頻分析及其運用－以「華人」觀念起源爲例), to appear, in *Series on Digital Humanities*, volume 3, J. Hsiang (Ed.), National Taiwan University Press. (in Chinese)

Jin, G., Yu, T.-S., & Liu, C.-L. (2011). Applications of digital methods to study social movements - Using the preparation of constitutional monarchy in the late Qing dynasty as an example (社會行動的數位人文研究：以清末預備立憲爲例), presented in the Third Conference of Digital Archives and Digital Humanities. (in Chinese)

Liu, J.-S., & Lai, L.-P. (2011). Rank promotion prediction on Taiwanese government officials, in (Hsiang, 2011a), 113-130. (in Chinese)

Ma, W.-Y. & Chen, K.-J. (2005). Design of CKIP Chinese word segmentation system. *Chinese and Oriental Languages Information Processing Society*, 14(3), 235-249.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005). A conditional random field word segmenter. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 32-39.

Tu, F.-E., Tu, H.-C., Chen, S.-P., Ho, H.-I., & Hsiang, J. (2011). Information technology and open problems in the Taiwan history digital library (THDL), in (Hsiang, 2011b), 21-44. (in Chinese)

Wu, C.-W., Jan, S.-Y., Tsai, R. T-H., & Hsu, W.-L. (2006). On using ensemble methods for Chinese named entity recognition. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 142-145.

Yang, C.-H., Kuo, H.-A., & Chen, H.-H. (2007). Emotion trend analysis using blog Corpora, *Proceedings of the Nineteenth Conference on Computational Linguistics and Speech Processing*, [http://aclweb.org/anthology-new/O/O07/O07-1015.pdf]. (in Chinese)

Yu, H. (鱼宏亮)。(2012)。范式的转变：重建观念史图像中的历史真实 - 新方法与中国近代观念史研究，*東亞觀念史集刊*，第三期，將出刊。政治大學出版社。(in Chinese)

Xiang, X., & Unsworth, J. (2006). Connecting text mining and natural language processing in a humanistic context, *Proceedings of the International Conference on Digital Humanities* 2006.

Xiao, H.. (2008). On the applicability of Zipf’s law in Chinese word frequency distribution, *Journal of Chinese Language and Computing*, 18(1), 33-46. (This journal has been renamed as *International Journal of Asian Language Processing* in 2009.)

Zipf, G. K.. (1949). *Human Behavior and the Principle of Least Effort*. Addison Wesley.

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 1-2, March/June 2011, pp. 47-68

以籠統查詢評估查詢擴展方法與線上搜尋引擎之資訊檢索效能

Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query

許志全*, 吳世弘*

Chih-Chuan Hsu and Shih-Hung Wu

摘要

當資訊檢索系統的使用者在不確知正確關鍵字時，可能會使用不精確的查詢詞描述其所需的訊息，嘗試著尋找所需要的資訊。我們稱這些不精確的查詢關鍵字爲籠統查詢(general query)。本文將評估線上搜尋引擎與資訊檢索研究者所研發的檢索系統對籠統查詢的檢索效能。現有的資訊檢索測試集不適合用於評估線上系統，因爲一來面對的文件集不相同，其次是測試集並不包含籠統查詢詞與精確查詢詞。爲了提供線上搜尋引擎與一般資訊檢索系統一個一致性的評比環境，我們利用維基百科建立一個測試集，這樣一來每個系統都可以檢索同樣的文件集容，同時可以比較籠統查詢詞與精確查詢詞的查詢結果。

在我們的檢索系統中，我們利用此測試集的鏈結結構特性，提出了新的查詢擴展方法。使用維基百科作爲查詢擴展方法的同義辭典，並與虛擬關聯回饋的查詢擴展方法結合，我們稱此方法爲維基百科查詢擴展。

實驗結果表明，本篇論文所建構的基於籠統查詢的資訊檢索測試集，能夠合理的評估線上搜尋引擎，且在籠統查詢與精確的關鍵字檢索效能的比較，可以明

* 朝陽科技大學資工系
Chaoyang University of Technology, Taichung, Taiwan
E-mail: shwu@cyut.edu.tw
The author for correspondence is Shih-Hung Wu.

顯的觀察到，籠統查詢的檢索效能的確較差。並且我們發現，在使用籠統查詢下，虛擬關聯回饋的檢索系統會優於主流的線上搜尋引擎，如:Google, Alta Vista。

而在查詢擴展方法的部分，適當的使用維基百科查詢擴展方法的確是可以提升檢索效能，而且只使用維基百科查詢擴展與只使用虛擬關聯回饋查詢擴展間效能的比較，顯示利用維基百科作為查詢擴展的同義辭典是很好的資源。

關鍵詞：資訊檢索、籠統查詢、測試集、維基百科、查詢擴展

Abstract

Users might use general terms to query the information in need, when the exact keyword is unknown. We treat these inexact query terms as general queries. In this paper, we consturct a test data set to evaluate the performance of online search engine on searching Wikipedia with general queries and exact queries.

We also proposed a new query expansion method that performs better on general queries. The Wikipedia query expansion method is regarding the Wikipedia as a thesaurus to find candidates of query expansion. The expanded queries are then combined with the pseudo relevance feedback. The performance of this method is better than online search engine on the general queries.

Keywords: Information Retrieval, General query, Test Collection, Wikipedia, Query Expansion.

1. 研究動機

在本篇論文中，我們將評估資訊檢索系統面對一個普遍發生的情況:「使用者不知道如何使用精確的關鍵字描述其資訊需求」。此時使用者只能使用籠統的關鍵字描述其資訊需求，並透過多次重新檢索之後，才找到精確的關鍵字，最後取得相關資訊。我們稱這些不精確的查詢關鍵字為籠統查詢(general query)。本文將評估線上搜尋引擎與我們所研發的檢索系統對籠統查詢的檢索效能。

現今的資訊檢索測試集(Test Collection)不適合評估使用籠統查詢與精確關鍵字的檢索效能，因為目前的各種大型測試集提供的是主題式的查詢資訊。如 TREC(Text REtrieval Conference)、CLEF(Cross Language Evaluation Forum)、NTCIR(NII Test Collection for IR Systems)等。各測試集典型的查詢主題(Topic)，都是使用主題當關鍵字描述其查詢的內容，而這些查詢主題是不區分為精確或籠統的。

在本研究主要可以分為兩個主題，第一部分是建置籠統查詢之資訊檢索測試集，主要是依據國際資訊檢索測試集機構(TREC、NTCIR)，以標準流程建構在 Web 文件上使用籠統查詢的資訊檢索測試集。第二部份是對查詢擴展作進一步的探討，其中我們提出

新的方法作查詢擴展，亦為虛擬關聯回饋與同義辭典的結合。

在建置測試集的標準流程，如圖 1 所示。標準的測試集是由三個文件集合所構成，即查詢主題(Query set)、文件集(Document set)、相關判斷集(Relevance Judgment set)，其中查詢主題與文件集是事先蒐集建構的，而相關判斷集則由利用各個不同的檢索系統的檢索結果中，透過 Pooling Method 建構出相關判斷的 Pool，最後再透過參與判斷之相關判斷者，對 Pool 中的每篇文件判斷所建構而成的。我們利用其維基百科所釋出的資料(Wikipedia Dump Data)，作為我們的測試集的內容。而查詢主題的建構，是由真實世界使用者對於維基百科全書知識需求所建構而成，其中包含了籠統查詢以及精確查詢等資訊。

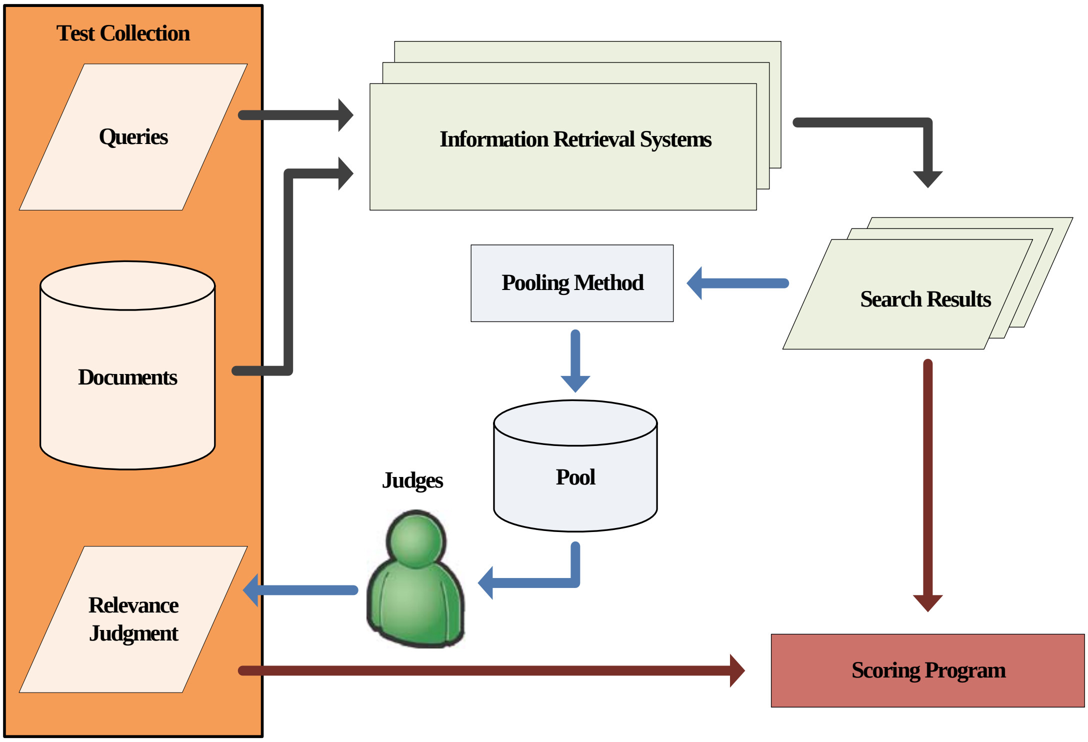


圖 1. 評估資訊檢索效能之流程圖

本研究的另一主題，探討查詢擴展檢索系統的效能。查詢擴展方法能夠有效提升資訊檢索系統的召回率(Recall)，而我們認為在籠統查詢上，由於使用者缺乏相關的資訊詳細描述其精確的資訊需求，我們可以透過查詢擴展的方法，幫助使用者找尋到更多更相關的文件。在查詢擴展的研究，我們將維基百科視為同義辭典，並結合虛擬關聯回饋的機制，自動的查詢擴展，以提升檢索的效能。維基百科擁有超鏈結的特性，每個條目(頁面)中都包含了許多超鏈結的鏈結文字(anchor text)，而這些字詞都是與條目擁有高度相關的字詞，我們將這些相關字詞視為查詢擴展的候選詞，並與虛擬關聯回饋中的字詞共同競選，以挑出更多相關的詞彙作查詢擴展。

2. 文獻探討

2.1 資訊檢索測試集 (Information Retrieval Test Collection)

資訊檢索評估是透過在一致性的評比環境中進行測試，衡量不同的資訊檢索技術或各檢索系統間的效能評比。此方法最早是由 Cleverdo 在 Cranfield II(Cleverdon, 1967)時提出，主要以文件集(Document set)、查詢主題集(Query set)、相關判斷集(Relevance Judgment set)建構測試集，作爲評估各系統的基礎資料，並訂定一套效能評估準則，評估各資訊檢索技術及檢索系統間的效能。在 1992 年美國 Defense Advanced Research Projects Agency (DARPA)與 National Institute of Standards and Technology(NIST)共同舉辦 Text REtrieval Conference (TREC) (Harman, 1993)，TREC 提供了當時最龐大的測試集，使得資訊檢索測試的環境更接近於真實。

繼 TREC 之後，各界對於提供資訊檢索一致性的評比環境，許多機構亦開始提供不同語系，相似於 TREC 的大型測試集，例如：Cross Language Evaluation Forum (CLEF) (Braschler, 2001)、NACSIS Test Collection for IR Systems (NTCIR) (Manning , Raghavan & Schütze, 2008)，這些機構與 TREC 每屆都會舉行各種不同的資訊檢索任務(Harman, 2005) (Ferro & Peters, 2008) (Kando, 2007)，如：單語言資訊檢索(Single Language IR, SLIR)、跨語言資訊檢索(Cross Language IR, CLIR)、跨多語言資訊檢索(Multi-Lingual IR, MLIR)。

在文件集中 TREC、CLEF、NTCIR 都會將收集的新聞文件加上各種不同的標記(Tag)，詳細的區分文件的特性，以利於系統進行剖析各種不同資訊，並有效的將其應用，表 1 爲 NTCIR 文件標記之說明(陳光華, 2001) (陳光華, 2004) (Chen & Chiang, 2000)，圖 2 爲 NTCIR 所使用的中文文件範例，其資訊是在陳光華教授所收集的 CIRB040r 文件集，本研究是透過參與 NTCIR-6 CLIR 與 NTCIR-7 IR4QA 之任務所取得。

表1. NTCIR使用文件標記說明

Tag		
<DOC>	</DOC>	The tag for each document
<DOCNO>	</DOCNO>	Document identifier
<LANG>	</LANG>	Language code: CH, EN, JA, K
<HEADLINE>	</HEADLINE>	Title of this news article
<DATE>	</DATE>	Issue date
<TEXT>	</TEXT>	Text of news article

近幾年 TREC、CLEF、NTCIR 等在不同的任務也使用了不同的文件集，如：TREC 的 Web Track(Craswell & Hawking, 2004)、Terabyte Track (Büttcher , Clarke & Soboroff 2006)、Blog Track (Ounis & Soboroff, 2008)，CLEF 的 WebCLEF (Jijkoun & Rijke, 2007)，NTCIR 的 WEB Task 等 (Eguchi , Oyama, Aizawa & Ishikawa, 2004)，其文件集的型態不同於以往的新聞文件，而是網際網路上的網路文本，即每一篇文件皆爲網頁。

```
<DOC>
<DOCNO>edn_xxx_20000101_0265056</DOCNO>
<LANG>CH</LANG>
<HEADLINE>總統府前升旗　喜迎千禧曙光</HEADLINE>
<DATE>2000-01-01</DATE>
<TEXT>
<P>記者蕭君暉／台北報導</P>
<P>全台陷入迎接千禧的狂熱！台東太麻里的海邊，數以萬計的人潮共同迎接台灣第一
道千禧曙光；...。</P>
</TEXT>
</DOC>
```

圖 2. NTCIR 中文文件範例

在查詢問題集中使用者描述其資訊需求稱之爲查詢問題(Query)，查詢問題中包含使用者所需的查詢關鍵字。TREC 提出使用查詢主題集(Topic Set)取代查詢問題集(Query Set)，其中的不同，在於查詢主題集以多欄位的方式陳述各種不同層次的查詢需求，而後 NTCIR、CLEF 等亦使用查詢主題集作爲各種查詢需求的陳述。表 2 爲 NTCIR 所使用 Topic 的標記及其意義(陳光華, 2001)(陳光華, 2004)(Chen & Chiang, 2000)。

TREC 在早期是使用模擬的方式建構查詢主題集，而 NTCIR 的查詢主題集則是來自於真實使用者的需求(CIRB010)，之後透過人工以及全文檢索工具的輔助，濾除敘述不清、不夠詳盡，或者主題涵蓋範圍太廣泛、主題不符合等(陳光華, 2001)。

表2. NTCIR 查詢主題標記及說明

<TOPOIC>	</TOPOIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language code: CH, EN, JA, KR
<TLANG>	</TLANG>	Target language code: CH, EN, JA, KR
<TITLE>	</TITLE>	The concise representation of information request, which is composed of noun or noun phrase.
<DESC>	</DESC>	A short description of the topic. The brief description of information need, which is composed of one or two sentences.
<NARR>	</NARR>	A much longer description of topic. The <NARR> has to be detailed, like the further interpretation to the request and proper nouns, the list of relevant or irrelevant items, the specific requirements or limitations of relevant documents, and so on.
<CONC>	</CONC>	The keywords relevant to whole topic.

相關判斷即爲由判斷者(人)判斷查詢主題與文件集中的每篇文件之相關程度。現今測試集的規模都相當龐大，無法閱讀所有文件，因此發展出 Pooling Method (Kageura & others, 1997)，此方法是假設真正相關的文件，會被多數的資訊檢索系統所檢索出，將所有資訊檢索系統檢索的結果，建構一個相關文件候選的 Pool，評估者只需要判斷相關候選 Pool 的文件，以此可降低建構相關判斷集的時間與人力。

NTCIR 在進行相關判斷時(陳光華, 2001)，每位判斷者必須詳細閱讀並瞭解查詢主題，並以查詢主題中<NARR>欄位作爲主要的判斷依據，將文件分到判斷者認爲最適當的相關類別。NTCIR 的相關判斷集分爲四個層級，如表 3.所示，然而 TREC 的相關判斷集是採取二元分層的方式(Harman, Braschler, Hess, Kluck, Peters & Schäuble, 2001)，TREC 的作法被視爲資訊檢索評估的標準流程，所以 NTCIR 亦採取二元分層的方式產生兩組相關判斷集，即嚴謹相關(Rigid Relevance)以及寬鬆相關(Relaxed Relevant)，嚴謹相關視”S”與”A”爲相關，寬鬆相關視”S”、”A”、”B”爲相關。

表3. NTCIR相關判斷層級

Label of Relevance	Sign	Score
Highly Relevant	S	3
Relevant	A	2
Partially Relevant	B	1
Irrelevant	C	0

NTCIR 中每一個查詢主題由 3 位判斷者做判斷，判斷者必須在一段連續的時間內完成一個查詢主題的判斷工作，以儘量確保判斷標準前後的一致性。之後透過以下公式結合三位判斷者的判斷:

$$R = \frac{avg(X_A + X_B + X_C)}{Z} \qquad (1)$$

其中 X 爲各判斷者對文件所給的類別等級，A, B, C 則爲三位判斷者之代號，Z 爲正規化參數(爲最高分數)。所得的值 R 介於 0 與 1 之間，若 R 愈接近 1，則表示二者愈相關。 其結合相關判斷分數時是視每位判斷者對於相關判斷的整體貢獻是相同的，所以不作特別加權，並且每個判斷都是獨立的。

嚴謹相關判斷集以及寬鬆相關判斷集的區分，則是透過訂立兩個門檻値 0.6667 與 0.3333，區分嚴謹以及寬鬆，如前述，嚴謹爲 B 以上之分數(2)，所以嚴謹的門檻值是透過以下的運算所取得，寬鬆則是 C 以上之分數(1)，亦透過相同的運算取得寬鬆之門檻値。

2.2 查詢擴展 (Query Expansion)

查詢擴展爲資訊檢索系統中常見的技術，最早由(Robertson & Sparck Jones, 1976) 所提出，主要的概念爲將原始的查詢詞擴展，將其加入至原始的查詢中，再使用擴展後的查詢作進一步的檢索，此種方法能夠有效提升資訊檢索系統的召回率(Recall)。

關聯回饋(Relevance Feedback)是一種藉由反覆查詢提高檢索精準度技術，其概念為透過第一次檢索出來的文件，取得其中與原始查詢的關聯程度，並回饋給檢索系統，而系統可以利用這些相關或是不相關的文件，修改檢索系統中的各種參數值或是修改原使的查詢(Query)，之後進行下一次的檢索時，即可得到較精準的檢索結果。

傳統關聯回饋方法以 Rocchio 等所提出的演算法(Joachims, 1997)最具代表性，其公式如下：

$$Q_{new} = \alpha Q_{current} + \frac{\beta}{|R'|}\sum_{D\in R'} D - \frac{\gamma}{|NR'|}\sum_{D\in NR'} D \tag{2}$$

Q_{new} 為經過關聯回饋後產生的新查詢，$Q_{current}$ 為舊有的查詢，*R'*表示與查詢相關的文件，NR'表示與查詢不相關的文件，α、β 及 γ 為參數比值，α+β+γ=1。其中 γ 通常設為 0，因為在真實世界中比較少會去區分文件為不相關，通常使用者最多只標注哪些文件為相關。

關聯回饋分為三種：顯性回饋（Explicit feedback）、隱性回饋（Implicit feedback）、和隱蔽的回饋（blind feedback）或 "虛擬"關聯回饋(Buckley, Salton, & Allan, 1994)(Harman, 1992)(Salton & Buckley, 1990)(Saracevic, 1970)(Sparck Jones & Rijsbergen, 1976)。顯性回饋指使用者主動標記哪些文件是相關或不相關。隱性回饋指系統監視使用者的行為，像是使用者有點選或沒點選哪些網頁、觀看網頁多久時間，收集這些資訊可以讓系統個人化。

隱蔽的回饋又稱為虛擬關聯回饋(Fan, Luo, Wang, Xi, &Fox, 2004)，由於使用者自己提供關聯回饋的意願不高，因此需要系統自動產生出模擬使用者所做的關聯回饋。系統會先進行一次檢索，擷取出 Top N 篇的文件當做虛擬關聯回饋文件，用來新增查詢字詞，讓最終檢索效能提高。

Okapi BM25 是一種排序的公式，搜索引擎在接受查詢句後，使用此公式排序相符合文件的高低，藉此找出相關文件出來。此公式是在 1970 年所發展出來屬於機率模式的演算法(Robertson & Sparck Jones, 1976)。現今許多資訊檢索的方程式都是改進自 BM25(Robertson, Walker, Sparck Jones, Hancock-Beaulieu & Gatford, 1995)。

Okapi 的公式如下：

$$Sim\left(Q, D_n\right) = \sum_{T\in Q} w^1 \frac{(k_1+1)tf\,(k_3+1)qtf}{(K+tf)(k_3+qtf)} \tag{3}$$

$$w^1 = \log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{4}$$

$$K = k_1((1-b)+b\frac{dl}{avdl}) \tag{5}$$

Okapi BM25 演算法將會進行兩次的檢索，第一次的檢索結果當成虛擬關聯回饋的文件，再從其中虛擬關聯回饋文件中，挑選出 n 個字詞加入查詢中，達到查詢擴展的作

用，再進行第二次的檢索。而在第一次的檢索時，其中的其小 R、r 的數值爲 0，第二次檢索時透過第一次檢索後取得 R、r 的數值。

第一次挑選出 n 個字詞加入原始的查詢中，挑選是透過 TF-IDF 的公式計算每個字詞的權重值：

$$w(i) = tf(i) \times idf(i) \tag{6}$$

$$idf(i) = \log \frac{N}{df(i)} \tag{7}$$

其中 tf(i)爲 i 字詞所出現的頻率(次數)，df(i)爲 i 字詞出現在多少篇文章的頻率，N 爲所有文章數量。

3. 建置籠統查詢的資訊檢索測試集

我們介紹如何利用維基百科全書的釋出資料(dump data)建置我們的資訊檢索測試集，以及建置測試集的成果。我們依據前述所介紹的測試集建構之標準流程建置測試集，在以下分別介紹我們所建構的文件集、查詢主題集、相關判斷集。

3.1 文件集 (Document Set)

我們希望建構一個繁體中文的文件集，所以必須將維基百科的內容作簡體中文轉換爲繁體中文的處理。在本研究中我們是透過 MediaWiki (http://www.mediawiki.org/)建置本地端的維基百科網站，將簡體中文與繁體中文參雜的內容，轉換爲只有繁體中文的內容。

文件收集流程如圖 3 所示，我們會將維基百科釋出資料中命名空間條目(Namespace Articles)以及重定向條目(Redirect Articles)濾除，這是因爲命名空間條目以及重定向條目，並沒有百科全書實質上的資訊，所以必須將其濾除。我們由維基百科釋出資料所擷取條目真正有內容的文件集 211,147 篇。

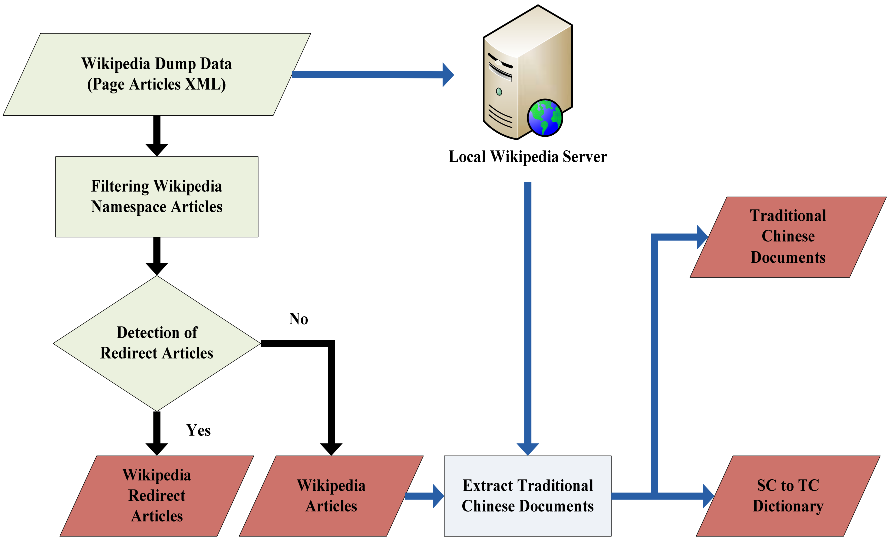


圖 3. 文件集收集流程

3.2 查詢主題集 (Topic set)

爲了使測試集的評比更接近真實的資訊檢索環境，查詢主題的建立通常都是透過真實使用者對於資訊的需求，並且須涵蓋多個不同的主題。所以我們收集了真實使用者對於維基百科的資訊需求。爲了評估現有的資訊檢索系統，是否能夠滿足使用者作籠統查詢，我們針對有缺乏精確關鍵字知識經驗的使用者收集其查詢 。我們透過訪談收集使用者在尋找資訊時，知道自己第一次查詢時使用的關鍵字是不精確的，並且經過幾次查詢可以找到可以滿足的資訊需求的文件。使用者修改其原始的查詢關鍵字最終找到需要的文件，這些修正前後的查詢就是我們所要蒐集的查詢主題。

挑選查詢總共有三個階段的篩選，第一階段，依據以上的假設蒐集，我們共收集了84 個籠統查詢。在第二階段，使用搜尋引擎輔助，使用籠統查詢到維基百科官方搜尋引擎作檢索，在前 Top 40 筆搜尋結果中必須有三篇以上使用者認爲非常相關或相關的文件，如果沒有則將此查詢濾除。最後再由使用者經由多次的檢索與閱讀其所需的資訊，修改籠統查詢成精確查詢，將這些資訊作更詳細的描述。如果發現此查詢主題不適用這個流程則將其刪除。

經由三階段篩選後，我們總共建立 34 個查詢主題。我們所建構查詢主題集各種使用的標記(Tag)及其定義如表 4 所示。

表4. 我們建置的查詢主題標記說明

Tags		Description
<TOPIC>	</TOPIC>	The tag for each topic
<NUM>	</NUM>	Topic identifier
<SLANG>	</SLANG>	Source language code: CH
<TLANG>	</TLANG>	Source language code: CH
< C-Query>	</ C-Query>	The concise representation of the general query
<DESC>	</DESC>	Description of this topic with one sentence
<NARR>	</NARR>	Length description of this topic, which contains two more tags: <BACK>, and<REL>
<BACK>	</BACK>	The background knowledge of the topic
<REL>	</REL>	How to judge the relevance
< EXACT >	</ EXACT >	The concise representation of information request, which is composed of noun or noun phrase.

3.3 相關判斷集 (Relevance Judgment set)

爲了降低建構相關判斷集所花費的時間以及人力，必須使用 Pooling Method 建構相關判斷候選集。我們的 Pool 是由 Google (http://www.google.com/)、Altavista (http://www.altavista.com/)、Wikipedia (http://zh.wikipedia.org/)、Wikigazer (http://wil.csie.cyut.edu.tw/Wikigazer/)等線上搜尋引擎，分別使用籠統查詢以及精確查詢作查詢，每次查詢取得最多 1000 筆的搜尋結果，查詢時間為 2009 年 1 月，所以每個查詢主題會由 8 個搜尋結果建構成一個 Pool。其中 Wikipedia 及 Wikigazer 為維基百科專屬的搜尋引擎，而 Google 及 Altavista 則是透過此兩種搜尋引擎指定網域搜尋中文維基百科，例如:兵法 site:zh.wikipedia.org。

當完成上述之前處理後，我們將建構每個查詢主題的 Pool(相關候選文件)，以提供相關判斷者作判斷。建構 Pool 之後，每個查詢主題皆有兩人參與相關判斷，每篇文件與查詢主題相關程度是依據 NTCIR 所訂定的四個層級(2.1 小節)，之後結合兩位相關判斷者之判斷分數，是透過公式(8)計算，並且參照 NTCIR 之寬鬆及嚴謹門檻值，0.3333 與 0.6667，產生兩組相關判斷集，即寬鬆相關判斷集以及嚴謹相關判斷集。

$$R = \frac{avg(X_A + X_B)}{Z} \qquad (8)$$

我們利用 kappa 統計量，統計每個查詢主題的兩位判斷者所做判斷之一致性分析，其公式如(9)所示，kappa 的假設為:判斷者在有意識的情況下所做的判斷，其一致性的結果應該大於隨機判斷的結果，其中 P(A)為兩位判斷者所做的判斷中，觀測到一致性判斷的機率，P(E)則代表為兩位判斷者偶然一致性判斷的機率。根據 An Introduction to Information Retrieval(Manning, Raghavan, & Schütze,2008)第八章所說明，kappa 值大於 0.8 為屬於好的一致性判斷，若借於 0.67~0.8 之間則屬於能認可的一致性判斷，如果小於 0.67 則屬於不好的判斷。

$$kappa = \frac{P(A) - P(E)}{1 - P(E)} \qquad (9)$$

圖 4 為我們各個查詢主題的 kappa 值統計分析，其總平均 kappa 值為 0.91，最差為查詢主題編號 22，其 kappa 為 0.63，而編號 5、17、20，是介於 0.67~0.8 之間，其餘 30 個查詢主題皆高於 0.8。基於 kappa 的統計分析，我們將捨棄查詢主題 022，其餘 33 個查詢主題則用於評估線上搜尋引擎與資訊檢索系統之效能。

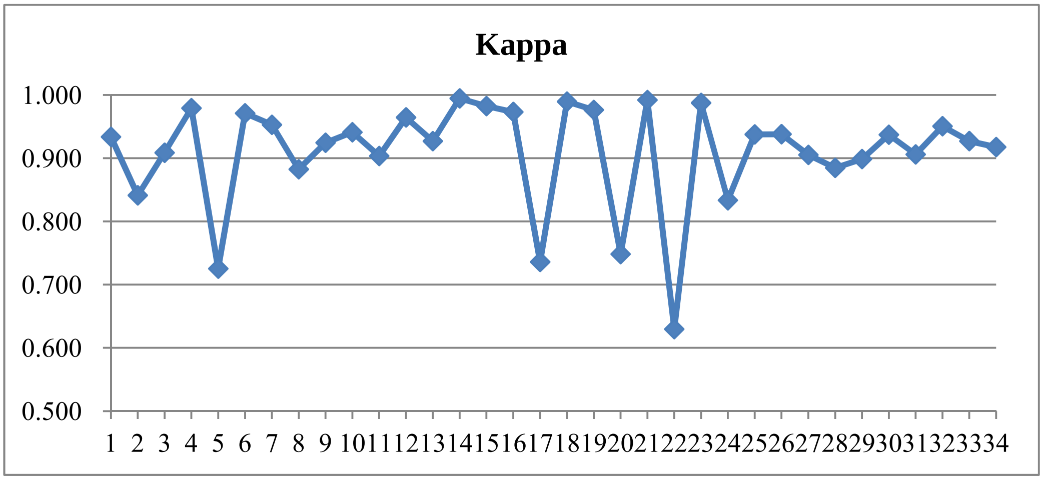


圖 4. 相關判斷的 Kappa 統計分析

4. 查詢擴展方法與系統

4.1 維基百科查詢擴展 (Wikipedia Query Expansion)

維基百科查詢擴展的概念，是利用維基百科擁有高品質、更新迅速的特性，當作是額外的同義辭典，以此輔助虛擬關聯回饋機制的查詢擴展方法。維基百科中每個頁面(條目)都有一個唯一的標題，而在每個頁面中都包含了各種超鏈結以及其鏈結文字連繫到其它相關的條目，而這些鏈結文字即為與此條目高度相關且重要的字詞，利用原始的查詢中的關鍵字，與維基百科中的條目名稱比對，並收集此條目中的鏈結文字作為查詢擴展的候選字詞。此方法在 NTCIR-7 IR4QA 的任務中，已證明適當的使用維基百科查詢擴展，可以有效的增進檢索效能(Hsu, Li, Chen & Wu, 2008)。

在以上的維基百科查詢擴展方法，只考慮到條目中的超鏈結，亦即只使用到鏈結結構中的鏈出鏈結，而在本篇論文中我們還使用了鏈入鏈結的鏈結文字的資訊，因為此資訊與鏈出鏈結相同，很可能也是有高度相關且重要的資訊，所以我們將其加入到維基百科查詢擴展方法中，以期能夠找出更多相關的字詞，增進檢索的效能。

圖 5 為我們查詢擴展檢索系統的系統流程圖，我們使用 OKAPI BM25 作為排序的演算法，BM25 參數設定為：k_1=1.2、k_3=7、b=0.75。我們系統的檢索可以分為三個步驟，首先由第一次檢索的結果中取出 Top 100 篇文件作為虛擬關聯回饋的文件，第二，由維基百科中擷取出與查詢字詞相關的字詞，第三，使用 TF-IDF 計算查詢擴展候選字的權重，最後由這 Top 100 篇虛擬關聯回饋文件和維基百科網站中，挑選出 n 個查詢擴展字詞加入原始查詢之中，作查詢擴展，最後再進行第二次的檢索

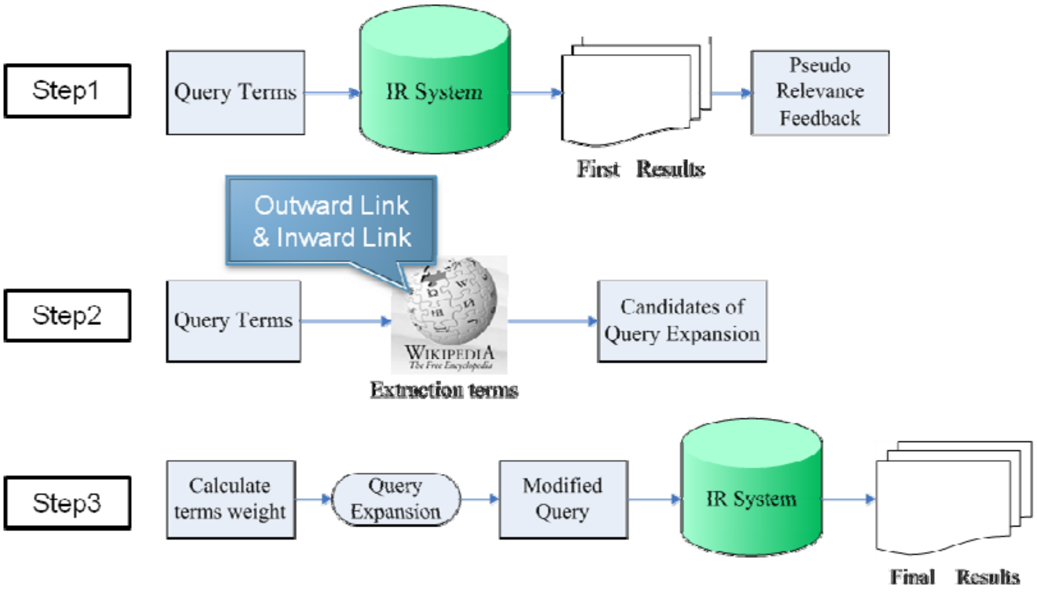


圖 5. 維基百科查詢擴展檢索系統流程圖

5. 實驗結果與分析

我們總共有三種實驗，這三種實驗所使用的資料爲我們所建置的籠統查詢的 Web 資訊檢索測試集。實驗一爲使用此測試集評估線上搜尋引擎以及查詢擴展檢索系統的檢索效能，而我們所使用的查詢擴展檢索系統爲 2.3 節所介紹的 Okapi BM25。實驗二爲 4.1 節所介紹的經由維基百科查詢擴展的效能評估。

5.1 實驗的目的與限制

在本文中，主要使用的資料爲 2009 年 1 月 16 日中文維基百科的釋出資料所建構的測試集，而我們所使用的評分方式爲 MAP，我們實驗的主要目的爲評估線上搜尋引擎的檢索效能，以及各種查詢擴展方法在檢索效能上的評估。在檢索時所使用的查詢有兩種，一爲籠統查詢，我們稱使用籠統查詢作檢索爲 C-run，另爲精確查詢，而使用精確查詢作檢索則爲 E-run。在評分時所使用的答案也分爲寬鬆(Relax)與嚴謹(Rigid)兩個答案集。所以在每個評估上會取得 4 種結果。我們的檢索系統主要是使用 Okapi BM25，其 Okapi 參數設定爲:K1=1.2、K3=7、B=0.75。

5.2 實驗的前處理

由於中文語系單字與單字之間並沒有以空格隔開，因此在中文語言的資訊檢索中，將句子斷詞這個前處理的步驟是相當重要的，在中文斷詞的部份，我們主要使用的是中研院 CKIP 小組開發的斷詞工具，其斷詞的平均準確度能夠達到 95%（http://ckipsvr.iis.sinica.edu.tw/），另外在處理爲中文維基百科中文斷詞時，我們避免將維基百科的條目名稱斷開，以提升檢索系統以及查詢擴展的效能。

5.3 評估方法

在本研究中主要以 MAP(Mean Average Precision) 評分公式，評估線上搜尋引擎以及我們所提出的新的 Okapi BM25 檢索系統的檢索效能 。線上搜尋引擎包含: Google、Wikipedia、Altavista、Wikigazer。MAP 公式如下:

$$MAP = \frac{1}{T}\sum_{j=1}^{T}\frac{\sum_{i=1}^{r_j}\frac{i}{Doc\ (i)}}{r_j} \qquad (10)$$

T 爲總共查詢主題數，r 表示在檢索文件中相關文件的數量，$Doc(i)$表示第 i 篇相關文件被檢索出來時，檢索文件的數量。

5.4 實驗1：評估線上搜尋引擎與查詢擴展演算法

5.4.1 實驗說明

本實驗主要評估的目標爲線上搜尋引擎與查詢擴展演算法，線上搜尋引擎有：Google、Wikipedia 官方搜尋引擎、Altavist、Wikigazer，查詢擴展演算法爲原始的 Okapi BM25，其 Okapi 參數設定爲:K1=1.2、K3=7、B=0.75，虛擬關聯回饋文件是第一次檢索的 Top 100 篇文件，並使用標準的 TFIDF 計算 Top 100 篇文件中的查詢擴展候選字詞，最後挑選 50 個字詞作查詢擴展。

5.4.2 實驗1結果

表5. C-run：使用寬鬆評估各搜尋引擎以及Okapi之MAP

	Okapi*	Google	Wikigazer*	Altavista	Wikipedia
Relax-MAP	**0.184**	0.145	0.133	0.087	0.082

表6. C-run：使用嚴謹評估各搜尋引擎以及Okapi之MAP

	Okapi*	Google	Wikigazer	Altavista	Wikipedia
Rigid-MAP	**0.185**	0.126	0.119	0.078	0.078

表7. E-run：使用寬鬆評估各搜尋引擎以及Okapi之MAP

	Google	Wikigazer	Okapi	Wikipedia	Altavista
Relax-MAP	**0.289**	0.287	0.259	0.198	0.189

表8. E-run：使用嚴謹評估各搜尋引擎以及Okapi之MAP

	Google	Wikigazer*	Okapi	Altavista	Wikipedia
Rigid-MAP	**0.326**	0.303	0.224	0.224	0.190

其中*代表此項數値與下一項(右側)數値，經由 T-test 統計檢定，其 P 値小於 0.05，則代表有顯著的差異。而由表 5~表 8 我們可以觀察到以下幾個重點：

(1)無論是使用寬鬆或嚴謹評估這些搜尋引擎或 Okapi，都可以發現 C-run 的檢索效能明顯的低於 E-run，這代表在現實生活中，在我們常使用的關鍵字搜尋引擎，如果使用了籠統查詢，其效能是不太能滿足使用者的需求，即使是著名的搜尋引擎 Google，其 C-run 的檢索效能亦不甚理想。

(2)在 C-run 中，我們發現基於虛擬關聯回饋的查詢擴展檢索系統 Okapi BM25 的檢索效能是最好的(高於 Google)，這代表在使用者使用籠統查詢時，這些相關但卻模糊的關鍵字，可以透過虛擬關聯回饋的機制，由其中挑選出正確的字詞，幫助使用者找到更多相關的文件。

(3)另外在寬鬆與嚴謹的結果中，會發現有幾個搜尋引擎的嚴謹會高於寬鬆的 MAP，經由我們分析之後發現，這是因爲嚴謹的相關文件數量低於寬鬆相關文件數量許多，在此情形下，使用 MAP 評估時其分母會縮小許多，而這些系統排序的方式又把嚴謹的答

案排序在很前面，所以嚴謹所得到的 MAP 值才會高於寬鬆的 MAP。

5.5 實驗2：維基百科查詢擴展

5.5.1 實驗說明

在本實驗中，我們將進行兩種實驗，實驗 2-1 爲使用維基百科中的鏈出鏈結與鏈入鏈結的鏈結文字同時作爲查詢擴展的候選字詞，在此稱其查詢擴展方式爲 WikiQE$_{out+in}$，並與 Okapi BM25 的 PRF 中的字詞共同競選，在此稱由 Okapi BM25 的 PRF 查詢擴展方式爲 OkapiQE，此實驗中會呈現使用 WikiQE$_{out+in}$ 與 OkapiQE 不同比重的查詢擴展效能。實驗 2-2 爲更詳細區分維基百科查詢擴展方法，將其區分爲鏈出鏈結 (WikiQE$_{out}$)、鏈入鏈結 (WikiQE$_{in}$)與 OkapiQE，針對不同比重的查詢擴展比較。

在這兩個實驗，我們將進行使用不同的查詢擴展字詞數，由 10、20 到 500，分別使用來自於不同查詢擴展方法的比重作查詢擴展，實驗中分別使用籠統查詢與精確查詢，且評分時也區分爲寬鬆與嚴謹。由於實驗的數據太過龐大，所以我們只針對 10~50 的查詢擴展字詞數的特別數據作呈現。

5.5.2 實驗2-1結果

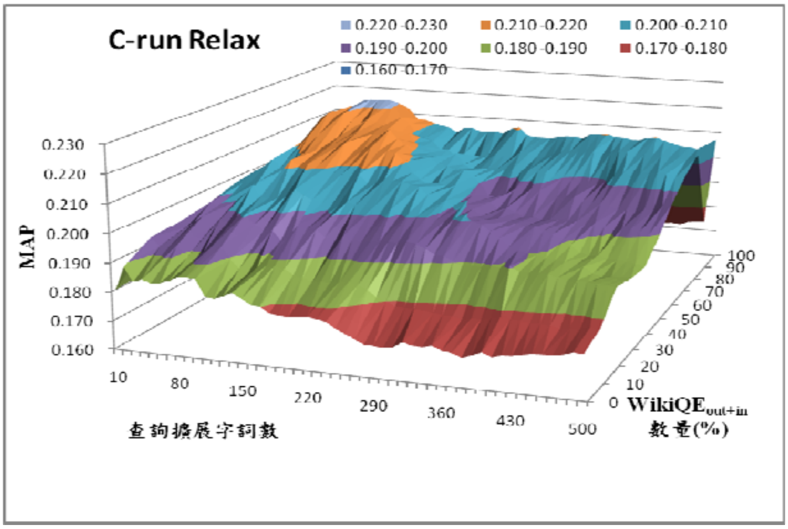


圖 6. C-run：不同查詢擴展方法比重與不同擴展字詞數之寬鬆 MAP

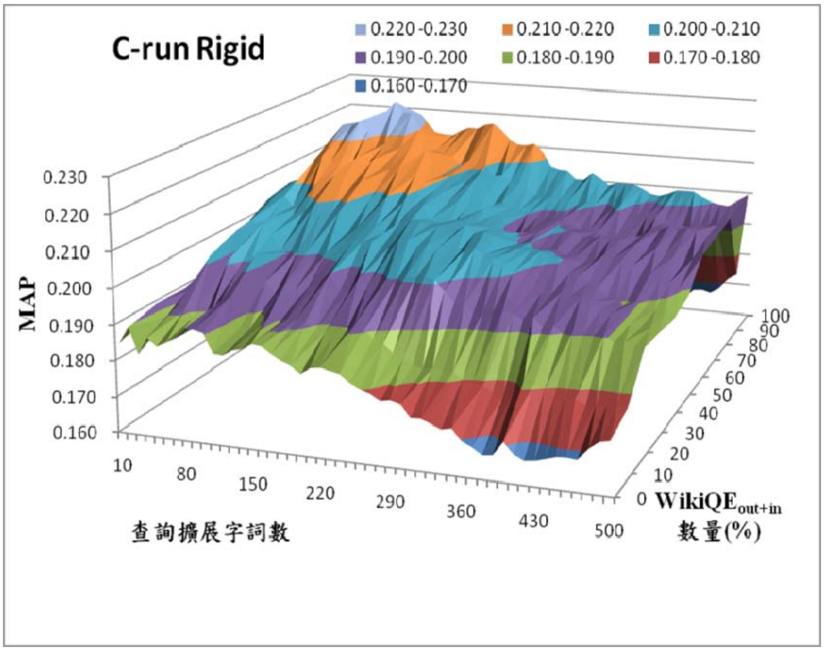


圖 7. C-run：不同查詢擴展方法比重與不同擴展字詞數之嚴謹 MAP

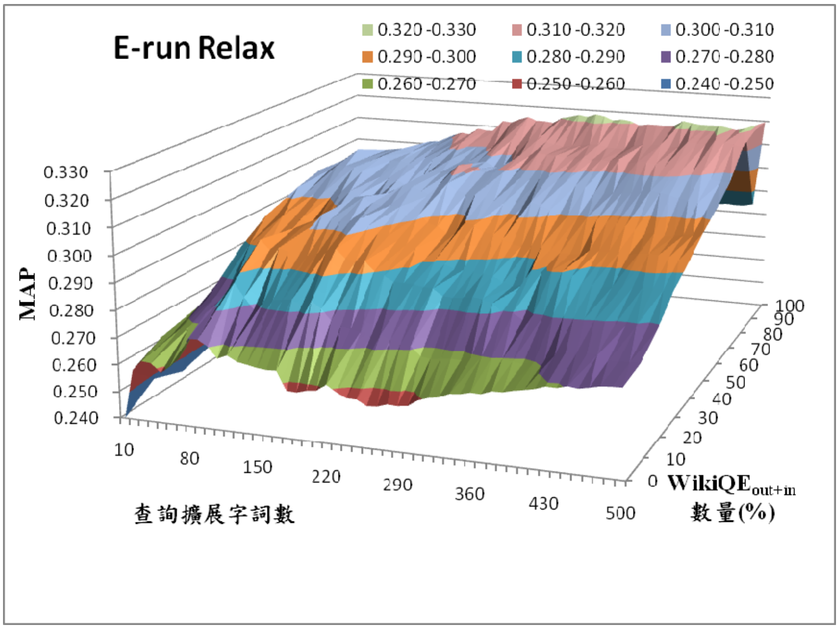


圖 8. E-run：不同查詢擴展方法比重與不同擴展字詞數之寬鬆 MAP

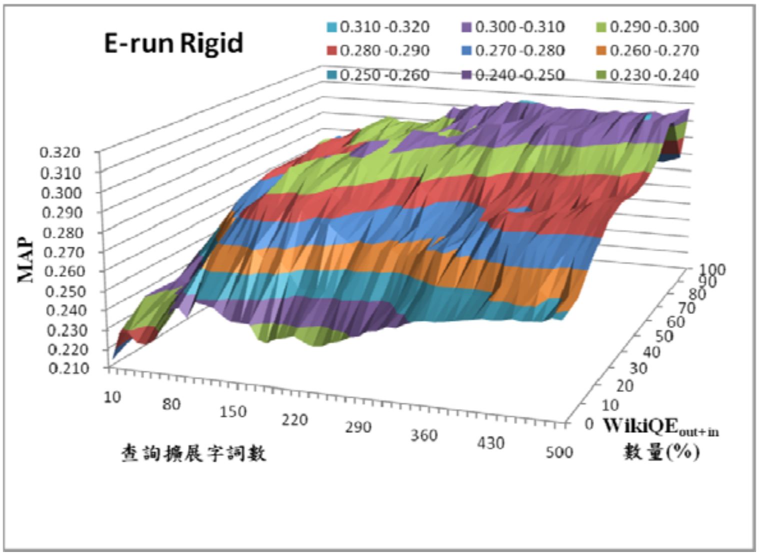


圖 9. E-run：不同查詢擴展方法比重與不同擴展字詞數之嚴謹 MAP

由圖 6~圖 9 的實驗結果中，我們歸納以下幾個重點：

(1)維基百科查詢擴展的確可以改進檢索的效能，無論是在使用籠統查詢或精確查詢，或是在寬鬆、嚴謹的評估下，適度的使用 OkapiQE 與 WikiQE$_{out+in}$ 的結合，能夠幫助挑選出更相關的查詢擴展字詞。在 C-run 中最佳查詢擴展字詞的比重是，80%來自於 WikiQE$_{out+in}$，20%來自於 OkapiQE。而在 E-run 中最佳查詢擴展字詞的比重是，60%來自於 WikiQE$_{out+in}$，40%來自於 OkapiQE。

(2)在不同查詢擴展字詞數的分析上，在有使用 WikiQE$_{out+in}$ 查詢擴展時，在使用較多的查詢擴展字詞上會得到比較好的 MAP，而單純使用 OkapiQE 查詢擴展時，對於使用不同字詞數的影響並不大。

(3)表 9 為 Topic 編號 009，C-run 檢索，而各種不同查詢擴展字詞的例子。首先，此查詢主題其精確關鍵字是”孫子兵法、孫武兵法、孫子、孫武”等字詞，而使用”風林火山、兵法、謀略”等籠統查詢，可以取得的字詞可由各個欄位作詳細檢視。由於我們所使用的 Okapi BM25 的演算法，因有 PRF 的查詢擴展，所以其演算法容許有重複的字詞出現，而我們的查詢擴展字詞來源總共有三種:PRF、維基百科的 Outward Link 與 Inward Link，所以同一個字詞其容許出現最多三次，由表中我們可以看到查詢擴展的效果，的確是可以有效的找出與籠統關鍵字高度相關的字詞。

表9. 查詢主題009；C-run；OkapiQE、WikiQE$_{out+in}$查詢擴展字詞

	OkapiQE:WikiQE$_{out+in}$		
Topic	100:0	30:70	0:100

General Query 風林火山、兵法、謀略	噴發、**孫子兵法**、武田、風林火山、**孫子**、聖海倫火山、爆發、山鵰、林峯、武田信玄、火星、岩漿、司馬佃、聖托里尼、火山爆發、海倫、helens、武者、地質、司馬法、公園、將軍、司馬、mount、加勒比板塊、火山碎屑、板塊、碎屑、活火山、軍事、大□、灰獴、拉森火山、戰略、孫臏兵法、孫臏、軍團、accessed、凌日、mandarin、mountains、里茲、板垣信方、次髎、國家、川中島、泥火山、形成、harris、九降風	噴發、**孫子兵法**、武田、風林火山、**孫子**、聖海倫火山、爆發、山鵰、林峯、武田信玄、火星、岩漿、司馬佃、聖托里尼、火山爆發、風林火山、(大河劇)、日本戰國、武田信玄、中國、**孫子兵法**、**孫子**、南北朝時代、奥州、北畠顯家、足利尊氏、村上源氏、井上靖、軍師、山本勘助、風林火山、(大河劇)、中國、兵家、中國、學術、知識份子、儒、法、晏武、修文、政治、文事、制度、宋、明、**孫子兵法**、中國、六韜、張良	風林火山、(大河劇)、日本戰國、武田信玄、中國、**孫子兵法**、**孫子**、南北朝時代、奥州、北畠顯家、足利尊氏、村上源氏、井上靖、軍師、山本勘助、風林火山、(大河劇)、中國、兵家、中國、學術、知識份子、儒、法、晏武、修文、政治、文事、制度、宋、明、**孫子兵法**、中國、六韜、張良、中國人民解放軍、中國共產黨、美國、前512年、吳楚之戰、黃帝、風後、握奇經、中國共產黨、抗日戰爭、八路軍
Exact Query 孫子兵法、孫武兵法、孫子、孫武			

5.5.3 實驗2-2結果

表10. C-run : OkapiQE 、WikiQE_{out} 、WikiQE_{in}之查詢擴展比較

	C-run									
	Only OkapiQE		Only WikiQE_{out}		Only WikiQE_{in}		Best		Worst	
	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid
50	0.184	0.185	0.191	0.197	0.158	0.167	0.212 (2:8:0)	0.212 (2:8:0)	0.158 (0:0:10)	0.167 (0:0:10)
40	0.188	0.188	0.188	0.197	0.159	0.169	0.207 (4:4:2 4:6:2)	0.211 (3:4:3 4:5:1)	0.159 (0:0:10)	0.169 (0:0:10)
30	0.187	0.184	0.173	0.184	0.161	0.167	0.204 (5:3:2 5:5:0)	0.210 (4:5:1)	0.161 (0:0:10)	0.167 (0:0:10)
20	0.191	0.190	0.173	0.187	0.157	0.165	0.199 (5:4:1 5:5:0)	0.198 (6:1:3)	0.157 (0:0:10)	0.162 (1:0:9)
10	0.183	0.186	0.164	0.176	0.155	0.165	0.184 (5:5:0 9:1:0)	0.188 (9:1:0)	0.152 (1:0:9)	0.157 (2:1:7)

表11. E-run : OkapiQE 、WikiQE$_{out}$ 、WikiQE$_{in}$之查詢擴展比較

	E-run									
	Only OkapiQE		Only WikiQE$_{out}$		Only WikiQE$_{in}$		Best		Worst	
	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid
50	0.259	0.224	0.253	0.256	0.250	0.243	0.300 (4:4:2*)	0.288 (0:4:6*)	0.250 (0:0:10)	0.226 (10:0:0)
40	0.262	0.225	0.256	0.260	0.245	0.237	0.292 (4:6:0)	0.278 (1:4:5*)	0.242 (0:1:9)	0.225 (10:0:0)
30	0.263	0.231	0.264	0.266	0.243	0.232	0.284 (3:7:0 4:6:0)	0.277 (0:6:4)	0.235 (0:1:9)	0.231 (8:0:2 10:0:0)
20	0.259	0.229	0.257	0.264	0.229	0.223	0.271 (3:7:0)	0.267 (1:9:0)	0.224 (0:1:9)	0.223 (0:0:10 0:1:9)
10	0.240	0.216	0.230	0.242	0.216	0.204	0.249 (5:5:0)	0.249 (1:6:3)	0.212 (0:4:6)	0.203 (1:0:9)

在表內*代表，此數值與只使用 OkapiQE 的 Relax 或 Rigid，其經由 T-test 統計檢定，其 P 值小於 0.05，亦即有顯著的差異。由表 10、表 11 的實驗結果中，我們歸納以下幾個重點：

(1)在 C-run 或 E-run 的結果顯示，只使用 WikiQE$_{out}$ 的效果與只經由虛擬關聯回饋 OkapiQE 的效果是很相近的，WikiQE$_{in}$ 的效果會比較差，這代表 WikiQE$_{out}$ 所找到的字詞是有很高度關連的字詞，而 WikiQE$_{in}$ 的輔助效果比較沒有 WikiQE$_{out}$ 好。在鏈結結構中，鏈出鏈結的鏈結文字通常是與此網頁非常相關的字詞，而鏈入鏈結對此網頁來說，其相關性並不一定很高，所以使用其鏈結文字查詢擴展的效能並不如使用鏈出鏈結的鏈結文字之效能好。

(2)在 Worst 的欄位中，可以看到最差的三種比重，以 OkapiQE 與 WikiQE$_{in}$ 占的比例最高，這代表此兩種擴展方式的效果是很接近的。在 C-run 中，Best 的欄位中，明顯的使用 WikiQE$_{out}$ 比重是偏高的，這代表在籠統查詢檢索中，透過維基百科內容中鏈出鏈結的鏈結文字，可以幫助我們由模糊、籠統的字詞來找到精確的字詞，以提升檢索效果。

(3)經由使用不同查詢擴展字詞數量的比較中，只使用 OkapiQE 的方式並沒有對效能有太大的影響，而只使用 WikiQE$_{out}$ 或 WikiQE$_{in}$ 的方式最佳的擴展字詞數約是 30~50 個擴展字詞。另外，我們可以看到在使用 50 個字詞擴展，E-run 使用 40%的 OkapiQE、40%WikiQE$_{out}$ 與 20%的 WikiQE$_{in}$ 效果會最好，寬鬆的 MAP 有 0.300，已經有 Google 的水平(見表 7)。

6. 結論

在本篇論文中，我們探討在真實世界的使用者，使用籠統查詢描述其資訊需求的情形。本研究依據建置測試集的標準流程，建構了一套基於籠統查詢的資訊檢索測試集。並且能夠使用在評估真實世界中線上的搜尋引擎，如 Google、Wikipedia、Altavista、Wikigazer 等。

由實驗結果證明，目前線上的搜尋引擎並不能滿足使用者使用籠統查詢，雖然 Google 使用精確查詢所檢索的效能的確很好，但是在籠統查詢上的表現並不是很好。另外我們發現基於虛擬關聯回饋的 Okapi BM25 演算法在籠統查詢的效果比 Google 還好，這證明基於 PRF 的查詢擴展方法能夠有效的幫助使用者在使用籠統查詢下的檢索效能。

在本文研究中，我們研究了維基百科鏈結結構的資訊，提出了維基百科查詢擴展方法，利用維基百科頁面中鏈出鏈結與鏈入鏈結的鏈結文字作為查詢擴展的候選字詞，並與 PRF 的集合結合，提升檢索的效能。由實驗結果證明，在使用籠統查詢的檢索中，透過維基百科鏈結資訊的鏈結文字，可以有效提升單純基於 PRF 查詢擴展方法(Okapi)的檢索效能；而在使用精確查詢的檢索中，此方法也能提升檢索的準確度，甚至可以達到 Google 的檢索水平。

參考文獻

陳光華(2001)。資訊檢索系統的評估 - NTCIR 會議。國立台灣大學圖書資訊學系四十週年系慶學術研討會論文集, 67-86, 台北：台灣大學。

陳光華(2004)。中文資訊檢索標竿測試集之建置, In *The Association for Computing Linguistics and Chinese Language Processing*, 15(4), 4-12.

Braschler, M. (2001). CLEF - Overview of Results. In Cross-Language Information Retrieval and Evaluation. *Lecture Notes in Computer Science 2069*, 89-101, Springer Verlag.

Buckley, C., Salton, G., & Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment, In *Proceedings of SIGIR 17.*, 292-300.

Büttcher, S., Clarke, C. L. A., & Soboroff, I. (2006). The TREC 2006 Terabyte track. In *TREC 2006*, Gaithersburg, Maryland USA.

Chen, K.H., & Chiang, Y.T. (2000). The Design and Implementation of the Chinese IR Benchmark. *Journal of Information,Communication, and Library Science*, 6(3), 61-80.

Cleverdon, C.W. (1967). The Cranfield Tests on Index Language Devices. In *Aslib Proceedings*, 19(6), 173-194.

Craswell, N., & Hawking, D. (2004). Overview of the trec-2004 web track. In *Proceedings of TREC-2004*, Gaithersburg, Maryland USA.

Eguchi, K., Oyama, K., Aizawa, A., & Ishikawa, H. (2004). Overview of the WEB task at the fourth NTCIR workshop. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization.*

Fan, W., Luo, M., Wang, L., Xi, W. & Fox, E. A. (2004). Tuning Before Feedback： Combining Ranking Discovery and Blind Feedback for Robust Retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,* 138-145.

Ferro, N., & Peters, C. (2008). From CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum. In *Proceedings of NTCIR-7 Workshop Meeting*, December 16-19, Tokyo, Japan.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipediabased explicit semantic analysis. In *IJCAI 2007. Proc. of International Joint Conference on Artificial Intelligence*, 1606-1611.

Harman, D.K. (1992). Relevance feedback revisited, In *Proceedings of SIGIR 15*, 1-10.

Harman, D.K. (1993). The First Text REtrieval Conference (TREC-1). *Information Processing and Management,* 29(4), 411-414.

Harman, D.K., Braschler, M., Hess, M., Kluck, M., Peters, C., & Schäuble, P. (2001). CLIR Evaluation at TREC. In *Peters(2001)*, S. 7-23.

Harman, D.K. (2005). *The TREC Test Collections*, Voorhees, E. M. and Harman, D. K. (eds.), TREC: Experiment and Evaluation in Information Retrieval, 21-52.

Hsu, C.C., Li, Y.T., Chen, Y.W., & Wu, S.H. (2008). Query Expansion via Link Analysis of Wikipedia for CLIR. In *Proceedings of NTCIR-7.*

Jijkoun, V., & Rijke, M. (2007). The University of Amsterdam at WebCLEF 2007: Using Centrality to Rank Web Snippets. In *CLEF 2007*, Budapest, Hungary, 2007.

Joachims, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proceedings of 14th International Conference on Machine Learning (ICML-97)*, 143-151.

Kageura, K. & others, eds. (1997). NACSIS Corpus Project for IR and Terminological Research. In *Natural Language Processing Pacific Rim Symposium '97*, 493-496, December 2-5, Phuket, Thailand.

Kando, N (2007). Overview of the Sixth NTCIR Workshop. In *Proceedings of the Sixth NTCIR Workshop*, May 15-18, NII, Tokyo.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University.

Ounis, I., & Soboroff, C.M.I. (2008). Overview of the TREC-2008 Blog Track. In *TREC 2008.*

Robertson, S.E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *The American Society for Information Science*, 27(3), 129-146.

Robertson, S.E., Walker, S., Sparck Jones, K., Hancock-Beaulieu, M., & Gatford, M. (1995). Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3).*

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288-297.

Saracevic, T. (1970). The Concept of ‘Relevance’ in Information Science: A Historical Review. In *Introduction to Information Science*, 111-151, New York, USA: R.R.Bowker.

Turmo, J., Comas, P.R., Rosset, S., Lamel, L., Moureau, N., & Mostefa, D. (2008). Overview of QAST 2008. In *Proceedings of the CLEF 2008 Workshop on Cross-Language Information Retrieval and Evaluation*.

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

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

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member：	NT$ 1,000 (US$50.-)
group membership：	NT$20,000 (US$1,000.-)

life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#：________________

Name：__________________________________ Date of Birth：_______________

Country of Residence：____________________ Province/State：____________________

Passport No.：___________________________ Sex: ________________________

Education(highest degree obtained)：____

Work Experience：____

____

Present Occupation：____

Address：____

____

Email Add：____

Tel. No：____________________________ Fax No：__________________________

Membership Category：□ Regular Member □ Life Member

Date： _____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）

Life Member ： US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號	（由本會填寫）		
姓　　名		性別		出生日期	年　月　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
申請人:　　　　（簽章） 中　華　民　國　　年　　月　　日					

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

PAYMENT FORM

Name :____(Please print)   Date: ___________________

Please debit my credit card as follows: US$_______________________

□ VISA CARD   □ MASTER CARD   □ JCB CARD   Issue Bank:_________________________

Card No.: __________- __________ - __________ - __________ Exp. Date:________________

3-digit code: ______________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE : ____________________________________

Tel.: ___________________________________ E-mail: ___________________________________

Add: ____

PAYMENT FOR

US$ _____ □ Computational Linguistics & Chinese Languages Processing (CLCLP)

Quantity Wanted: ______

US$ _____ □ Publications:____________________________________

US$ _____ □ Text Corpora: __________________________________

US$ _____ □ Speech Corpora:_________________________________

US$ _____ □ Others: _______________________________________

US$ __________ □Life Member Fee □ New Member □Renew

US$ __________ = Total

Fax : 886-2-2788-1638 or Mail this form to :

ACLCLP
℅ Institute of Information Science, Academia Sinica
R502, 128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會

信用卡付款單

姓名: ________________________________(請以正楷書寫)　日期:：______________

卡別：❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD　發卡銀行：________________

卡號:_________-_________-_________- ________　有效日期：_________________

卡片後三碼：_____________________（卡片背面簽名欄上數字後三碼）

持卡人簽名：__________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：_______________________ E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________❑ 中文計算語言學期刊(IJCLCLP)

NT$__________❑ 中研院詞庫小組技術報告

NT$__________❑ 中文（新聞）語料庫

NT$__________❑ 平衡語料庫

NT$__________❑ 中文詞庫八萬目

NT$__________❑ 中文句結構樹資料庫

NT$__________❑ 平衡語料庫詞集及詞頻統計

NT$__________❑ 中英雙語詞網

NT$__________❑ 中英雙語知識庫

NT$__________❑ 語音資料庫______________________________

NT$__________❑ 會員年費　❑續會　❑新會員　❑終身會員

NT$__________❑ 其他: __________________________________

NT$_______________=　合計

填妥後請傳真至 02-27881638 或郵寄至:
115台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

			AIR	AIR		
		Surface	(US&EURP)	(ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： □ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會
相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期）　年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：　______________________　收據抬頭：______________________

地　　址：____

電　　話：______________________　E-mail: ______________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.
1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.
2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.
3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.
4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).
5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript
6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.
7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

> Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

> Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).
Please visit the following websites for details.
(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)
(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

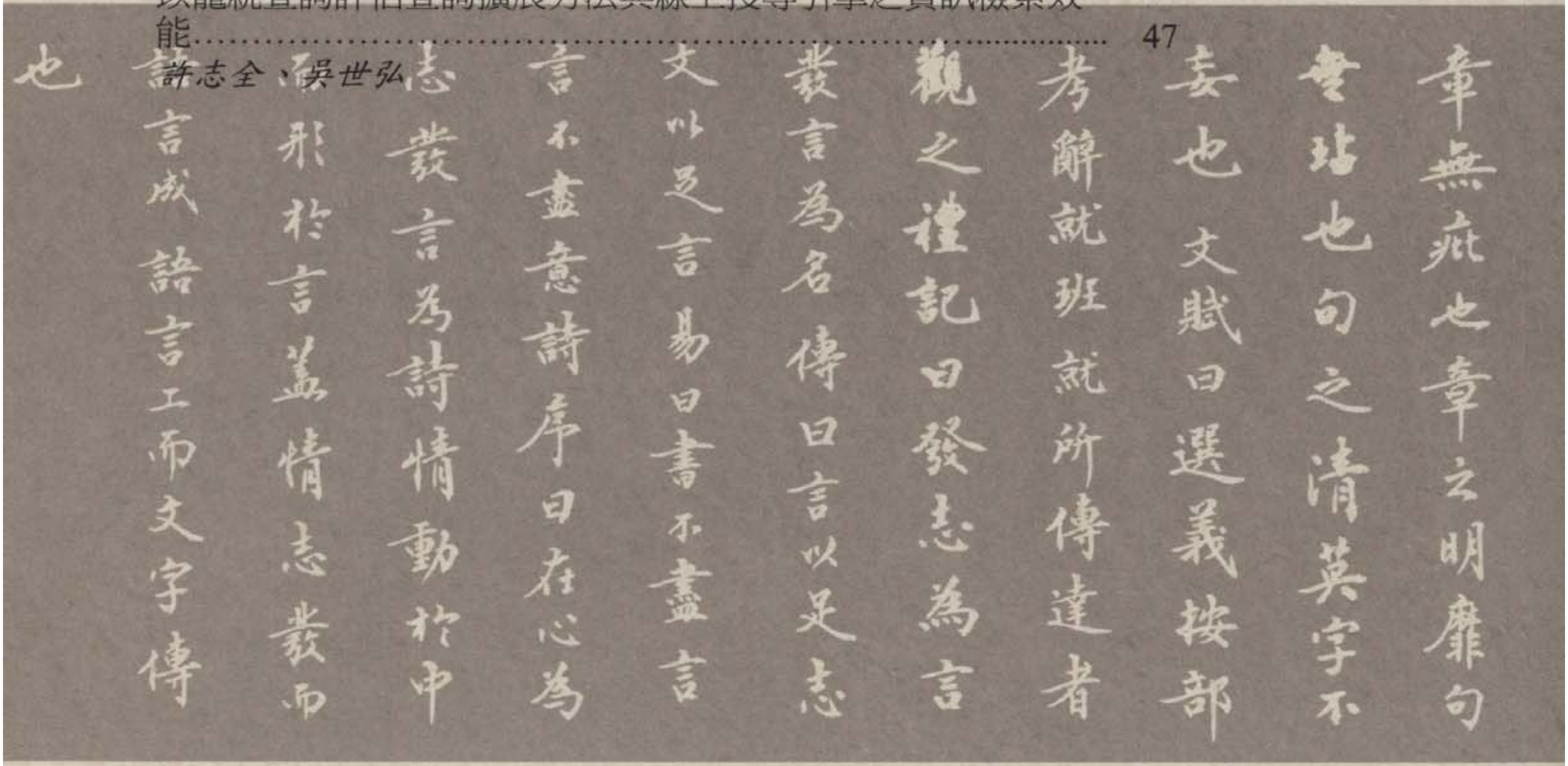
Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.16 no.1-2 March/June 2011

Contents

Papers



ISSN: 1027-376X
The Association for Computational Linguistics and Chinese Language Processing