Expert Systems with Applications 36 (2009) 7544–7551

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Zero anaphora resolution by case-based reasoning and pattern conceptualization

Dian-Song Wu, Tyne Liang *

Department of Computer Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan, ROC

ARTICLE INFO

Keywords:
Zero anaphora resolution
Case-based reasoning
Conceptual patterns
Knowledge resources

ABSTRACT

Effective anaphora resolution is helpful to many applications of natural language processing such as machine translation, summarization and question answering. In this paper, a novel resolution approach is proposed to tackle zero anaphora, which is the most frequent type of anaphora shown in Chinese texts. Unlike most of the previous approaches relying on hand-coded rules, our resolution is mainly constructed by employing case-based reasoning and pattern conceptualization. Moreover, the resolution is incorporated with the mechanisms to identify cataphora and non-antecedent instances so as to enhance the resolution performance. Compared to a general rule-based approach, the proposed approach indeed improves the resolution performance by achieves 78% recall and 79% precision on solving 1051 zero anaphora instances in 382 narrative texts.

1. Introduction

Anaphora resolution denotes the process of determining the antecedent of an anaphor (Lee, 2002; Mitkov, 1998; Xu, 2003). Among different types of antecedents, noun phrases are the most common ones representing the objects referred by an anaphor which can be a pronoun, a definite noun phrase, a quantifier or even unspecified in discourses. In recent literature, most of anaphora resolution approaches are presented to tackle pronominal anaphora, which is the most frequent one appearing in English texts (Hobbs, 1976; Kennedy & Boguraev, 1996; Lappin & Leass, 1994; Mitkov, 1999; Yang, Su, & Tan, 2006). The approaches are implemented by measuring the syntactic and semantic agreement between an anaphor and its antecedent candidates with or without the help of some outer resources like WordNet or the Web (Liang & Wu, 2004; Markert & Nissim, 2005; Mitkov, Richard, & Orasan, 2002; Modjeska, Markert, & Nissim, 2003).

Contrast to pronominal anaphora in English texts, zero anaphora is the major anaphora occurring in Chinese texts (Xu, 2003). It means that most of the anaphors appearing in Chinese texts can be unspecified if they are inferable from the contexts. The omitted grammatical constituent is called a zero anaphor (ZA). Zero anaphors may occur in a single sentence or in consecutive sentences. Essentially, the recovery of zero anaphors relies on contextual information, semantic inference, and world knowledge (Lee, 2002; Tao & Healy, 2005). However, efficient Chinese ZA resolution has not been widely addressed. Hence, an effective ZA resolution is presented in this paper with the aim to facilitate Chinese message understanding.

Basically, the challenges to resolving ZA in Chinese texts are difficulties of constructing proper reasoning mechanisms and insufficiency of lexical features useful for resolution. Recently, Yeh and Chen (2007) presented ZA resolution with partial parsing based on centering theory and obtained 66% *F*-score in 150 news articles. On the other hand, Converse (2005) applied full parsing results but obtained unsatisfactory ZA resolution since only few features were used by the Hobbs algorithm which is originally designed for resolving English anaphora. Zhao and Ng (2007) presented a decision tree classification approach to Chinese anaphoric zero pronouns resolution and obtained 43% *F*-score in 205 texts. The disadvantage of classification approach is that a binary classifier can only distinguish between two classes. During antecedent identification, a classifier will essentially ignore the remainder antecedent candidates whenever one candidate is tagged to be positive.

In this paper, a novel ZA resolution approach is proposed by applying case-based reasoning (CBR) and pattern conceptualization. This is because CBR is able to exploit the previous experience that might be useful for the novel problem. In this paper, we utilize the antecedent features of the retrieved cases to predict the antecedent of a novel case. As all cases are represented with the patterns containing semantic tags for their nouns and grammatical tags for the verbs, such pattern conceptualization will be able to efficiently reduce data sparseness in the case base. Moreover, the presented resolution is incorporated with a filtering mechanism

* Corresponding author. Tel.: +886 3 5712121x31365.
E-mail addresses: diansongwu@cs.nctu.edu.tw (D.-S. Wu), tliang@cs.nctu.edu.tw (T. Liang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.09.065

to identify those non-anaphoric cases such as cataphora and non-antecedent instances in order to enhance the overall resolution performance. The experimental results show that our proposed approach achieved competitive resolution by yielding 79% *F*-score on 1051 ZA instances and yielded 13% improvement while compared to the general rule-based approach presented by Yeh and Chen (2007).

The remainder of this paper is organized as follows: Section 2 introduces the commonly seen zero anaphora instances in Chinese texts. Section 3 describes the resolution approach by using CBR-based learning. Section 4 describes the procedure of zero anaphora resolution and the experimental results. Section 5 presents the final conclusions.

2. Chinese zero anaphora

According to Huang (2000) and Li (2004), a Chinese sentence is generally integrated by complete syntactic components and expresses an intact meaning. It is composed of one or more clauses and is explicitly identified with punctuation marks like "○, !, ?". A Chinese clause is an utterance which is identified with punctuation marks like "„ ;, :, ○, ?" and grammatically it may or may not be a complete syntactic component. As mentioned above, ZA is the most common anaphora displaying in Chinese texts and it can be intra-sentential when a ZA appears in a single-clause sentence or inter-sentential when it appears in multiple-clause sentences. In the following examples, we list some typical ZA and use "ϕ" to denote zero anaphors which may play as subject or object roles in Chinese sentences and their referents are noun phrases.

(A) Inter-sentential ZA:

1. *Subject-role case:* The subject (like "Xiaoming" in the example) appears overtly once in the first clause, but later mentions of the same subject are left unspecified in a multiple-clause sentence.

(ex. 1) 小明$_1$ 打開 在 地上 的 箱子 ϕ_1 拿出 兩 本 故事書 後, ϕ_1 回到 自己 的 房間。
(Xiaoming opened the box on the ground. (Xiaoming) took out two storybooks. (Xiaoming) went back to his room.)

2. *Object-role case:* The object (like "new album" in the example) is unspecified in the second clause if it can be understood or inferred from the first clause in a multiple-clause sentence.

(ex. 2) 張三 買 了 新 唱片$_2$, 許多 朋友 都 向 他 借 ϕ_2。
(Zhangsan bought a new album. Many of his friends borrowed (a new album) from him.)

(B) Intra-sentential ZA:

3. *Subject-role case:* The same subject (like "Lisi" in the example) is unspecified if it is shared from the previous verb in a single-clause sentence with one more verbs.

(ex. 3) 李四$_3$ 參加 演講 比賽 ϕ_3 贏得 冠軍。
(Lisi participated in a lecture contest and (Lisi) won the first honor.)

4. *Object-to-subject case:* The subject (like "Wangwu" in the example) of the second verb is unspecified if it is the object of the first verb in a pivotal sentence.

(ex. 4) 李四 允許 王五 ϕ_4 再 重做 一 份 報告。
(Lisi allowed Wangwu (and Wangwu) redo a report again.)

As mentioned previously, a Chinese sentence expresses one complete meaning. However, it is usually observed that a sentence might be incorrectly segmented into a sequence of clauses with punctuations like "," and some of them are just a noun phrase or a prepositional phrase as shown in the following examples (ex. 5 and ex. 6). So it is required for a ZA resolver to identify such kind of anaphoric relations in the adjacent clauses for a multiple-clause sentence.

Table 1
The positional distribution of anaphor–antecedent pairs

Relative position[a]	(1)	(2)	(3)	(4)
Number of pairs	710	57	22	4
Ratio (%)	89.5	7.2	2.8	0.5

[a] Relative position: (1) pairs are in the same complex sentence; (2) pairs are in two complex sentences; (3) pairs are in the same paragraph; (4) pairs are not in the same paragraph.

(ex. 5) 總理 斯洛德$_5$, ϕ_5 宣布 德國 將 舉行 議會 選舉。
(Premier Schroeder, (Premier Schroeder) declared that Germany will hold a council election.)

(ex. 6) 人 的 生活 空間$_6$, ϕ_6 和 自然 環境 發生 了 對立。
(Human living space, (human living space) and environment brought about conflict.)

As mentioned above, the antecedent of a zero anaphor occurs in the previous expressions. However, there are also cases that antecedents are not specified in the previous context, called non-anaphoric zero anaphora (as shown in example (ex. 7)). Therefore, effective zero anaphora resolution relies on not only the identification of antecedents but also the elimination of non-anaphoric cases.

Non-anaphoric zero anaphora case: In this example, ϕ_7 refers to "time" but the antecedent "time" is not specified previously.

(ex. 7) ϕ_7 過 了 兩 天 , 警察 找到 了 犯罪 的 證據。

(After two days, the police found the criminal evidence.)

Table 1 lists the positional distribution of 793 anaphor–antecedent pairs in our training data and it shows that 96.7% of antecedents are in a distance of two sentences.

3. The approach

Fig. 1 illustrates the proposed ZA resolution at the training and testing phases. At training phase, the kernel case-based reasoning module is built in three major steps, namely, feature extraction, pattern conceptualization, and feature weight learning. As a result, a case base, which contains both anaphoric and non-anaphoric ZA cases, is constructed for case retrieval at testing phase. At the testing phase, an input text is processed by a pipeline of text preprocessing, zero anaphor detection, and antecedent (ANT) identification. Moreover, a weighted *k*-nearest-neighbor (WKNN) algorithm is presented to measure the similarities of cases at case retrieval. The antecedent features of the retrieved cases are applied for antecedent selection. The following subsections describe each component and the resolution procedure in detail.

3.1. CBR approach

CBR is an incremental learning technique that has been successfully used for building knowledge systems and aiding knowledge acquisition (Aamodt & Plaza, 1994; Cardie, 1999; Liu & Ke, 2007). The main concept of CBR is to exploit the previous experience that might be useful for the novel problem. In this paper, we utilize the antecedent features of the retrieved cases to predict the antecedent of a novel case. In the case base, those anaphoric cases (treated as positive cases) will be encoded with more features than the non-anaphoric cases (treated as negative cases) and all the cases will

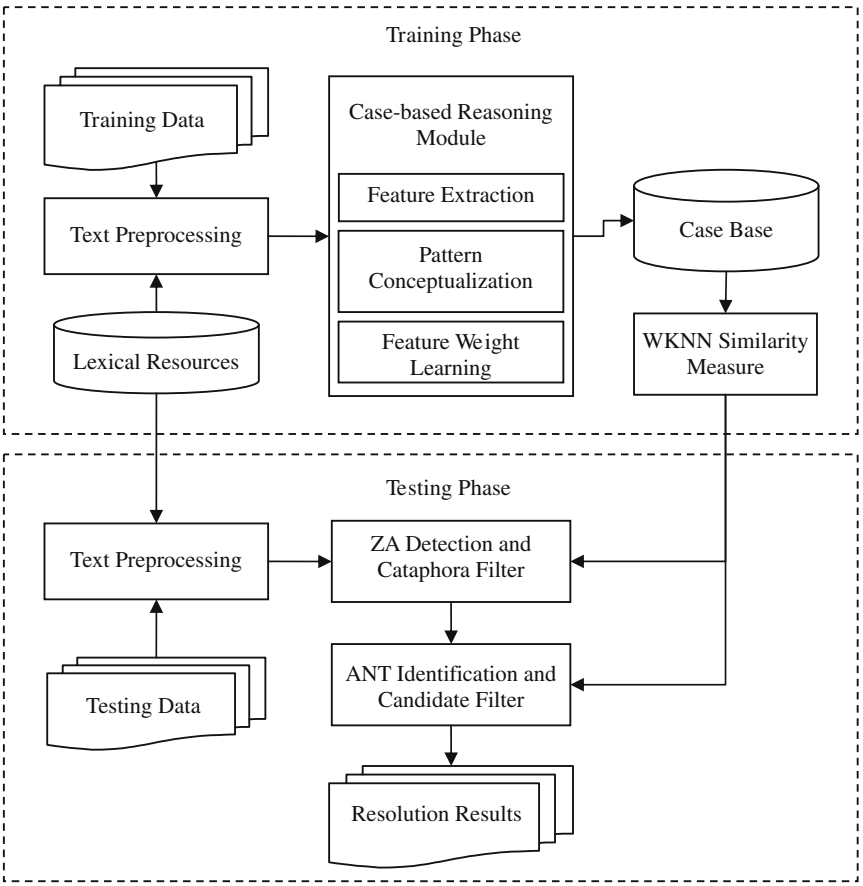

Fig. 1. The presented Chinese zero anaphora resolution procedure.

be transformed into conceptual patterns. By measuring the similarity between the novel case and the stored cases, we can check whether a give sentence contains a ZA or not. The most similar case will be reused for antecedent selection if it is a positive case.

For instance, an omission occurs before the verb "宣布" (announce) in the following example (ex. 8). A positive case is extracted from case base, as shown in example (ex. 9), to infer the corresponding antecedent.

(ex. 8) 議員(Na)[1]討論(VE)細節(Na)後(Ng),
ϕ 宣布(VE) 明年(Nd) 將(D)舉行(VC)大選 (Na)。
(After discussing the details, (Councilor) announced that there will be an election next year.)

(ex. 9) 主席(Na)整理(VC)意見(Na)後(Ng),
ϕ 決定(VE)明天(Nd)表決(VE)修正案(Na)。
(After collecting opinions, (Chairman) decided that the amendment will be decided by vote tomorrow.)

3.2. The corpus and its preprocessing

The ASBC corpus (CKIP, 1996) is used in the training phase with 46 kinds of POS tags. Each text of the corpus is segmented into sentences, and each word is tagged with its part-of-speech. For noun phrase chunking, we built up a finite state machine chunker to chunk noun phrases which will be treated as antecedent candidates. In Chinese, the head noun occurs at the end of a noun phrase. Therefore, in a noun phrase, words preceding the head noun are regarded as modifiers. The head noun is assigned with feature values such as gender or animate, since it dominates the fundamental property of the noun phrase. There are five types of head nouns defined in Yu and Chen (2000); they are common nouns, proper nouns, location nouns, temporal nouns, and pronouns. Several examples of noun phrases recognized by the presented chunker are as follows:

(ex. 10) 每(Nes)位(Nf)用戶 (Na)的(DE)個人(Na) 資料(Na)
(the individual information of each subscriber)
(ex. 11) 委員會(Nc)主席(Na)劉生明(Nb)
(committee chairman Liushengming)
(ex. 12) 相當(Dfa)有名(VH)的(DE)公園(Nc)
(a very famous park)
(ex. 13) 十月(Nd)六日(Nd)早上(Nd)
(the morning of October 6)

The presented chunker is also able to recognize verbal nominalization and transformation by utilizing heuristics discussed in Ding, Huang, and Huang (2005). These cases are handled by the following heuristics:

1. If the preceding word of the verb is tagged with DE, then the verb is treated as a noun during the chunking phase.
2. If the verb is followed by a word tagged with DE, then the verb is regarded as a modifier of a noun phrase.
3. If the verb is followed by the word "地", then the verb is treated as an adverb.

[1] Each word of the sentence is followed by its part-of-speech tag. A detail description of part-of speech tag set used in this paper is available at http://ckipsvr.iis.sinica.edu.tw/category_list.doc.

Table 2
Semantic classes selected from CKIP lexicon

Entity	Semantic classes
Physical	Mankind, places, artifacts, and matter
Nonphysical	Events, temporal, principles, and mental

Table 3
Case representation in the case base

Content of a case in the case base	主席(Na)整理(VC)意見(Na)後(Ng), ϕ 決定(VE)明天(Nd)表決(VE)修正案(Na)。 (After collecting opinions, (Chairman) decided that the amendment will be decided by vote tomorrow)
Implementation level (ZA template)	PRE_NPS: N ROLE: subject POS: VE FIRST_VERB: Y CLASS_VERB: report SEN_DIST: 2 PRE_VERB: Y PRE_PREP: N PRE_CONJ: N PRE_ZA: N CON_PAT: ϕ (VE)[temporal](VE)[events]
Implementation level (ANT template)	TOPIC: Y ROLE: subject NUM: singular GND: neutral POS_HEAD: Na NE: N DEF: N EMB: N CLASS: mankind SEN_DIST: 1 OFFSET: 1 CUR_ZA: N RPT: N CON_PAT: [mankind](VC)[mental]

3.3. Outer lexical resources

Two outer resources are used to acquire informative features such as semantic classes of nouns and verbs during ZA resolution. The resources used are CKIP lexicon (CKIP, 1995) and the Academia Sinica Bilingual WordNet (SinicaBOW).[2] There are four kinds of verbs regarded as animate verbs; namely, {cognition}, {communication}, {emotion}, and {social}. CKIP lexicon contains 80,000 entries annotated with syntactic categories and corresponding semantic classes. There are eight semantic classes selected from CKIP lexicon. During processing of noun phrases, head nouns of noun phrases are tagged with semantic classes. The classes can be divided into physical entities and nonphysical entities as listed in Table 2.

3.4. Feature extraction

A case for example (ex. 9) in the case base is represented in the form as shown in Table 3. It contains both ZA template and ANT template used as a ZA resolution method. During the training phase, the case base contains examples collected from the training corpus and annotated with ZA markers (denoted as "ϕ") by human experts. Table 4 shows an input test case in which ϕ occurs before the verb "宣布" (announce). A detail description of features for ZA template and ANT template is shown in Table 5.

Table 4
Input case representation

Content of an input case	議員(Na)討論(VE)細節(Na)後(Ng), ϕ 宣布(VE)明年(Nd)將(D)舉行(VC)大選(Na)。 (After discussing the details, (Councilor) announced that there will be an election next year)
Implementation level (ZA template)	PRE_NPS: N ROLE: subject POS: VE FIRST_VERB: Y CLASS_VERB: report SEN_DIST: 2 PRE_VERB: Y PRE_PREP: N PRE_CONJ: N PRE_ZA: N CON_PAT: ϕ (VE)[temporal](VC)[events]

Table 5
Description of template features

	Feature	Description
ZA template	PRE_NPS	If the preceding clause is a noun phrase then Y; else N
	ROLE	Grammatical role of the ZA: subject, object, or other
	POS	Part-of-speech of the related verb
	FIRST_VERB	If the related verb is the first one then Y; else N
	CLASS_VERB	Semantic class of the related verb
	SEN_DIST	The ZA occurs in the *i*th clause of a complex sentence
	PRE_VERB	If the ZA is followed by a verb then Y; else N
	PRE_PREP	If the ZA is followed by a preposition then Y; else N
	PRE_CONJ	If the ZA is followed by a conjunction then Y; else N
	PRE_ZA	If a ZA occurs in the preceding clause then Y; else N
	CON_PAT	The conceptual pattern of a sentence in which a ZA occurs
ANT template	TOPIC	If the ANT is the first noun phrase of a complex sentence then Y; else N
	ROLE	Grammatical role of the ANT: subject, object, or other
	NUM	Single, plural, or unknown
	GND	Male, female, neutral, or unknown
	POS_HEAD	Part-of-speech of the ANT head noun
	NE	The ANT is a person name or an organization name
	DEF	If the ANT is a definite noun phrase then Y; else N
	EMB	If the ANT is a embedded noun phrase then Y; else N
	CLASS	Semantic class of the ANT
	SEN_DIST	The ANT occurs in the *i*th clause of a sentence
	OFFSET	Distance between the ANT and the ZA in terms of clauses
	CUR_ZA	If a ZA occurs in the current clause then Y; else N
	RPT	If the ANT repeats more than once then Y; else N
	CON_PAT	The conceptual pattern of a sentence in which the ANT occurs

During the antecedent identification phase, features are assigned to the target candidate based on syntactic, semantic, contextual, and positional properties. However, information regarding number or gender of noun phrases cannot be obtained during text preprocessing. We design two procedures to identify these two features as described below:

1. Number identification procedure:

Step 1: We define symbols as follows:
NP = noun phrase;

[2] SinicaBOW is a Mandarin–English bilingual database based on the framework of English WordNet and language usage in Taiwan. A detailed description is available at http://bow.sinica.edu.tw/.

HNP = head noun of the noun phrase;
Q = the set of quantifiers;
P = the set of collective quantifiers such as {{群, 夥, 堆, 對, 批}};
R = {都, 全, 全部, 全體, 皆, 所有,
每個, 雙方, 多數, 一些, 某些,;
若干,幾個, 數個, 許多, 諸多}

Step 2: If NP satisfies any of the following conditions, then return singular.
i. HNP is a person name;
ii. NP contains a title;
iii. NP $\in$ {[[這|那|該|某|一]+{$Q - P$} + noun};

Step 3: Else if NP satisfies any of the following conditions, then return plural.
i. HNP is an organization name;
ii. The last character of NP $\in$ {們, 倆};
iii. NP contains plural numbers + Q;
iv. NP follows r, where $r \in R$;

Step 4: For other cases, the number feature is marked unknown.

2. Gender identification procedure:

Step 1: We define symbols as follows:
NP = noun phrase;
F = the set of female titles such as "太太", "女友";
M = the set of male titles such as "先生", "男友";
C = the set of common characters for female names;

Step 2: If semantic tag of NP $\notin$ mankind, then return neutral;

Step 3: Else if NP satisfies any of the following conditions; then return male;
i. NP = person name + 先生;
ii. the first character of NP is "男";
iii. the last character of NP is "父";
iv. NP + 的 + f, where $f \in F$;
v. NP + 他;

Step 4: Else if NP satisfies any of the following conditions; then return female;
i. NP = person name + 女士;
ii. the first character of NP is "女";
iii. the last character of NP is "母";
iv. NP + 的 + m, where $m \in M$;
v. NP + 她;
vi. any character of the first name $\in C$;

Step 5: For other cases, the gender feature is marked unknown.

3.5. Pattern conceptualization

The ZA template contains ten features as well as its conceptual pattern. Conceptual patterns are utilized to measure the similarity between sentences in which zero anaphors occur. Each sentence is expressed as a pattern composed of semantic classes of nouns and grammatical categories of verbs. In the following examples, sentences (ex. 14) and (ex. 15), the corresponding conceptual patterns are represented. Each field bracketed by [ ] or ( ) indicates an item in the conceptual pattern.

(ex. 14) ϕ 宣布(VE)明年(Nd)將(D)舉行(VC) 大選(Na)。
((Councilor) announced that there will be an election next year.)
Concept pattern representation: ϕ (VE) [temporal] (VC) [events]

(ex. 15) ϕ 表決(VE)明天(Nd)表決(VE)修正案(Na)。
((Chairman) decided that the amendment will be decided by vote tomorrow.)
Conceptual pattern representation: ϕ (VE) [temporal] (VE) [events]

Similarity between the input test sentence with a ZA and the case sentence with a ZA in the case base is described in Eq. (1). For a given input test sentence I and a case sentence C, *CPSIM(I,C)* calculates the similarity value of sentences in which zero anaphors occur. For example, the similarity value of examples (ex. 14) and (ex. 15) is given as $(2 \times 4)/(5 + 5) = 0.8$,

$$CPSIM(I,C) = \frac{2 \times LENLCS(I,C)}{(LEN(I) + LEN(C))} \tag{1}$$

where I: the input test sentence with a ZA; C: the case sentence with a ZA in case base; $LENLCS(I,C)$: number of items in the longest common subsequence of I and C; $LEN(I)$: number of items in I; $LEN(C)$: number of items in C.

3.6. Feature weight learning

Eq. (3) is the similarity function used to compute the similarity between the input case and the stored case examples. The similarity computation concerns the similarity between ZA template features and conceptual patterns as described above. Subsequently, the ANT template with the highest similarity value is retrieved from the case base and used to identify the antecedent with respect to a given test case. For instance, to identify the antecedent of ϕ as shown in Table 4, the most similar case, shown in Table 3, is extracted by Eq. (3). According to the ANT template in Table 3, "議員" (councilor) is selected as the antecedent because it matches the most features than other candidates such as "細節,'" (details).

We conduct a weighted k-nearest-neighbor algorithm in the case retrieval phase. The case retrieval phase captures the most similar case in the case base and employs the antecedent features to resolve the test case. The process is shown as follows:

1. Calculate the weight w_{f_i} of each feature f_i by the following equation:

$$w_{f_i} = \inf(S) - \inf_{f_i}(S) \tag{2}$$

$$\inf(S) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$\inf_{f_i}(S) = \sum_{j=1}^{v}\frac{|S_j|}{|S|} \times \inf(S_j)$$

where f_i: the ith feature with v distinct values; S: the set of cases in the case base; S_j: the subset of S for which feature f_i has value j; p: the number of cases belonging to positive ones; n: the number of cases belonging to negative ones.

2. Calculate the similarity $SIM(I,C)$ of the test case I and each case C in the case base by the following equation:

$$SIM(I,C) = \frac{\sum_{i=1}^{|f|} w_{f_i} \times match(I_{f_i}, C_{f_i}) \times \alpha}{\sum_{i=1}^{|f|} w_{f_i}} + CPSIM(I,C) \times \beta \tag{3}$$

where $|f|$: the number of test case features f; w_{f_i}: the weight of the ith feature in f; I_{f_i}: the value of feature f_i of the test case; C_{f_i}: the value of feature f_i of the case in the case base; $match(I_{f_i}, C_{f_i})$: returns 1 if feature value of I_{f_i} and C_{f_i} are equal; otherwise returns 0; $CPSIM(I,C)$: conceptual pattern similarity as shown in Eq. (1); α, β: weighting factors where $\alpha + \beta = 1$.

3. Retrieve k cases with the highest similarity value.
4. Let k retrieved cases vote on the antecedent features as a solution for the test case.

4. Zero anaphora resolution

Zero anaphora resolution is decomposed into two subtasks, namely zero anaphor detection and antecedent identification. Heuristics are incorporated to identify cataphora and non-antecedent instances so as to enhance the resolution performance.

4.1. ZA detection

During the ZA detection phase, verbs in sentences are examined sequentially. If there is any omission of subjects or objects with respect to a verb, ZA detection will submit the sentence to reasoning module to decide whether there is a ZA. If there is any positive case retrieved, the ANT identification phase will be performed using the resolution template returned from the case base. If the retrieved case belongs to negative one, then the case is regarded as a non-anaphoric instance.

We must be mindful of the cataphora cases that may be mistakenly treated as a ZA. So we observe the following properties which can be utilized by our cataphora filter:

1. It often occurs after verbs tagged with VE.
2. There is no patient after the verb.
3. It occurs frequently in the first clause of a complex sentence.
4. The related verbs are followed by punctuation marks like ", " and ":".
5. It often refers to the succeeding description rather than noun phrases.

The cataphora filter algorithm is shown as follows:

Step 1: For a ZA candidate, we define symbols as follows:
V = the verb preceding the ZA candidate;
VE = the set of reporting verbs;
W = the set of any words;
$M = W -$ {nouns and verbs};
Step 2: If $V \in$ {VE} and all the following conditions are satisfied, then return cataphora;
i. sentence pattern = [W^*VM+, |W^*VM+:];
ii. a ZA candidate occurs in the first clause of a complex sentence.
Step 3: For other cases, return ZA.

Moreover, it must be noted that the following conditions will not be considered while detecting ZAs around verbs (Liu, Pan, & Gu, 2002). The conditions are described as follows:

(ex. 16) "把" (Ba) sentence:
張三(Nb)已經(D)把(P)工作(Na)完成(VC)。
(Zhangsan has made the work done.)
(ex. 17) "被" (Bei) sentence:
工作(Na)已經(D)被(P)張三(Nb)完成(VC)。
(The work has been finished by Zhangsan.)
(ex. 18) In an adverbial case: when the verb functions as a part of an adverb as described in Section 3.2, it is not the verb related to a ZA.

4.2. Antecedent identification

During the ANT identification phase, we select the most likely antecedent by applying the ANT template returned by the case-based reasoning module described in Section 3.6. Furthermore, the following heuristics are applied to filter out candidates with respect to a corresponding zero anaphor. CAN denotes an item in the candidate set preceding the ZA. If CAN satisfies any of the following patterns, it is regarded as a non-antecedent instance:

1. Conjunction pattern: ZA[c]CAN or CAN[c]ZA
$c \in$ {跟, 和, 與, 同, 及, 向, 對, 面對, 或, 或是, 或者, 亦或, 以及, 還 是, 還有}
2. Verb pattern: ZA[Vt]CAN or CAN[Vt] ZAVt denotes a transitive verb in a sentence.
3. Preposition pattern: ZA[p]CAN or CAN[p]ZA
$p \in$ {在, 對, 到, 朝, 給, 向, 比;}

4.3. The resolution comparison and analysis

Our resolution is justified by 382 narrative report articles selected from ASBC corpus (CKIP, 1996). In experimental evaluation, fivefold cross-validation was conducted over the selected data set. The positive and negative zero anaphora cases were annotated

Table 6
Statistical information of evaluation data

	Data set
Articles	382
Sentences	12,775
Words	126,119
Zero anaphors	5255

Table 7
Performance at various thresholds

Threshold α	0.1	0.2	0.3	0.4	0.5	0.6	**0.7**	0.8	0.9	1.0
Recall	0.42	0.48	0.55	0.59	0.68	0.74	0.78	0.75	0.71	0.66
Precision	0.41	0.47	0.49	0.61	0.69	0.75	0.79	0.78	0.72	0.67

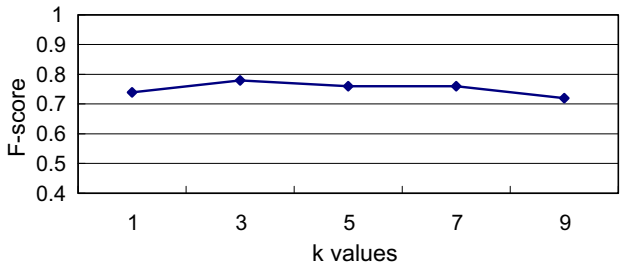

Fig. 2. *F*-score over different k values.

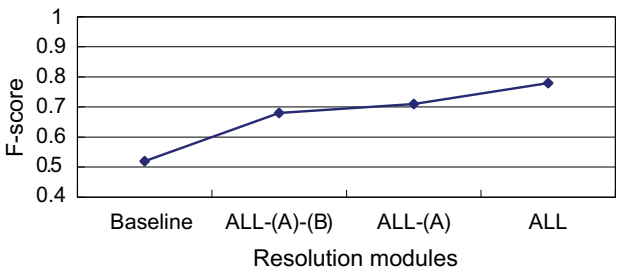

Fig. 3. *F*-score after applying resolution modules.

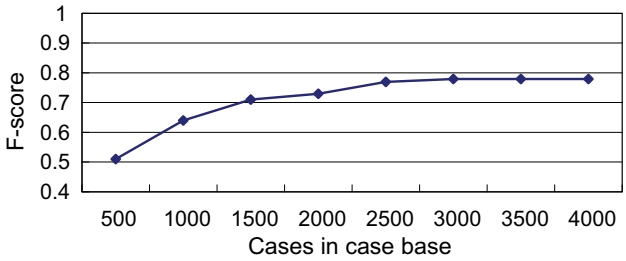

Fig. 4. *F*-score over different case base scale.

Table 8
Performance evaluation with different methods

Method	*F*-score (%)
CT	66
CT + VN	69
CT + VN + CF	71
Our method	79

Table 9
Error analysis

Error types	Ratio (%)
POS tagging/chunking error	20
Semantic class mismatch	16
Inappropriate ANT template	14
Exceeding window size	13
Number mismatch	11
Gender mismatch	10
Multiple antecedents	9
Others	7
Total	100

manually by domain experts. There are 5255 zero anaphors in total, which contains 3217 anaphoric cases and 2038 non-anaphoric cases respectively. During the testing phase, CKIP Chinese word segmentation system[3] is utilized for tagging POS. Table 6 lists the statistical data regarding both the training and the testing corpora. Table 7 lists the results in terms of precision and recall at various matching thresholds. It is observed that optimal performance (in terms of *F*-score) is achieved when the α and β values are 0.7 and 0.3, respectively. Moreover, we employ the presented weighted *k*-nearest-neighbor algorithm during resolution. According the result shown in Fig. 2, the value of *k* is set to be 3 in our experiments. In order to verify the impact of the extracted features, a baseline model is built in such a way that only grammatical features are used in ANT identification. Fig. 3 shows that the highest *F*-score is obtained when all the ZA template features and conceptual patterns (denoted as "ALL") are concerned and the baseline yields the worst performance by comparison. Additionally, the resolution performance can be enhanced significantly by applying semantic class features (denoted as "A") and conceptual pattern mapping (denoted as "B"). We verify the sensitivity of training case size in our presented CBR approach for resolving zero anaphora. It is found from Fig. 4 that feasible performance results can be obtained when the training corpus is two times the size of the testing corpus. If the training case size is half of the testing case size, performance may decrease by 25%. We also re-implement the approach proposed by Yeh and Chen (2007) for the same data in our work. In their method, centering theory (CT) is adopted as the frame work to resolve zero anaphora. Since only grammatical roles and constraints are major criteria used for resolution, the performance is not satisfactory. In addition, numerous errors are caused due to misjudgment of verbal nominalization (VN) and lack of cataphora filter (CF). Table 8 illustrates that the performance is indeed improved if VN and CF are incorporated in resolution. The result indicates that our proposed method significantly outperforms the CT approach by 13%. Finally, a summary of errors of our proposed method is listed in Table 9. Seven types of errors are listed and the proportion of each error is calculated.

[3] CKIP Chinese word segmentation system is available at http://ckipsvr.iis.sinica.edu.tw/.

5. Conclusions

In this paper, we present a case-based reasoning approach to Chinese zero anaphora resolution. Compared to rule-based resolution methods, the presented approach turns out to be promising for dealing with both intra-sentential and inter-sentential zero anaphora. The contributions of our work are revealed from two aspects. First, a case-based reasoning approach with weighted KNN retrieval is demonstrated to be an effective method in comparison with the state-of-the-art rule-based approach. Second, we introduced two new features, semantic classes acquired from outer resources and conceptual patterns, for both ZA detection and ANT identification. Experimental results show that these two features can improve overall resolution performance by 11%. The drawback to this approach is that a case base must be constructed in advance. However, our experimental analysis shows that feasible performance results can be obtained when the training corpus is two times the size of the testing corpus. The future work will be directed toward definite anaphora and event anaphora by exploiting more useful resources. The web corpus will be investigated to identify associated patterns that could be useful in anaphora resolution.

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues. *methodological variations. and system approaches* (Vol. 7). AI Communications, IOS Press (pp. 39–59).
Cardie, C. (1999). Integrating case-based learning and cognitive biases for machine learning of natural language. *Journal of Experimental and Theoretical Artificial Intelligence, 11*(3), 297–337.
CKIP (1995). *The content and illustration of Sinica corpus of Academia Sinica*. Technical report no. 95-102, Institute of Information Science, Academia Sinica.
CKIP (1996). *A study of Chinese word boundaries and segmentation standard for information processing*. Technical report, Taiwan, Taipei, Academia Sinica.
Converse, S. P. (2005). Resolving pronominal references in Chinese with the Hobbs algorithm. In *Proceedings of the 4th SIGHAN workshop on Chinese language processing* (pp. 116–122).
Ding, B. G., Huang, C. N., & Huang, D. G. (2005). Chinese main verb identification: From specification to realization. *International Journal of Computational Linguistics and Chinese Language Processing, 10*, 53–94.
Hobbs, J. (1976). *Pronoun resolution*. Research report 76-1, Department of Computer Sciences, City College, City University of New York.
Huang, Y. (2000). *Anaphora: A cross-linguistic study*. Oxford, England: Oxford University Press.
Kennedy, C., & Boguraev, B. (1996). Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th international conference on computational linguistics* (pp. 113–118).
Lappin, S., & Leass, H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics, 20*(4), 535–561.
Lee, C. L. (2002). *Zero anaphora in Chinese*. Taipei: Crane publication company.
Li, W. (2004). Topic chains in Chinese discourse. *Discourse Processes, 37*, 25–45.
Liang, T., & Wu, D. S. (2004). Automatic pronominal anaphora resolution in English texts. *International Journal of Computational Linguistics and Chinese Language Processing, 9*, 1–20.
Liu, D. R., & Ke, C. K. (2007). Knowledge support for problem-solving in a production process: A hybrid of knowledge discovery and case-based reasoning. *Expert Systems with Applications, 33*(1), 147–161.
Liu, Y. H., Pan, W. Y., & Gu, W. (2002). *Shiyong xiandai hanyu yufa (Practical modern Chinese grammar)*. The Commercial Press.

Markert, K., & Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics, 31*(3), 367–402.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th international conference on computational linguistics* (pp. 869–875).

Mitkov, R. (1999). Multilingual anaphora resolution. *Machine Translation, 14*(3-4), 281–299.

Mitkov, R., Richard, E. & Orasan, C. (2002). A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the 3rd international conference on computational linguistics and intelligent text processing* (pp. 168–186).

Modjeska, N. N., Markert, K. & Nissim, M. (2003). Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 176–183).

Tao, Liang, & Healy, A. F. (2005). Zero anaphora: Transfer of reference tracking strategies from Chinese to English. *Journal of Psycholinguistic Research, 34*(2), 99–131.

Xu, J. J. (2003). *Anaphora in Chinese texts*. Beijing: China Social Science.

Yang, X. F., 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* (pp. 41–48).

Yeh, C. L., & Chen, Y. C. (2007). Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing, 17*(1), 41–56.

Yu, C. H. & Chen, H. H. (2000). *A study of chinese information extraction construction and coreference*. Unpublished master's 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* (pp. 541–550).