Using Zero Anaphora Resolution to Improve Text Categorization[1]

Ching-Long Yeh and Yi-Chun Chen
Department of Computer Science and Engineering
Tatung University
40 Chungshan N. Rd. 3rd. Section
Taipei 104
Taiwan
chingyeh@cse.ttu.edu.tw, d8806005@mail.ttu.edu.tw

Abstract

In Chinese, anaphors are frequently omitted, termed zero anaphor (ZA), from text due to their prominence. Thus the information carried by ZAs in text can not be used to contribute the calculation of text categorization. In this paper, we employ a ZA resolution method to recover the omissions of anaphors in text. Then the resulting text is used as the input of a text categorization system. The experiment result shows that ZA resolution method enhances the accuracy of text categorization from 79% to 84%.

1 Introduction

In Chinese text, anaphoric expressions are frequently eliminated, termed zero anaphor (ZA) hereafter, due to their prominence in discourse (Li and Thompson, 1981). Zero anaphors are generally noun phrases that are understood from the context and do not need to be specified. Anaphors in Chinese can be classified as zero, pronominal and nominal forms, as exemplified in (1) by φ^i, 他^i and 那 人^j, respectively[2]. Zero anaphors are generally noun phrases that are understood from the context and do not need to be specified.

(1) a. 張三^i 驚慌 的 往 外 跑，
Zhangsan frightened and ran outside.

b. φ^i 撞到 一個 人^j，
(He) bumped into a person.

c. 他^i 看清 了 那 人^j 的 長相，
He saw clearly that person's appearance.

d. φ^i 認出 那 人^j 是 李四。

(He) recognized that man as Lisi.

The methods of anaphora resolution can be classified into traditional and alternative approaches. The former integrates 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 (Grosz *et al.*, 1995; Lappin and Leass, 1994; Okumura and Tamura, 1996; Walker *et al.*, 1998; Yeh and Chen, 2001). 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. The latter employs statistical models or AI techniques, such as machine learning, to compute the most likely candidate

[1] This research was supported by the Taiwan National Science Council under Contract No. NSC92-2422-H-036-322

[2] We use a φ^a to denote a zero anaphor, where the superscript *a* is the index of the antecedent. Also note that superscripts attached to NPs and pronouns are used to represent the indices of their antecedents.

(Aone and Bennett, 1995; Connoly *et al.*, 1994; Ge *et al.*, 1998; Seki *et al.*, 2002). This approach 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 (Stuckardt, 2002).

A recent trend is in search of inexpensive, fast and reliable procedures for anaphora resolution (Baldwin, 1997; Ferrández *et al.*, 1998; Kennedy and Boguraev, 1996; Mitkov, 1998). The approach relies on cheaper and more reliable NLP tools such as part-of-speech (POS) tagger and shallow parsers. Without complex syntactic, semantic and discourse analysis, we work on the output of a part-of-speech tagger, AUTOTAG (CKIP, 1999), which is developed by CKIP, Academia Sinica and use a partial parsing instead of a complex parsing to resolve zero anaphors in Chinese text.

Text categorization is the task of classifying documents into a certain number of predefined categories. A list of keywords is used to represent a document or a class, so that a free document can be categorized by comparing its keyword list and those of document classes. A wide range of supervised learning algorithms has been applied to this issue, using a training data set of pre-classified documents. The naive Bayes, k-nearest neighbors and Rocchio are well-known algorithms (Joachims, 1997; Schapire *et al.*, 1998; Tsay and Wang, 2000, Yang *et al.*, 2002). Without training data set, the unsupervised training methods use techniques of clustering which group similar documents into one cluster that no longer distinguishes between constituent documents (Ko and Seo, 2000). The main difference between these two training methods is that supervised training method needs the pre-classified documents for the training data set. In general, the accuracy of text categorization based on supervised learning is better than based on unsupervised learning.

In this paper, we employ a ZA resolution method to recover the omissions of anaphors in text in order to improve the accuracy of text categorization. The text with ZA resolved is used as the input of a text categorization system. The new text categorization system works as below: First an input document with ZA resolved is taken as a ZA-resolved input document which each zero anaphor in the text is replaced by its antecedent. Second the ZA-resolved input document is categorized by the *k*-NN classifier. The result shows the new text categorization system increases the accuracy from 79% to 84%. In the following sections we first describe our ZA resolution method that works on the output of a POS tagger, and use a partial parsing instead of a complex parsing to resolve zero anaphors in Chinese text. In section 3 we describe the detail of the text categorization system and how to apply the resulting text of ZA resolution to it. In section 4 we illustrate the result of applying our method of ZA resolution to the sample text for categorization. In the last section the conclusions and future works are made.

2 ZA resolution

The ZA resolution method we develop is divided into three parts. First we use a POS tagger to produce the tagged result of an input document. Second is ZA detection that identifies occurrences of ZA within utterances by employing detection rules based on the result of partial parsing. Third is antecedent identification that identifies the antecedent of each detected ZA using rules based on the centering theory.

2.1 Partial parsing

Partial (or shallow) parsing does not deliver full syntactic analysis but is limited to parsing smaller constituents such as noun phrases or prepositional phrases (Abney, 1996; Mitkov, 2002). For example, the sentence (2) can be divided as follows:

(2) 花蓮 成爲 熱門 的 旅遊 地點。
Hualien became the popular tourist attraction.
→ [NP 花蓮] [VP 成爲] [NP 熱門 的 旅遊 地點]
[NP Hualien] [VP became] [NP the popular tourist attraction]

In our work, we use a number of simple noun phrase rules to identify the noun phrases in the output produced by AUTOTAG which is a POS tagger developed by CKIP, Academia Sinica (CKIP, 1999). For example, the result of (2) produced by AUTOTAG is as below.

(3) 花蓮 成爲 熱門 的 旅遊 地點。
→ [花蓮(Nc) 成爲(VG) 熱門(VH) 的(DE) 旅遊(VA) 地點(Na) 。]

There are about 47 POS tags used in AUTOTAG, in which 13 POS tags belong to the noun tag set and 17 POS tags belong to the verb tag set, as shown in Table 1.

Table 1: Noun and verb tag set

Set	**POS tag**
Noun	Na, Nb, Nc, Ncd, Nd, Nep, Neqa, Neqb, Nes, Neu, Nf, Ng, Nh
Verb	VA, VAC, VB, VC, VCL, VD, VE, VF, VG, VH, VHC, VI, VJ, VK, VL, V_2, V_11

Since a noun phrase consists of at least one noun and its POS is included in the noun tag set, we create simple noun phrase rules in DCG (Gazdar and Mellish, 1989) as shown below.

Noun phrase rules:
n(N)→ na; nb; nc; ncd; nd; nep; neqa; neqb; nes; neu; nf; ng; nh.
vah(V)→ va; vh.
np([N])→ n(N).
np([N1,Ns])→ n(N1), (de, np(Ns) ; np(Ns) ; [ ]).
np([V1,Ns])→ vah(V1), (de;[]), np(Ns) .

2.2 Centering Theory

In the centering theory (Groze *et al*, 1995; Walker *et al.*, 1994; Strube and Hahn, 1996), 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. The most highest ranked 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 theory specifies a set of constraints and rules (Groze *et al*, 1995; Walker *et al.* 1994).

Constraints

For each utterance U_i in a discourse segment U_1, ..., 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 U_1, ..., 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 (Groze *et al*, 1995).

Rule II reflect 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.

2.3 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 ZA resolutions is divided into two phases: first ZA detection and then antecedent identification. In this paper, we focus on the cases of ZA occurring in the topic or subject position.

In the ZA detection phase, we use simple syntactic relations to detect omitted cases as ZA candidates:

ZA detection rules:

1. For each utterance U_i in a discourse segment $U_1, \dots, U_m$: If a verb, coordinating conjunction, or preposition appears in the initial position of U_i then an omitted case is detected as a ZA candidate.
2. For each ZA candidate detected in an utterance: If an intransitive verb appears in the initial position of an utterance and the leftmost elements of the utterance are an intransitive verb and a noun in order then ZA does not occur.

Note that in ZA detection rules, we employ the rule 1 to detect omitted cases as ZA candidates, and the rule 2 is an exceptional case which a verb appears in the initial position of an utterance but the verb is an intransitive verb tagged as VA or VH.

In the antecedent identification phase, we employ the concept, 'backward-looking center' of centering theory to identify the antecedent of each ZA. First we use noun phrase rules to obtain noun phrases in each utterance, and then the antecedent is identified as the most prominent noun phrase of the preceding utterance (Yeh and Chen, 2001):

Antecedent identification rule:

1. For each utterance U_i in a discourse segment $U_1, \dots, U_m$: If a ZA occurs in U_i then choose the noun phrase in the initial position of U_{i-1} as the antecedent.

Due to topic-prominence in Chinese (Li and Thompson, 1981), topic is the most salient grammatical role. In general, if the topic is omitted, the subject will be in the initial position of an utterance. If the topic and subject are omitted concurrently, the ZA occurs. Since the antecedent identification rule is corresponding to the concept of centering theory.

We do not intend to resolve all kinds of zero anaphor in our preliminary experiment but focus on the most important cases of ZA that occur in either the topic or subject positions of utterances. For example, (5) is the result of applying ZA resolution to (4).

(4) a. 基隆醫院^i 爲 擴大 服務 範圍，
Kee-lung General Hospital aims to increase service coverage.

b. φ^i 積極 提升 醫療 服務 品質 及 標準化，
(It) actively improves the service quality of medical treatment and standardization.

c. φ^i 獲 衛生署 認可 爲 辦理 外勞體檢 醫院。
(It) is certified by Department of Health as a hospital which can handle physical examinations of foreign laborers.

(5) a. 基隆醫院 爲 擴大 服務 範圍，

Kee-lung General Hospital aims to increase service coverage.

b. *基隆醫院* 積極 提升 醫療 服務 品質 及 標準化，
Kee-lung General Hospital actively improves the service quality of medical treatment and standardization.

c. *基隆醫院* 獲 衛生署 認可 爲 辦理 外勞體檢 醫院。
Kee-lung General Hospital is certified by Department of Health as a hospital which can handle physical examinations of foreign laborers.

3 Text categorization

We implement a text categorization system based on a supervised learning algorithm, *k*-nearest neighbor (*k*-NN) algorithm. In this section, we describe the approach to term extraction, calculation of the term weights and the *k*-NN Algorithm.

3.1 Term extraction

Due to the nature of Chinese language, there is no blank between words. In the phase of term extraction, we employ Bi-gram model to extract terms from documents that belong to each class (Yang *et al.*, 1993; Chen, 2001). The process starts from the first character of the sentence and combines 2 consecutive characters to form a bi-gram. Then it goes on to the second and repeats the grouping further on until the end. Consequently, all possible overlapping bi-grams are obtained. For example, the terms extracted from sentence (2) are: [花蓮, 蓮成, 成爲, 爲熱, 熱門, 門的, 的旅, 旅遊, 遊地, 地點].

3.2 *k*-NN Algorithm

As an instance-based classification method, *k*-NN has been known as an effective approach to a broad range of pattern recognition and text classification problems (Yang *et al.*, 2002; Ko and Seo, 2002). In *k*-NN algorithm, a new input instance should belong to the same class as their k nearest neighbors in the training data set. After all the training data is stored in memory, a new input instance is classified with the class of *k* nearest neighbors among all stored training instances.

For the distance measure and the document representation, we uses the conventional vector space model, which represents each document as a vector of term weights, and the distance between two documents is measured using the cosine value of the angle between the corresponding vectors. We compute the weight vectors for each document using one of the conventional TF-IDF schemes. The weight of term *t* in document *d*, *w*(*t*,*d*), is calculated as frequency and inverse document frequency (TF-IDF) value by formulas (1) and (2).

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

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

where

i) $tf_{t,d}$ is the within-document Term Frequency (TF).

ii) N is the number of all training document s.

iii) df_t is the number of training documents in which t occurs.

Given a test document *d*, the *k*-NN classifier assigns a relevance score to each candidate category c_j using the following formula (3):

$$s(c_j, d) = \sum_{d' \in R_k(d) \cap D_j} \cos(wd, wd') \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

where $R_k(d)$ denotes a set of the k nearest neighbors of document d and D_j is a set of training documents in class c_j.

4 Experiment and Result

We collect 300 news articles pre-classified into 8 categories as a test corpus which contains about 132 thousands Chinese character for the experiments of ZA resolution and text categorization.

In the ZA detection rules mentioned in section 3.3, zero anaphors may occurs in three cases: a verb, a coordinating conjunction, or a preposition appearing in the initial position of an utterance. By applying the ZA detection rules and antecedent identification rule to half the test corpus, which is a collection of 150 news articles contained 998 paragraphs, 4631 utterances, and 77537 Chinese characters, the recall rates and precision rates of ZA resolution of these three cases are shown in Table 2. The main errors of ZA resolution occur in the experiment when parsing inverted sentences and non-anaphoric cases (e.g. exophora[3], cataphora[4]) (Hu, 1995; Mitkov, 2002). Cataphora is similar to anaphora, the difference being the direction of the reference. That is, we cannot detect the referent of a ZA in the following utterances, and we do not treat cataphora in this paper.

Table 2: Results of ZA resolution

Cases Accuracy	Verb	Coordinating conjunction	Preposition	Total
Recall Rate	67.4%	80%	67.3%	67.4%
Precision Rate	64.2%	66.7%	53.9%	61.8%

In the experiment of text categorization, half of 300 news articles are taken as the training data set and the other half as the test data set. Without applying ZA resolution on test data set, the accuracy of categorization is 79%.

To improve the accuracy of text categorization, we integrate the ZA resolution method mentioned previously into the text categorization system. The new text categorization system works as below: First an input document with ZA resolved is taken as a ZA-resolved input document which each zero anaphor in the text is replaced by its antecedent. Second the ZA-resolved input document is categorized by the k-NN classifier. The result shows the new text categorization system increases the accuracy from 79% to 84%.

5 Conclusions

In this paper, we develop an inexpensive method of Chinese ZA resolution that work on the output of a part-of-speech tagger and use a partial parsing instead of a complex parsing to resolve zero anaphors in Chinese text. Then the resulting text is used as the input of a text categorization system. In our preliminary experiment, we deal with the cases of topic or subject omission. The result shows that ZA resolution method enhances the accuracy of text categorization from 79% to 84%.

We have show that the result of employing information carried by ZAs in text to contribute the calculation of text categorization is promising to some extent. We will further extend our approach to dealing with other omission cases, such as indirect object omission, and apply the results to text categorization systems with other classification methods in the future.

Acknowledgements

We give our special thanks to CKIP, Academia Sinica for making great efforts in computational linguistics and sharing the Autotag program to academic research.

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

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

References

Steven Abney. 1996. 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.

Chinatsu Aone and Scott William Bennett. 1995. Evaluating automated and manual acquisition of anaphora res-olution strategies. In *Proceedings of the 33rd Annual Meeting of the ACL*, Santa Cruz, New Mexico, pages 122–129.

Baldwin B. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. *ACL/EACL workshop on Operational factors in practical, robust anaphor resolution.*

Hongbiao Chen. 2001. *Looking for Better Chinese Indexes: A Corpus-based Approach to Base NP Detection and Indexing*, Ph.D. thesis, Guangdong University of Foreign Studies.

CKIP. 1999. 中文自動斷詞系統 Version 1.0 (Autotag), *http://godel.iis.sinica.edu.tw/CKIP/*, Academia Sinica.

Connoly, Dennis, John D. Burger & David S. Day. 1994. A Machine learning approach to anaphoric reference. *Proceedings of the International Conference on New Methods in Language Processing*, 255-261, Manchester, United Kingdom.

A. Ferrández, Manuel Palomar, Lidia Moreno. 1998. Anaphor Resolution in Unrestricted Texts with Partial Parsing. *Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pages 385-391. Montreal, Canada.

G. Gazdar and C. Mellish. 1989. *Natural Language Processing in PROLOG – An Introduction to Computational Linguistics*, Addison- Wesley.

Niyu Ge, John Hale and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161 –170

B. J. Grosz and C. L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics,* No 3 Vol 12, pp. 175-204.

B. J. Grosz, A. K. Joshi and S. Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics,* 21(2), pp. 203-225.

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

Thorsten Joachims. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 143-151, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US.

Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113-118. Copenhagen, Denmark.

Youngjoong Ko and Jungyun Seo. 2000. Automatic Text Categorization by Unsupervised Learning. *Proceedings of COLING-00, the 18th International Conference on Computational Linguistics.*

Youngjoong Ko and Jungyun Seo. 2002. Text Categorization using Feature Projections. *Proceedings of COLING-2002, the 19th International Conference on Computational Linguistics.*

Lappin S. and Leass H. 1994. An algorithm for pronominal anaphor resolution. *Computational Linguistics*, 20(4).

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

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

Mitkov, Ruslan. 1999. Anaphora resolution: the state of the art. Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution). University of Wolverhampton, Wolverhampton.

Mitkov, Ruslan. 2002. *Anaphora Resolution*, Longman.

Okumura, Manabu and Kouji Tamura. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, 871-876.

Robert E. Schapire, Yoram Singer, and Amir Singhal. 1998. Boosting and rocchio applied to text filtering. *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information.*

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution. *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp.911-917.

C. L. Sidner. 1979. *Toward a Computational Theory of Definite Anaphora Comprehension in English Discourse*, Ph.D. thesis, MIT.

C. L. Sider. 1983. Focusing in the comprehension of definite anaphora. *Computational Models of Discourse*, MIT Press.

S. Soderland. 1996. CRYSTAL: Learning Domain-specific Text Analysis Rules. *CIIR Technical Report*, Nov.

Strube, M. and U. Hahn. 1996. *Functional Centering. Proc. Of ACL '96*, Santa Cruz, Ca., pp.270-277.

Roland Stuckardt. 2002. Machine-Learning-Based vs. Manually Designed Approaches to Anaphor Resolution: the Best of Two Worlds. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2002)*, University of Lisbon, Portugal, pages 211-216.

Jyh-Jong Tsay and Jing-Doo Wang. 2000. Design and Evaluation of Approaches to Automatic Chinese Text Categorization. *Computational Linguistics and Chinese Language Processing (CLCLP)*, 5(2): 43-58.

Walker, M. A. 1989. Evaluating Discourse Processing Algorithms. *Proc. Of ACL '89*, Vancouver, Canada.

Walker, M. A., M. Iida and S. Cote. 1994. Japan Discourse and the Process of Centering. *Computational Linguistics*, 20(2): 193-233.

Walker, M. A. 1998. Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, Centering in Discourse. Oxford University Press.

R.Z. Wang and J.S. Chang. 2001. 適應性文件分類系統. *Proc. Of ROCLING XIV*, Cheng Kung University, Tainan, Taiwan.

Yang Y., Slattery S., and Ghani R. 2002. A study of approaches to hypertext categorization, *Journal of Intelligent Information Systems, Volume 18, Number 2.*

Yun-Yen Yang, Keh-Jiann Chen, Ching-Chun Hsieh, and Shu-Mei Chen. 1993. A Study of Document Auto-Classification in Mandarin Chinese. In *Proceedings of ROCLING VI*, Hsinchu, Taiwan.

Ching-Long Yeh. 1995. *Generation of Anaphors in Chinese*, Ph.D. thesis, University of Edinburgh.

Ching-Long Yeh and Yi-Chun Chen. 2001. An empirical study of zero anaphora resolution in Chinese based on centering theory. In *Proceedings of ROCLING* XIV, Tainan, Taiwan.