Using Topic Identification in Chinese Information Retrieval

Ching-Long Yeh, Yi-Chun Chen
Department of Computer Science and Engineering
Tatung University
Taiwan
chingyeh@cse.ttu.edu.tw, yjchen7@ms7.hinet.net

Abstract

Information retrieval is to identify documents, from text collections, which are relevant with respect to some query. In current information retrieval systems, users can query with an unordered set of keywords, a question or a sentence. A list of document links matching the query can be retrieved and ordered by relevancy between the query and the documents. In this article, we are concerned with a hypothesis that the discourse-level element, topic, could be used to contribute the calculations of information retrieval. Due to the phenomenon of zero anaphora frequently occurring in Chinese texts, the topics may be omitted and are not expressed on the surface text. The key elements of the centering model of local discourse coherence are employed to extract structures of discourse segments. We propose a topic identification method using the local discourse structure to recover the omissions of topics and identify the topics of documents in the text collection. Then the topic information is inserted into the text for creating better indices. The experiment results are demonstrated on a test collection which is taken from Chinese Information Retrieval Benchmark, version 3.0.

Keywords: Natural Language Processing, Shallow Parsing, Topic Identification, Information Retrieval

1 Introduction

One of the most striking characteristics in a topic-prominent language like Chinese is the important element, "topic," in a sentence which can represent what the sentence is about [1]. That is, if we can identify the topics of utterances, we can obtain the most information embedded in text. In this paper, we tend to identify the topic of each utterance within a discourse based on the centering model [2]. However, in many natural languages, elements that can be easily deduced by the reader are frequently omitted from expressions in texts. The elimination of anaphoric expressions is termed zero anaphor which often occurs in Chinese texts, due to their prominence in discourse [1]. Accordingly, to identify the topic of each utterance in a discourse, we have to solve the problem of zero anaphora resolution.

There are several methods of anaphora resolution. One method is to integrate different knowledge sources or factors (e.g. gender and number agreement, c-command constraints, semantic information) that discount unlikely candidates until a minimal set of plausible candidates is obtained [2-6]. Anaphoric relations between anaphors and their antecedents are identified based on the integration of linguistic and domain knowledge. However, it is very labor-intensive and time-consuming to construct a domain knowledge base. Another method employs statistical models or AI techniques, such as machine learning, to compute the most likely candidate [7-10]. This method can sort out the above problems. However, it heavily relies upon the availability of sufficiently large text corpora that are tagged, in particular, with referential information [11]. Our method is an inexpensive, fast and reliable procedure for anaphora resolution, which relies on cheaper and more reliable NLP tools such as part-of-speech (POS) tagger and shallow parsers [12-16]. The resolution process works from the output of a POS tagger enriched with annotations of grammatical function of lexical items in the input text stream. The shallow parsing technique and the centering model are used to detect zero anaphors and identify the noun phrases preceding the anaphors as antecedents.

In this paper, an information retrieval system for retrieving news articles is implemented based on the TFIDF (term frequency/inverse document frequency) word weighting scheme [17] as the baseline. Then we identify and insert the topic information into the documents to create indices and see the improvement of information retrieval. Though in the previous similar works in English [18], they report on that resolution of anaphors may not help in information retrieval, the recent investigations shows that the pronominal anaphora resolution can contribute the improvement of information retrieval or question-answering systems [27-29]. In Chinese, topics or zero anaphors may carry more important information than pronouns do, due to their frequent occurrence in a Chinese discourse. For evaluating our approach, we took a subset of the articlesdocuments and topic sets form CIRB030 (Chinese Information Retrieval Benchmark version 3.0) [19] as the test collection for performing the experiment.

2 Shallow Parsing

Shallow (or partial) parsing which is an inexpensive, fast and reliable method does not deliver full syntactic analysis but is limited to parsing smaller syntactical related constituents [20]. For example, the sentence (1a) can be divided as (1b).

(1)a 花蓮 成為 熱門 的 旅遊 地點。
Hualian chengwei remen de luyou didian。
Hualian become popular NOM tour place
Hualien became the popular tourist attraction.
b [NP 花蓮] [VP 成為] [NP 熱門 的 旅遊 地點]
[NP Hualian] [VP chengwei] [NP remen de luyou didian]
[NP Hualien] [VP became] [NP the popular tourist attraction]

Given a Chinese sentence, our method of shallow parsing is divided into the following steps: First the sentence is divided into a sequence of POS-tagged words by employing a word segmentation program, Zhong wen duan ci xi tong (Chinese word segmentation system), which is a POS tagger developed by CKIP, Academia Sinica [21]. Second the sequence of words is parsed into smaller constituents such as noun phrases and verb phrases with phrase-level parsing. Each phrase is represented as a word list. Then the sequence of word lists is transformed into triples, [*S,P,O*]. For example in (2), (2b) is the output of sentence (2a) produced by the POS tagger and (2c) is the *triple* representation.

(2)a [花蓮(Nc) 成為(VG) 熱門(VH) 的(DE) 旅遊(VA) 地點(Na)]
b [[花蓮], np], [[成為], vp], [[熱門,的,旅遊,地點], np]
c [[花蓮], [成為], [熱門,旅遊,地點]]

The definition of *triple* representation is illustrated in Definition 1. The *triple* here is a simple representation which consists of three elements: *S*, *P* and *O* which correspond to the *Subject* (noun phrase), *Predicate* (verb phrase) and *Object* (noun phrase) respectively in a clause.

Definition 1

A Triple T is characterized by a 3-tuple:
T = [*S*, *P*, *O*] where
- *S* is a list of nouns whose grammatical role is the subject of a clause.
- *P* is a list of verbs or a preposition whose grammatical role is the predicate of a clause.
- *O* is a list of nouns whose grammatical role is the object of a clause.

In the step of *triple* transformation, the sequence of word lists as shown in (2b) is transformed into triples by employing the Triple Rules. The Triple Rules is built by referring to the Chinese syntax. There are four kinds of Triples in the Triple Rules, which corresponds to four basic clauses: subject + transitive verb + object, subject + intransitive verb, subject + preposition + object, and a noun phrase only. The rules listed below are employed in order:

Triple Rules

Triple1(S,P,O) → np(S), vtp(P), np(O).
Triple2(S,P,*none*) → np(S), vip(P).
Triple3(S,P,O) → np(S), prep(P), np(O).
Triple4(S,*none*,*none*) → np(S).

The vtp(P) denotes the predicate is a transitive verb phrase, which contains a transitive verb in the rightmost position in the phrase; likewise the vip(P) denotes the predicate is an intransitive verb phrase, which contains an intransitive verb in the rightmost position in the phrase. In the rule Triple3, the prep(P) denotes the predicate is a preposition. The Triple4 is employed if only a sentence contains only one noun phrase and no other constituent. If all the rules in the Triple Rules failed, the ZA Triple Rules are employed to detect zero anaphor (ZA) candidates.

ZA Triple Rules

Triple1^{z1}(*zero*,P,O)→ vtp(P), np(O).
Triple1^{z2}(S,P,*zero*)→ np(S), vtp(P).
Triple1^{z3}(*zero*,P,*zero*)→ vtp(P).
Triple2^{z1} (*zero*,P,*none*)→ vip(P).
Triple3^{z1}(*zero*,P,O) → prep(P), np(O).
Triple4^{z1}(*zero*,P,O) → co-conj(P), np(O).

The zero anaphora in Chinese generally occurs in the topic, subject or object position. The rules Triple1^{z1}, Triple2^{z1}, and Triple3^{z1} detect the zero anaphora occurring in the topic or subject position. The rule Triple1^{z2} detects the zero anaphora in the object position and Triple1^{z3} detect the zero anaphora occurring in both subject and object positions. In the Triple4^{z1}, the co-conj(P) denotes a coordinating conjunction appearing in the initial position of a clause. For example in (3), there are two *triples* generated. In the second *triple*, *zero* denotes a zero anaphor according to Triple1^{z1}.

(3)a張三 參加 比賽 贏得 冠軍。
Zhangsan canjia bisai yingde guanjun。
Zhangsan enter competition win champion
Zhangsan entered a competition and won the champion.
b [[[張三], [參加], [比賽]], [[zero], [贏得], [冠軍]]]
[[[Zhangsan], [enter], [competition]], [[zero], [win], [champion]]]

3 Topic Identification

Topic identification is similar to theme identification in [22]. The theme clearly corresponds to the backward-looking center of the centering model: the theme, under a general definition, is what the current utterance is about; what utterances are about provides a link to previous discourse, since otherwise the text would be incoherent. The role of the backward-looking center is precisely to provide such a link. In our approach, in addition to the centering model, we further employ the antecedent identification rule to identify the topic.

3.1 Centering Model

In the centering theory [23][2][5], the "attentional state" was identified as a basic component of discourse structure that consisted of two levels of focusing: global and local. For Grosz and Sidner, the centering theory provided a model for monitoring local focus and yielded the centering model which was designed to account for the difference in the perceived coherence of discourses. In the centering model, each utterance U in a discourse segment has two structures associated with it, called forward-looking centers, $C_f(U)$, and backward-looking center, $C_b(U)$. The forward-looking centers of U_n, $C_f(U_n)$, depend only on the expressions that constitute that utterance. They are not constrained by features of any previous utterance in the discourse segment (DS), and the elements of $C_f(U_n)$ are partially ordered to reflect relative prominence in U_n. Grosz et al., in their paper [2], assume that grammatical roles are the major determinant for ranking the forward-looking centers, with the order "*Subject* > *Object(s)* > *Others*". The superlative element of $C_f(U_n)$ may become the C_b of the following utterance, $C_b(U_{n+1})$. In addition to the structures for centers, C_b, and C_f, the centering model specifies a set of constraints and rules [2][24].

Constraints

For each utterance U_i in a discourse segment consisting of utterances $U_1, \ldots, U_m$:

1. U_i has exactly one C_b.
2. Every element of $C_f(U_i)$ must be realized in U_i.
3. Ranking of elements in $C_f(U_i)$ guides determination of $C_b(U_{i+1})$.
4. The choice of $C_b(U_i)$ is from $C_f(U_{i-1})$, and can not be from $C_f(U_{i-2})$ or other prior sets of C_f.

Backward-looking centers, C_bs, are often omitted or pronominalized and discourses that continue centering the same entity are more coherent than those that shift from one center to another. This means that some transitions are preferred over others. These observations are encapsulated in two rules:

Rules

For each utterance U_i in a discourse segment consisting of utterances $U_1, \ldots, U_m$:

I. If any element of $C_f(U_i)$ is realized by a pronoun in U_{i+1} then the $C_b(U_{i+1})$ must be realized by a pronoun also.

II. Sequences of continuation are preferred over sequence of retaining; and sequences of retaining are to be preferred over sequences of shifting.

Rule I represents one function of pronominal reference: the use of a pronoun to realize the C_b signals the hearer that the speaker is continuing to talk about the same thing. Psychological research and cross-linguistic research have validated that the C_b is preferentially realized by a pronoun in English and by equivalent forms (i.e. zero anaphora) in other languages [2]. Rule II reflects the intuition that continuation of the center and the use of retentions when possible to produce smooth transitions to a new center provide a basis for local coherence. For example in (4), the subjects of the utterance (4b) and (4d) are eliminated, and their antecedents are identified as the subjects of the preceding utterances (4a) and (4c) respectively[1] according to the centering model.

(4)a 電子股^{i}受美國高科技股重挫影響，
dianzigu^i shou meiguo gaokejigu zhongcuo yingxiang，
Electronics stock receive USA high-tech stock heavy-fall affect
Electronics stocks were affected by high-tech stocks fallen heavily in America.

b φ_1^i持續下跌。
φ_1^i chixu xiadie。
(Electronics stocks) continue fall
(Electronics stocks) continued falling down.

c 證券股^j 也有相對 回應，
zhengquangu^j ye you xiangdui huiying，
Securities stocks also have relative respondence
Securities stocks also had respondence.

d φ_1^j陸續下殺至跌停。
φ_1^j lixu xiasha zhi dieting。
(Securities stocks) continue fall by close.
(Securities stocks) fell by close one after another.

3.2 Zero Anaphora Resolution

The process of analyzing Chinese zero anaphora is different from general pronoun resolution in English because zero anaphors are not expressed in discourse. The task of zero anaphora resolution is divided into two phases: first zero anaphor detection and then antecedent identification. In the ZA detection phase, we use the ZA Triple Rules described in the Section 2 to detect omitted

[1] We use an φ_a^b to denote a zero anaphor, where the subscript a is the index of the zero anaphor itself and the superscript b is the index of the referent. A single φ without any script represents an intrasentential zero anaphor. Also note that a superscript attached to an NP is used to represent the index of the referent.

cases as ZA candidates denoted by *zero* in *triples*. Table 1 shows some examples corresponding to the ZA Triple Rules.

Table 1 Examples of Zero Anaphora

ZA Triple Rule	Example
Triple1^{z1} (*zero*,P,O)	φ zhuangdao yi ge ren (he) bump-to a person (He) bumped into a person.
Triple1^{z2} (S,P,*zero*)	Zhangsan xihuan φ ma Zhangsan like (somebody) Q[2] Does Zhangsan like (somebody)?
Triple1^{z3} (*zero*,P,*zero*)	φ xihuan φ (he) like (somebody or something) (He) likes (somebody or something).
Triple2^{z1} (*zero*,P,*none*)	φ qu gouwu le (he) go shopping ASPECT (He) has gone shopping.
Triple3^{z1} (*zero*,P,O)	φ zai nabian (he) in there (He) is there.
Triple4^{z1} (*zero*,P,O)	φ gen xiaopengyou wan (he) with child play (He) is playing with little children.

In the phase of antecedent identification, we concentrate on the resolution of ZA, and we first design the ZA identification constraints for filtering out the non-anaphoric cases[3] from the ZA candidates which are detected in the phase of ZA detection. In the case of cataphora,[4] because the first utterance has neither preceding utterances nor previous elements to be referred to as antecedents, the candidates detected in this utterance cannot be anaphors. By the observation of the test data, a news article sometimes has jushuo **"**it is said**"** as its first utterance, which is a case of expohora.[5] Therefore, the ZA identification constraint 1 is employed to eliminate the exophora or cataphora. In addition, the constraint 2 includes some cases might be incorrectly detected as ZAs, such as passive sentences or inverted sentences [25].

ZA identification constraints

For each ZA candidate *c* in a discourse:

1. *c* can not be in the first utterance in a discourse segment.
2. ZA does not occur in the following case:
   NP + *bei* + NP + VP + *c*
   NP (topic) + NP (subject) + VP + *c*

Most lexical knowledge such as person, number and gender employed in pronoun resolution in English cannot be utilized in zero anaphora resolution because the ZA itself is not expressed in text. In the antecedent identification, we employ the concept of *centers*[6] which are of the key elements of the centering theory [2][26] to establish the antecedent identification rule for identifying the antecedent of each ZA.

Antecedent identification rule

For each ZA *z* in a discourse segment $U_1, \ldots, U_m$:

If *z* occurs in U_i, and no ZA occurs in U_{i-1}

then choose the *preferred center* of U_{i-1} as the antecedent

Else if only one ZA occurs in U_{i-1}

then choose the antecedent of the ZA in U_{i-1} as the antecedent of *z*

Else if more than one ZA occurs in U_{i-1}

then choose the antecedent of the ZA in U_{i-1} as the antecedent of *z* according to the *forward-looking center ranking criterion*

End if

Forward-looking center ranking criterion

Topic > *Subject* > *Object* > *Others*

In the centering model as mentioned in Section 3.1, Grosz et al. assume that grammatical roles are the major determinant for ranking the forward-looking centers, with the order "*Subject>Object(s)>Others*". In Chinese, the concept of subject seems to be less significant while the topic in a sentence appears to be crucial in explaining the structure of ordinary sentences in the language [1]. By adopting the concept of grammatical roles and topic-prominence in Chinese, we order the grammatical roles in Chinese with topic having the highest priority and the order is referred to as the forward-looking center ranking criterion. This criterion is not only used to rank forward-looking centers but also employed to choose the antecedent of the ZA in the antecedent identification rule.

3.3 Topic Extraction

Grosz et al., in their paper [2], reported on that psychological research and cross-linguistic research have validated that the backward-looking center is preferentially realized by a pronoun in English and by equivalent forms (i.e. zero anaphora) in other languages. By adopting this notion, the key elements of the centering model of local discourse coherence and the vital characteristic, topic-prominence, in Chinese, we establish the topic identification rule for identifying the topics in text.

[2] We use a Q to denote a question (ma); an ASPECT to denote aspect marker.

[3] The non-anaphoric cases such as exophora or cataphora are the different research issues from the zero anaphora resolution. In our work, we do not intend to eliminate all non-anaphoric cases but to filter out some less complicated ones.

[4] Cataphora arises when a reference is made to an entity mentioned subsequently.

[5] Exophora is reference of an expression directly to an extralinguistic referent and the referent does not require another expression for its interpretation.

[6] The centers include forward-looking centers, the backward-looking center [2] and the preferred center [26].

Topic identification rule

For identifying each topic t in a discourse segment consisting of utterances $U_1, \ldots, U_m$:
If at least one ZA occurs in U_i
then refer to forward-looking center ranking criterion to choose the antecedent of the ZA as the t
Else if no ZA occurs in U_i
then refer to forward-looking center ranking criterion to choose one element of U_i as the t
End if

When a zero anaphor occurs in the utterance U_i, the antecedent of the zero anaphor is identified as the topic of U_i. Otherwise, if the transition relation, center shifting, occurs, topic will not be identified as any of the element in the preceding utterance but the element in the current utterance according to forward-looking center ranking criterion. We now take the example (4) to identify each topic of the utterances (4a) to (4d) by employing the topic identification rule. The topic of (4a) is dianzigu 'Electronics stocks,' and the topic of (4b) is omitted identified as the antecedent of φ_1^i, dianzigu 'Electronics stocks.' Similarly, the topic of (4d) is zhengquangu 'Securities stocks,' which is referred to as the antecedent of the zeroed topic of (4c).

4 Information Retrieval System

For evaluating the work of topic identification, we implement a word-based information retrieval system which uses English and Chinese words as indexing terms.[7] Each document of the test collection is first segmented by the Chinese word segmentation system [21], and each utterance of a document is transformed into a list of POS-tagged words separated by blanks. After the transformation accomplished, each output document is taken as input to the system and is assigned a document number as identification (docID). Every word in an input document d is taken as an indexing term t, whose weight $w(t,d)$ is calculated as term frequency and inverse document frequency (TFIDF) [17] value by equations (1) and (2).

$$w(t,d) = tf_{t,d} \times idf_t \quad (1)$$

$$idf_t = \log(\frac{N}{df_t}) \quad (2)$$

where
i) $tf_{t,d}$ is the within-document term frequency (TF).
ii) N is the number of all training documents.
iii) df_t is the number of training documents in which t occurs.

The system creates an index data file which stores an indexing term list in the order of the ASCII (American Standard Code for Information Interchange) code of the each indexing term's first character. An indexing term was followed a sequence of docID-weight value pair. In addition to the index data file, the system also builds an ASCII index file, which stores the ASCII codes of all indexing terms' first characters and records their positions in the index data file. By referring the ASCII index file, the system can obtain the starting position for efficiently searching the indexing terms. For example, the ASCII code of the Chinese character 汽 is *A854*, whose position is *355* recorded in the ASCII index file. The position *355* is the starting position of indexing terms having the first character, 汽, like 汽水 "soda-water" and 汽車 "automobile" in the index data file. By referring the ASCII index file, the system can obtain the starting position for efficiently searching the indexing terms.

5 Experiment and Result

We use topic identification to improve information retrieval. For evaluating our approach, we take part articles of China Times Express and Central Daily News form CIRB030 [19] as the test collection.[8] We selected 592 news articles of China Times Express containing more than 30,000 utterances as the test collection *A* according to the CIRB030 Answer Set. The Answer Set is a list of document numbers (DOCNO) assigned the topic IDs and their relevance in four categories: "Highly Relevant", "Relevant", "Partially relevant", and "Irrelevant." Because many topic categories do not have enough relevant articles for observing the results, we put additional 30 relevant news articles of Central Daily News into the test collection *A* as the test collection *B*.

5.1 Zero Anaphora Resolution

Topic identification is to identify and extract topics of utterances in text. Due to the phenomenon of zero anaphora occurring in Chinese texts frequently, we need to solve the problem of zero anaphora resolution. The test collection *B* including 622 news articles is used to test our method of zero anaphora resolution. In the test collection, the average number of utterances of an article is 52, while 17ZAs occurs in these utterances. The recall rates (RR) and precision rates (PR) of zero anaphora resolution are 0.67 and 0.64 respectively calculated using equation (3) and equation (4). Most errors occur when a zero anaphor refers to an entity other than the corresponding grammatical role in the preceding utterance, or refers to other entity in the more previous utterance.

[7] A Chinese word here is a meaningful word consisting of one or more Chinese characters, such xuexiao "school" and jiaru "join".

[8] CIRB030 developed by Chen, K. H., National Taiwan University is a test collection designed to be used for performance evaluation of Chinese document retrieval. The test collection contains three parts: Document Set, Topic Set and Answer Set. It is a helpful and powerful tool for investigation of the developing systems and the developed systems.

$$\text{PR of ZA resolution} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of ZA candidates}} \quad (3)$$

$$\text{RR of ZA detection} = \frac{\text{No. of antecedent correctly identified}}{\text{No. of ZA occurred in text}} \quad (4)$$

5.2 Information Retrieval

We performed an experiment to examine the effectiveness of using topic identification for information retrieval. In the experiment, we take the test collection *A* as input to the information retrieval system as the baseline, and then insert topic information for all utterances in an article into the article to see the result. The keywords relevant to topics of CIRB030 Topic Set are taken as the queries for test. The recall rates and R-precision rates of information retrieval are calculated using equation (5) and equation (6) respectively. Table 2 shows the result of the experiment on the test collection *A*.

$$\text{RR of information retrieval} = \frac{\text{No. of relevant articles retrieved}}{\text{No. of articles retrieved for a query}} \quad (5)$$

$$\text{R-PR of information retrieval} = \text{the PR at the top R articles retrieved in the ranking} \quad (6)$$

$$\text{PR of information retrieval} = \frac{\text{No. of relevant articles retrieved}}{\text{No. of articles retrieved}} \quad (7)$$

Table 2 Results of the Experiment on the Test Collection *A*

	Articles retrieved	RR	R-PR
Baseline	5	0.40	0.35
	10	0.56	
	20	0.72	
After topic identification	5	0.42	0.40
	10	0.64	
	20	0.82	

The experiment is performed repeatedly by replacing the test collection with the test collection *B*, and the keywords of six topic categories, which have ten relevant articles each, are taken as the queries. The recall rates and precision rates of information retrieval are calculated using equations (5) and (7). Table 3 shows the result of the experiment on the test collection *B*.

Table 3 Results of the Experiment on the Test Collection *B*

	Articles retrieved	RR	PR
Baseline	10	0.58	0.58
	20	0.80	0.40
After topic identification	10	0.65	0.65
	20	0.85	0.43

6 Conclusions

In this paper, we propose a method of topic identification based on the centering model to improve information retrieval in Chinese. According to observations on real texts, we found that to identify the topics in Chinese context is much related to the issue of zero anaphora resolution. Our method of zero anaphora resolution works on the output of a POS tagger and employs a shallow parsing instead of a complex parsing to resolve zero anaphors in Chinese text.

Due to the work of verifying the result of zero anaphora resolution is very laborious, we could only perform the experiment on a small test collection. However, the preliminary experimental result is promisingly to some extent even though the precision rate of zero anaphora resolution is 0.64. Our future work will proceed with increasing the accuracy of topic identification and further investigate the other related problems, such as pronoun resolution and co-reference resolution in Chinese.

Acknowledgements

We give our special thanks to CKIP, Academia Sinica for sharing the Chinese word segmentation system. We would like to thank to LIPS Lab. of LIS department and NLP Lab. of CSIE department in NTU, National Science Council, and ROCLING for providing CIRB030 for information retrieval research.

References

[1] Charles N. Li and Sandra A. Thompson, *Mandarin Chinese -- A Functional Reference Grammar*, University of California Press, 1981.

[2] B. J. Grosz, A. K. Joshi and S. Weinstein, *Centering: A Framework for Modeling the Local Coherence of Discourse*, *Computational Linguistics,* Vol.21, No.2, 1995, pp.203-225.

[3] S. Lappin and H. Leass, *An Algorithm for Pronominal Anaphora Resolution*, *Computational Linguistics*, Vol.20, No.4, 1994, pp.535-561.

[4] M. Okumura and K. Tamura, *Zero Pronoun Resolution in Japanese Discourse Based on Centering Theory*, *Proc. of the COLING'96*, Copenhagen, Denmark, 1996, pp.871-876.

[5] M. A. Walker, Centering, *Anaphora Resolution, and Discourse Structure*, In Walker, M. A., Joshi, A. K. and Prince, E. F. (eds.), *Centering in Discourse*, Oxford University Press, 1998.

[6] Ching-Long Yeh and Yi-Chun Chen, *An Empirical Study of Zero Anaphora Resolution in Chinese Based*

on Centering Theory, *Pro. of ROCLING XIV*, Tainan, Taiwan, 2001, pp.237-251.

[7] C. Aone and S. W. Bennett, *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies*, *Proc. of the 33rd Annual Meeting of the ACL*, Santa Cruz, New Mexico, 1995, pp.122-129.

[8] D. Connoly, John D. Burger and David S. Day, *A Machine Learning Approach to Anaphoric Reference*, *Proc. of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 1994, pp.255-261.

[9] N. Ge, J. Hale and E. Charniak, *A Statistical Approach to Anaphora Resolution*, *Proc. of the Sixth Workshop on Very Large Corpora*, Montreal, Quebec, Canada, 1998, pp.161-170.

[10] K. Seki, A. Fujii and T. Ishikawa, *A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution*, *Proc. of the COLING 2002*, Taipei, Taiwan, 2002, pp.911-917.

[11] R. Stuckardt, *Machine-Learning-Based vs. Manually Designed Approaches to Anaphor Resolution: The Best of Two Worlds*, *Proc. of the DAARC2002*, University of Lisbon, Portugal, 2002, pp.211-216.

[12] B. Baldwin, *CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources*, *Proc. of the ACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution*, Madrid, Spain, 1997, pp.38-45.

[13] A. Ferrández, M. Palomar and L. Moreno, *Anaphor Resolution in Unrestricted Texts with Partial Parsing*, *Proc. of the COLING'98)/ACL'98*, Montreal, Canada, 1998, pp.385-391.

[14] C. Kennedy and B. Boguraev, *Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser*, *Proc. of the COLING'96*, Copenhagen, Denmark, 1996, pp.113-118.

[15] Mitkov, Ruslan, *Anaphora Resolution*, Longman, 2002.

[16] C. L. Yeh and Y. C. Chen, *Zero Anaphora Resolution in Chinese with Partial Parsing Based on Centering Theory*, *Proc. of IEEE NLP-KE03*, Beijing, China, 2003, pp.683-688.

[17] G. Salton and C. Buckley, *Term Weighting Approaches in Automatic Text Retrieval*, *Information Processing and Management*, Vol.24, No.5, 1988, pp.513-523.

[18] S. Bonzi and E. D. Liddy, *The Use of Anaphoric Resolution for Document Description in Information Retrieva*l, *Information Processing and Management*, Vol.25, No.4, 1990, pp.429-441.

[19] K. H. Chen and H. H. Chen, *Overview of CIRB030 Information Retrieval Test Collection*, 2004, http://lips.lis.ntu.edu.tw/cirb/index.htm

[20] S. Abney, *Tagging and Partial Parsing,* In Ken Church, Steve Young and Gerrit Bloothooft (eds.), *Corpus-Based Methods in Language and Speech*, An ELSNET Volume, Kluwer Academic Publishers, Dordrecht, 1996.

[21] CKIP, *Zhong Wen Duan Ci Xi Tong* (Chinese word segmentation system) (in Chinese), *Academia Sinica*, 2003, *http://blackjack.iis.sinica.edu.tw/uwextract/*

[22] O. Rambow, *Pragmatic Aspects of Scrambling and Topicalization in German: A Centering Approach*, *Proc. of IRCS Workshop on Centering in Discourse,* Univ. of Pennsylvania, 1993.

[23] B. J. Grosz and C. L. Sidner, *Attention, Intentions, and the Structure of Discourse*, *Computational Linguistics,* Vol.12, No.3, 1986, pp.175-204.

[24] M. A. Walker, M. Iida and S. Cote, *Japanese Discourse and the Process of Centering*, *Computational Linguistics*, Vol.20, No.2, 1994, pp.193-232.

[25] Wenze Hu, *Functional Perspectives and Chinese Word Order*, Ph. D. dissertation, The Ohio State University, 1995.

[26] S. Brennan, M. Friedman and C. Pollard, *A Centering Approach to Pronouns*, *Proc. of the 25th Annual Meeting of the ACL*, 1987, pp.155-162.

[27] Richard J. Edens, L. Gaylard Helen, J. F. Jones Gareth and M. Lam-Adesina Adenike, *An Investigation of Broad Coverage Automatic Pronoun Resolution for Information Retrieval*, *Proceedings of SIGIR*, 2003, pp.381-382.

[28] R. Watson, J. Preiss and E. J. Briscoe, *The Contribution of Domain-Independent Robust Pronominal Anaphora Resolution to Open-Domain Question-Answering*, *Proc. of Int. Symposium on Reference Resolution and Its Application to Question-Answering and Summarisation*, 2003, p.75-82.

[29] J. L. Vicedo and A. Ferrandez, *Importance of Pronominal Anaphora Resolution in Question Answering Systems*, *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, Hong-Kong, 2000, pp.555-562.

Biographies

Ching-Long Yeh was born on Dec. 24, 1960 in Taipei, Taiwan. He is an associate professor of the Computer Science and Engineering Department, Tatung University, Taipei, Taiwan. He receives Ph.D. degree in Artificial Intelligence from the University of Edinburgh in 1995. His research interests include emergent

web technology (Web Services and Semantic Web), knowledge engineering and knowledge management, electronic business, e-learning and natural language processing.

Yi-Chun Chen obtained his M.S. and Ph.D. degree from Computer Science Engineering, Tatung University in 1999 and 2005, respectively. He was a technical manager of SimpleAct Inc. from Aug. 2000 to Oct. 2004. In SimpleAct, he worked on the NLP related applications such as question-answering system, e-mail routing system, information retrieval system and so on. He joined defend military service and was employed by QNAP Systems, Inc. in 2005. His research interest includes natural language processing, anaphora resolution and discourse analysis in Mandarin Chinese. After he joined QNAP, he is also interested in the techniques of Linux based embedded systems.