Chinese Pronominal Anaphora Resolution Using Lexical Knowledge and Entropy-Based Weight

Dian-Song Wu and Tyne Liang
Department of Computer Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan.
E-mail: {diansongwu, tliang}@cs.nctu.edu.tw

Pronominal anaphors are commonly observed in written texts. In this article, effective Chinese pronominal anaphora resolution is addressed by using lexical knowledge acquisition and salience measurement. The lexical knowledge acquisition is aimed to extract more semantic features, such as gender, number, and collocate compatibility by employing multiple resources. The presented salience measurement is based on entropy-based weighting on selecting antecedent candidates. The resolution is justified with a real corpus and compared with a rule-based model. Experimental results by five-fold cross-validation show that our approach yields 82.5% success rate on 1343 anaphoric instances. In comparison with a general rule-based approach, the performance is improved by 7%.

Anaphora denotes the phenomenon of referring back to previously mentioned items in a text. The reference is called an anaphor and the entity to which it refers is its antecedent. Effective anaphora resolution facilitates natural language applications, such as machine translation, text summarization, and information extraction. In this article, pronominal anaphora resolution in Chinese texts is addressed.

Pronominal anaphora resolution can be approached in different ways. Most traditional approaches are based on hand-crafted rules concerning constraints like syntactic parallelism, semantic parallelism, proximity, or parsing results (Converse, 2005; Lappin & Leass, 1994; Mitkov, 1999; Wang & Mei, 2005; Wang, Yuan, Wang, & Li, 2002; Yang, Su, & Tan, 2006). For example, a set of filtration and evaluation rules were used to resolve anaphora in Chinese financial texts (Wang et al., 2002). Another rule-based approach was described in (Wang & Mei, 2005) to resolve pronominal anaphora in Chinese texts by using number, gender, grammatical roles, and distance features. To obtain further structured relationship between anaphors and antecedents, Converse (2005) used full parsing results from the Penn Chinese Treebank and obtained 77.6% accuracy. Similarly, Yang et al. (2006) proposed pronominal resolution using the syntactic information extracted from the parse trees. The main drawbacks of rule-based approaches are attributed to intuitive observations and subjective biases in selecting feature weight. The accuracy is not always guaranteed by heuristics. Moreover, it takes laborious effort to designate new rules whenever the test data vary from original ones.

Recently, machine learning techniques have been employed in anaphora resolution (Ng, 2005; Ng & Cardie, 2002; Strube & Muller, 2003). To deal with insufficient knowledge acquired from a given corpus, the World Wide Web has been widely used as a corpus (Bergsma & Lin, 2006; Bunescu, 2003; Markert & Nissim, 2005; Modjeska, Markert, & Nissim, 2003). For example, Modjeska et al. (2003) utilized Web search and lexico-syntactic patterns to solve the out-of-vocabulary problem in hand-crafted lexicon. Bergsma and Lin (2006) presented a support vector machine (SVM)-based approach by using general features and path—coreference data that were extracted from a large parsed corpus to compensate for a paucity of data. Such approach successfully resolves 75% of 1078 anaphoric instances in English texts.

In contrast to profound studies of English texts, efficient Chinese anaphora resolution has not been widely addressed (Converse, 2005). The challenges involved are difficulty of proper noun identification and insufficiency of explicit contextual features. For example, morphological clues are rare for determining gender or number of Chinese nouns (Wang & Mei, 2005).

In this article, a hybrid approach using two strategies is presented to resolve pronominal anaphors in Chinese written texts. One is a Web-based acquisition model to extract useful lexical knowledge, such as gender, number, and collocate compatibility. Another is an adaptive weight salience measurement for antecedent identification. The experimental results show that our proposed approach yields 82.5% success rate on 1343 anaphoric instances, enhancing 7% improvement while compared with the general rule-based approach presented by Wang and Mei (2005).

Received October 27, 2007; revised March 17, 2008; accepted June 5, 2008

© 2008 ASIS&T • Published online 17 July 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20922

TABLE 1. A collocate compatibility example.

示威(VA)群眾(Na)和(Caa)警察(Na)對峙(VH), 他們$_1$(Nh)遊行(VA)抗議(VE)拆除(VC)行動(Na), 並(Cbb)與(P)警察(Na)發生(VJ)衝突(Na)。一些(Neqa)群眾(Na)在(P)衝突(Na)中(Ng)毆打(VC)警察(Na), 更(D)接著(D)搶走(VC)了(Di)他們$_2$(Nh)的(DE)配槍(Na)。

The demonstrating people confronted the policemen. They$_1$ paraded and protested the dismantling action and had conflicts with the policemen. Some people beat up the policemen in the conflict and then took away their$_2$ guns.)

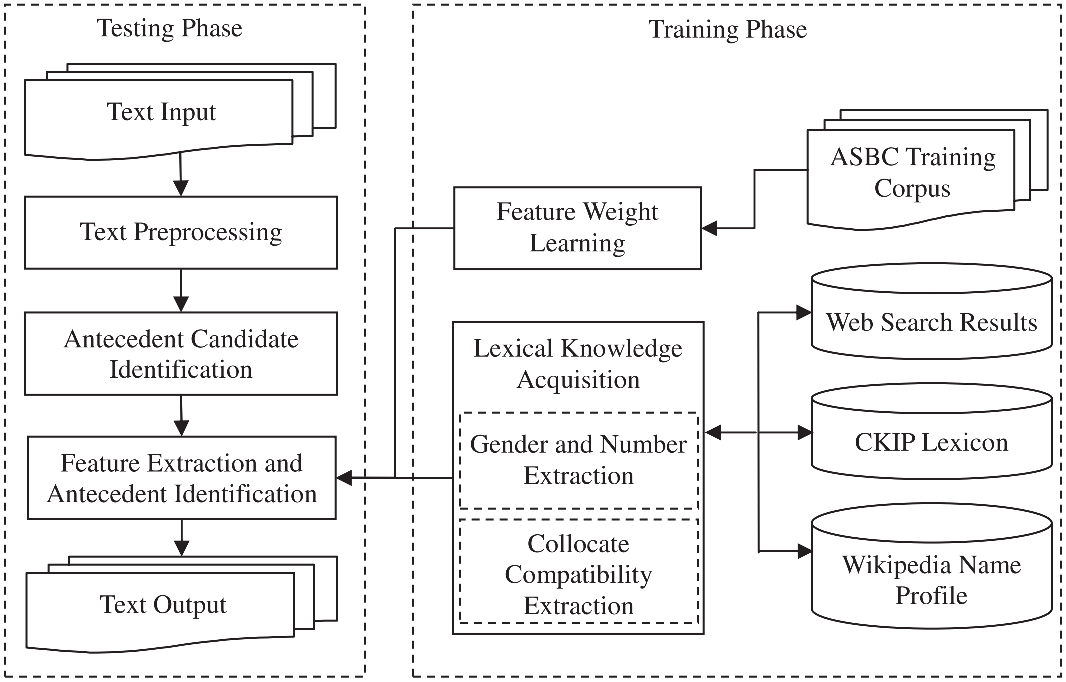
Testing Phase
Training Phase
Text Input
Text Preprocessing
Antecedent Candidate Identification
Feature Extraction and Antecedent Identification
Text Output
ASBC Training Corpus
Feature Weight Learning
Lexical Knowledge Acquisition
Gender and Number Extraction
Collocate Compatibility Extraction
Web Search Results
CKIP Lexicon
Wikipedia Name Profile

FIG. 1. The presented Chinese pronominal anaphora resolution procedure.

The rest of the article is organized as follows. The Chinese Pronominal Anaphora section introduces pronominal anaphora in Chinese texts and some of the problems encountered. The Approach section describes the proposed method by using lexical knowledge and entropy-based weight in detail. The Experiments and Analysis section present the resolution results and error analysis. The Conclusions section give a summary of our study.

Chinese Pronominal Anaphora

Pronominal anaphora resolution relies on the constraints between pronouns and antecedents, such as gender, number, grammatical role and animacy. As mentioned above, a general Chinese person's name does not always carry gender information and a Chinese noun does not have morphological differences for indicating its singularity or plurality.

In addition, identifying the referent of a pronoun in Chinese texts is not always trivial if insufficient real-world knowledge is incorporated. Table 1 lists two subsequent sentences where each word is followed by its part-of-speech[1], the first pronoun "他們$_1$" (they$_1$) refers to "示威群眾" (demonstrating people) while the second pronoun "他們$_2$" (their$_2$) refers to "警察" (policemen). So it is necessary for an anaphora resolver to check collocate compatibility between anaphors and their candidates.

The Approach

Figure 1 illustrates the presented pronominal anaphora resolution, which is incorporated with three external resources, namely, Web search results, Chinese knowledge information processing (CKIP) lexicon[2], and Wikipedia name profile. The resolution is implemented in the training phase and the testing phase. The training phase involves lexical knowledge acquisition and feature weight learning. Three kinds of lexical knowledge are addressed, namely, gender, number, and collocate compatibility. In feature weight learning, an entropy-based approach is employed. The testing phase concerns text preprocessing, antecedent candidate identification, feature extraction, and antecedent identification. The following subsections describe each component and the resolution procedure.

[1] A detailed description of part-of-speech tag set used in this article is available at http://ckipsvr.iis.sinica.edu.tw/category_list.doc. For example, "Na" denotes a common noun and "VA" means an intransitive verb.

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

DOI: 10.1002/asi

TABLE 2. Chinese noun phrase examples.

Types	Noun phrase examples
Common noun	每(Nes)位(Nf)用戶(Na) 的 (DE)個人(Na)***資料****(Na)* (every subscriber's individual ***information***)
Proper noun	委員會(Nc)主席(Na)***劉生明****(Nb)* (committee chairman ***Liu Shengming***)
Location noun	相當(Dfa)有名(VH)的(DE)***公園****(Nc)* (a very famous ***park***)
Temporal noun	十月(Nd)六日(Nd)***早上****(Nd)* (in the ***morning*** of October 6)
Verbal nominalization	心情(Na) 的 (DE)***放鬆***(VHC) (the ***release*** of mood)
Transformation case	放鬆 (VHC) 的(DE)***狀態***(Na) (the relaxed ***condition***)

Text Preprocessing

Text preprocessing includes sentence segmentation, POS tagging, named entity identification, and noun phrase chunking. The sentence segmentation and POS tagging are processed by CKIP Chinese word segmentation system[3]. The named entity identification is done by applying the hybrid approach presented in (Liang, Yeh, & Wu, 2003). In an experiment of 150 news documents selected from Academia Sinica Balanced Corpus (ASBC)[4], this approach yields 94% precision and 93% recall on person name identification, and 89% precision and 84% recall on organization name identification.

In this article, a finite state machine chunker is constructed to recognize noun phrases by their head nouns which may be common nouns, proper nouns, location nouns, temporal nouns, or pronouns (Yu & Chen, 2000). In Chinese, a head noun (as indicated in italics in Table 2) occurs at the end of a noun phrase. Except for noun phrase chunks, the chunker is also able to recognize verbal nominalization and transformation by utilizing heuristics described in (Ding, Huang, & Huang, 2005). As shown in Table 2, all the chunks, including the one containing the verb "放鬆" (relax), will be treated as antecedent candidates and will be assigned with semantic feature values like gender, animate and number useful at antecedent candidate identification.

Antecedent Candidate Identification

Table 3 lists the target pronominal anaphors to be resolved in this article. Unlike English pronouns, Chinese pronouns remain the same in expressing nominative and accusative cases. Table 4 lists the positional distribution of 692 anaphor-antecedent pairs in our training data and it shows that 94% of antecedents are in two sentences ahead of anaphors. Some antecedent candidates can be explicitly filtered out by applying the following heuristics. Here, CAN denotes an item in the candidate set preceding the corresponding pronominal anaphor (PA). A CAN will be filtered if it satisfies any of the following patterns.

1. Conjunction pattern: PA[c]CAN or CAN[c]PA
   c ∈ {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: PA[Vt]CAN or CAN[Vt] PA
   Vt denotes a transitive verb in a sentence.
3. Preposition pattern: PA[p]CAN or CAN[p]PA
   p ∈ {在，對，到，朝，給，向，比}

Lexical Resources

We use two lexical resources to acquire number and gender features for anaphora resolution. One is the CKIP lexicon (Chinese Knowledge Information Processing Group, 1995) and out of which we annotated 5,697 nouns with gender and number. For example, "硬漢" (tough guy) and "姑丈" (uncle) are marked as male nouns; "反對黨" (opposition party) and "考察團" (investigation group) are marked as plural. Table 5 shows the statistics of the annotated data in the tagged lexicon. The other resource, denoted as "Wikipedia Name Profile," was constructed by extracting 780 common Chinese person names from Wikipedia[5] and, for each name, the gender and role are tagged by hands. For instance, ("羅大佑" (Luo Da You), "男" (male), "歌手" (singer)) and ("劉墉" (Liu Yong), "男" (male), "作家" (writer)) are two entries stored in the Name Profile.

Lexical Knowledge Acquisition

Lexical knowledge acquisition involves the extraction of gender, number, and collocate compatibility from reliable patterns constructed at training phase.

Gender extraction. The gender extraction is to classify each chunk to be masculine, feminine or unknown with the help of the two lexical resources as described above or the so-called gender-indicating pattern (GP).

There are six kinds of GPs (denoted as "GP_i" and $1 \leq i \leq 6$) and each GP is tagged with masculine class (C_m) or feminine class (C_f) as follows:

1. *Attachment titles pattern (GP_1)*: N_i is followed by a gender-marked title.
   (a). If GP_1 is $N_i +$ [先生], then $N_i \in C_m$.
   (b). Else If GP_1 is $N_i +$ [女士|小姐|夫人], then $N_i \in C_f$.
2. *Opposite roles pattern* (GP_2): N_i acts as a possessive of some specific nouns.
   (a). If GP_2 is $N_i +$ [的] + [太太|妻子|夫人|老婆|女友|未婚妻], then $N_i \in C_m$.
   (b). Else If GP_2 is $N_i +$ [的] + [先生|丈夫|老公|男友|未婚夫], then $N_i \in C_f$.
3. *Reflexives pattern* (GP_3): N_i is followed by a reflexive.
   (a). If GP_3 is $N_i +$ [他自己], then $N_i \in C_m$.
   (b). Else If GP_3 is $N_i +$ [她自己], then $N_i \in C_f$.

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

[4] Academia Sinica Balanced Corpus is available at http://www.sinica.edu.tw/SinicaCorpus/

[5] Common Chinese person names are available at http://zh.wikipedia.org/w/index.php?title=%E4%BA%BA%E5%90%8D%E8%A1%A8&variant=zh-tw

TABLE 3. The target pronominal anaphors.

	Singular	Plural	Possessive(singular)	Possessive(plural)
Male	他(he, him)	他們(they, them)	他的(his)	他們的(their, theirs)
Female	她(she, her)	她們(they, them)	她的(her, hers)	她們的(their, theirs)
Neutral	它(it)	它們(they, them)	它的(its)	它們的(their, theirs)

TABLE 4. The positional distribution of anaphor-antecedent pairs.

Relative Position[a]	(1)	(2)	(3)	(4)	(5)
Number of pairs	48	319	284	21	20
Ratio	6.9%	46.0%	41.0%	3.0%	2.8%

[a] Relative Position:
(1) Pairs are in the same clause.
(2) Pairs are in the same sentence.
(3) Antecedents are in the previous sentence.
(4) Pairs are in the same paragraph.
(5) Pairs are not in the same paragraph.

TABLE 5. Gender and number statistics in the CKIP lexicon.

	Gender			Number	
Type	Male	Female	Neutral	Singular	Plural
Number of entries	502	515	4860	5345	352

4. *Possessives pattern* (GP_4): N_i is followed by a possessive.
   (a). If GP_4 is $N_i + [\text{他的}]$, then $N_i \in C_m$.
   (b). Else If GP_4 is $N_i + [\text{她的}]$, then $N_i \in C_f$.
5. *Complement derivation pattern* (GP_5): Person nouns are subjects and gender-marked nouns are in the predicate position. Gender-marked nouns are identified by using the tagged CKIP lexicon.
   (a). If GP_5 is $N_i + [\text{是}] + \text{Modifier} + \text{Male-noun}$, then $N_i \in C_m$.
   (b). Else If GP_5 is $N_i + [\text{是}] + \text{Modifier} + \text{Female-noun}$, then $N_i \in C_f$.
6. *Gender-modifier pattern* (GP_6): N_i is modified by a gender-modifier.
   (a). If GP_6 is gender-modifier $+ N_i$ and gender-modifier like "英俊" (handsome), then $N_i \in C_m$.
   (b). Else If GP_6 is gender-modifier $+ N_i$ and gender-modifier like "溫柔" (tender), then $N_i \in C_f$.

All the gender modifiers will be mined from the Web in advance by implementing following steps:

(i) Randomly select 100 male and female names, respectively.
(ii) Submit each name to the underlying search engine Google and acquire at most 50 snippets.
(iii) Retain nouns, verbs, adjectives, and adverbs in snippets.
(iv) Use Bayesian Parameter Learning (Equation 1; Russell & Norvig, 2003, chap. 20) and rank words in the ascending order of σ^2. The frequencies of words collocating with male names and female names are $\alpha - 1$ and $\beta - 1$, respectively.

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{1}$$

(v) Extract top 5 clue words for each gender.

Figure 2 illustrates the overall three-layer gender feature extraction for each N_i of an input text T_i and it is described as follows:

Step 1: If N_i is matched with the tagged CKIP lexicon or Wikipedia Name Profile, then return the corresponding gender.

Step 2: Else Search T_i with the help of gender-indicating patterns and gender information statistics *Gender*(N_i) defined in Equation 2. If the gender feature can be decided as male or female, then return the corresponding gender.

$$Gender(N_i) = \begin{cases} male, & if\ \rho_{male} > \rho_{female} \\ female, & if \rho_{female} < \rho_{male} \\ unknown, & otherwise \end{cases} \tag{2}$$

$$\rho_{male} = \frac{freq_{Cm}}{freq_{Cm} + freq_{Cf}}$$

$$\rho_{female} = \frac{freq_{Cf}}{freq_{Cm} + freq_{Cf}}$$

where
N_i: the noun to be assigned with gender feature
$freq_{Cm}$: the total number of all matched GP_i that belongs to C_m
$freq_{Cf}$: the total number of all matched GP_i that belongs to C_f

Step 3: Else transform N_i to queries according to each kind of *GPs*. For example, "$N_i + [\text{先生}]$", "$N_i + [\text{他自己}]$". Search the Web corpus for each gender-indicating pattern and calculate *Gender*(N_i). If the gender feature can be decided as male or female, the return the corresponding gender.

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

Number extraction. The number extraction is presented to facilitate resolving plural anaphors. With the collection of numerical clue words, the extraction is implemented in the following steps:

Step 1: Define symbols as follows:

NP = noun phrase;
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 = the set of plural numerals, such as {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 ∈ {[這|那|該|某|一] + {Q-P} + noun};

DOI: 10.1002/asi

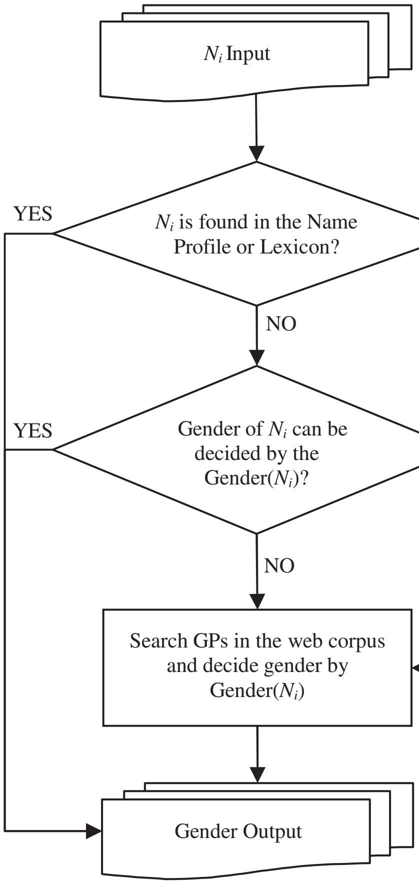
N_i Input
Wikipedia Name Profile
YES
N_i is found in the Name Profile or Lexicon?
NO
CKIP Lexicon
YES
Gender of N_i can be decided by the Gender(N_i)?
Gender-indicating Patterns
NO
Search GPs in the web corpus and decide gender by Gender(N_i)
Gender Output

FIG. 2. Gender identification procedure.

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 ∈ {們, 倆};
(iii) NP contains plural numbers + Q;
(iv) NP follows r, where r ∈ R;

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

Collocate compatibility extraction. The presented collocate compatibility extraction measures binding strength between candidates and anaphors. We consider three types of collocate patterns, namely agent-verb, verb-patient, and possessive-noun, and use collocate statistics to evaluate the preference of candidates. The collocate compatibility extraction is implemented as follows:

1. For each pronominal anaphor, replace it with its antecedent candidates accordingly.
2. According to the role (agent or patient) of the anaphor in its context, one collocate pattern is extracted for each candidate.

For instance, consider Table 1 mentioned above, anaphors “他們$_1$” and “他們$_2$” are the agent-verb and possessive-noun patterns, respectively. Therefore the collocate patterns for “他們$_1$” are “群眾遊行’ and “警察遊行”. Accordingly, “群眾的配槍” and “警察的配槍” are patterns for “他們$_2$”. For each candidate, its collocate compatibility with the anaphor is calculated by Equation 3. In the case of the anaphor “他們$_1$”, three queries are formed for each candidate and they are submitted to Google search engine. For candidate “群眾”, the *pattern query* is “群眾遊行’. Accordingly, the *candidate query* and the *attach query* are “群眾” and “遊行”, respectively.

$$Col\_Com(candidate, anaphor) \approx \log \frac{freq(pattern) \times N}{freq(candidate) \times freq(attach)} \quad (3)$$

where

N: total number of Google pages
freq(pattern): the number of pages retrieved with a pattern query
freq(candidate): the number of pages retrieved with a candidate query
freq(attach): the number of pages retrieved with a attach query

Feature Set

There are seventeen features concerned at our antecedent identification as follows. C denotes an antecedent candidate and P denotes the pronominal anaphor. For each feature, we set its value to be 1 if an antecedent candidate satisfies the feature constraint; otherwise we set its value to be 0.

1. *Same_Pro*: *C* and *P* are the same pronouns, for example, *C* is “她” (she) and *P* is “她” (she) as well.
2. *Reflexive*: *P* is a reflexive of *C*, such as “劉生明他自己” (Liu Shengming himself) in which “劉生明” (Liu Shengming) is an antecedent candidate.
3. *Role*: *C* is the agent of a verb, namely, *C* precedes a transitive verb or an intransitive verb.
4. *Parallel*: *C* and *P* are the same grammatical roles. For example, *C* and *P* are both subjects of sentences.
5. *Gender*: *C* and *P* are the same gender. The gender feature is identified by the way mentioned in the previous subsection *gender extraction*.
6. *Number*: *C* and *P* are the same number. The number feature is determined by the way mentioned in the previous subsection *number extraction*.
7. *Animate*: *C* is an animate entity and *P* is a male or female pronoun. We utilize the semantic class of CKIP lexicon to annotate animate entities. In addition, person names and organization names are regarded as animate entities, too.
8. *NE_Per*: *C* is a person name and *P* is a male or female pronoun. A person name is identified by using a classifier presented in (Liang, Yeh, & Wu, 2003).
9. *NE_Org*: *C* is an organization name and *P* is a plural pronoun. An organization name is identified by the way described above.
10. *Col_Com*: The value of *Col_Com*(*C*, *P*) is maximum. Equation 3 is used to calculate the value for each antecedent candidate.
11. *Same_Clause*: *C* and *P* are in the same clause. A clause is bounded by punctuation marks like “,”, “。”, “;”, “!”, and “?”.
12. *Same_Sent*: *C* and *P* are in the same sentence. A sentence is bounded by punctuation marks like “。”, “;”, “!”, and “?”.
13. *Same_Para*: *C* and *P* are in the same paragraph.
14. *Clause_Lead*: *C* is the first noun phrase in the clause.
15. *Sent_Lead*: *C* is the first noun phrase in the sentence.
16. *Repeat*: *C* repeats more than once in the context.
17. *Definite*: *C* is a definite noun phrase. For example, “這本雜誌” (the magazine) is a definite noun phrase.

Table 6 shows the feature vectors associated with some antecedent candidates of the anaphor 他們$_2$(they) in the example of Table 1. “警察” (policemen) is selected as the antecedent by applying the weighted salience measurement described in the following subsections.

Entropy-Based Weight

The weight function in Equation 4 is motivated from the decision tree learning, which utilizes the concept of entropy to select an attribute (Mitchell, 1997, chap. 3). The entropy value denotes the uncertainty associated with a random variable. In our case, a feature with lower entropy denotes that it can reduce uncertainty in selecting correct antecedents. Therefore, a feature with lower entropy is given a higher weight, and vice versa. During the training phase, positive instances were annotated manually. Other candidates between the positive pairs were used to form the negative instances.

$$\begin{aligned} weight_i &= 1 - entropy_i(S) \\ entropy_i(S) &= \sum_{j=1}^{v} \frac{|S_j|}{|S|} \times entropy(S_j) \\ entropy(S) &= -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \end{aligned} \tag{4}$$

where

S: the set of training instances
S_j: the subset of training instances in which $fval_i$ has value j
p: the number of positive instances
n: the number of negative instances

TABLE 6. Feature vectors of antecedent candidates.

Antecedent candidate	Feature vector
“一些群眾” (some people)	(0,0,1,0,0,1,1,0,0,0,0,1,1,1,1,1,0)
“衝突” (conflict)	(0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,1,0)
“警察” (policemen)	(0,0,0,1,0,1,1,0,0,1,0,1,1,0,0,1,0)

Antecedent Identification

The antecedent identification is implemented as follows:

Step 1: An input text is processed by the CKIP Chinese word segmentation system. An internal representation data structure encodes essential information, such as sentence offset, word offset, and word POS.

Step 2: Noun phrases in each sentence are identified by the NP chunker described above and stored in a global data structure. Each noun phrase will be tagged with number and gender features by applying the presented lexical knowledge acquisition.

Step 3: The target pronominal anaphors are identified throughout an input text. They are stored in a list containing their sentence offset and word offset.

Step 4: The remaining noun phrases, which are in a distance of two sentences ahead of an anaphor, are collected as antecedent candidates. Each candidate is further filtered by checking its gender, number, and animate agreement.

Step 5: For each candidate, its salience is evaluated by Equation 5. It is noticed that the salience will be set to be zero if one of the three agreements mentioned in Step 4 is negative. A candidate with the highest salience is selected for an anaphor.

$$\begin{aligned} Salience(can, ana) = {} & \frac{\sum_{i=1}^{n} (fval_i \times weight_i)}{\sum_{j=1}^{n} (max(fval_j) \times weight_j)} \\ & \times \prod_{k=1}^{3} agreement_k \end{aligned} \tag{5}$$

where

can: a candidate for a specified anaphor
ana: an anaphor to be resolved
$fval_i$: the ith feature value
$max(fval_i)$: the maximum value of the ith feature value
$agreement_k$: number, gender, and animate agreement
$weight_i$: the ith feature weight is computed by Equation 4

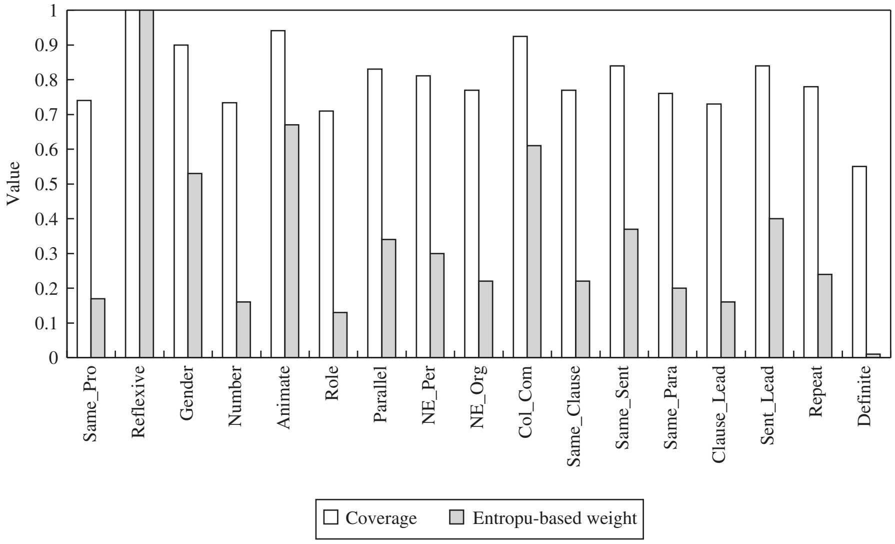

FIG. 3. The coverage ratio and entropy-based weight for each feature.

Experiments and Analysis

We extract 307 news documents from ASBC as our resolution corpus and from this corpus 1343 anaphor-antecedent pairs are identified by experts. The resolution performance is evaluated in terms of success rate defined by Equation 6 and is implemented by five-fold cross-validation. Each feature in the presented method is justified at testing phrase by the coverage ratio that is defined by Equation 7. Figure 3 illustrates the value of coverage ratio and entropy-based weight for each feature. It is found that features with top five weights are *Reflexive*, *Animate*, *Col_Com*, *Gender* and *Sent_Lead*, respectively. This result indicates that Reflexive, Animate and Gender features are three dominant features for animate entities if anaphors are gender-marked. In addition, the Col_Com feature shows the significance of collocate compatibility in selecting antecedents. Sent_Lead justifies the fact that Chinese is a topic prominent language.

$$success\ rate = \frac{number\ of\ correct\ resolution\ cases}{total\ number\ of\ anaphora\ cases\ identified} \tag{6}$$

$$coverage_i = \frac{\begin{array}{c}number\ of\ correct\ instances\ when\\ the\ ith\ feature\ is\ applied\end{array}}{\begin{array}{c}total\ number\ of\ instances\\ \text{that apply ith feature}\end{array}} \tag{7}$$

We implemented five resolution models for comparison. The baseline model was implemented by using number and gender agreement only, and the most recent subject was selected as the antecedent from a candidate set. The second model assigned equal weight to all 17 features and selected the top-weight candidate as the antecedent. The third and fourth models were implemented by considering four features only, namely number, gender, grammatical, and distance features. However, the third model assigned the features the same manual weight as described in Wang and Mei (2005), while the fourth model adopted our presented entropy-based weight. Table 7 shows that our method yields 82.5% success rate on 1343 anaphoric instances by employing entropy-based weight scheme and lexical knowledge. It improves about 7% success rate while compared with a rule-based model like the one presented in (Wang & Mei, 2005). Table 8 lists the distribution of each type of anaphors and individual success rate. It is found that anaphors with gender-mark are more easily to be resolved than the neutral ones. Similar conclusion can be found for those singular anaphors.

TABLE 7. Performance evaluation.

Method	Success rate
Baseline model	51.6%
Rule-based (equal-weighted)	72.5%
Wang & Mei (2005)	75.7%
Wang & Mei (2005) + entropy weight	78.2%
Our method	82.5%

TABLE 8. Anaphoric types and their success rate.

Type of anaphor	他(的)	他們(的)	她(的)	她們(的)	它(的)	它們(的)
# of identified instances	825	207	162	30	88	31
# of correctly resolved instances	697	162	134	24	69	22
Success rate	84.4%	78.2%	82.7%	80.0%	78.4%	73.3%

TABLE 9. Error analysis.

Error types	# of error instances	Ratio
POS tagging/chunking error	70	30%
Gender mismatch	49	21%
Inappropriate salience	38	16%
Exceeding window size	29	12%
Number mismatch	28	12%
Multiple antecedents	12	5%
Others	9	4%
Total	235	100%

The resolution errors are summarized in Table 9. As we can see, most of errors are attributed to preprocessing and gender constraints. One reason is that pronoun "他" (he) is often incorrectly used to identify a female entity in Chinese texts. So the presented resolution can be further improved by considering this phenomenon.

Conclusions

Our contributions are that we proposed three innovative methods for lexical knowledge acquisition, and our study is the first one that utilizes entropy-based weight in anaphora resolution. Compared with the manual weight scheme, the presented entropy-based weight scheme is more capable to estimate the likelihood that a candidate turns out to be an antecedent. Moreover, the presented lexical knowledge acquisition is indeed able to acquire more semantic information from contexts and Web resources. In comparison with a general rule-based approach, the presented resolution can achieve 7% improvement when lexical knowledge learning and entropy-based weight are implemented.

References

Bergsma, S. & Lin, D. (2006). Bootstrapping path-based pronoun resolution. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (pp. 33–40).

Bunescu, R. (2003). Associative anaphora resolution: A Web-based approach. Proceedings of the EACL Workshop on the Computational Treatment of Anaphora (pp. 47–52).

Chinese Knowledge Information Processing Group. (1995). The content and illustration of Sinica corpus of Academia Sinica (Report No. 95–102), Institute of Information Science, Academia Sinica.

Converse, S.P. (2005). Resolving pronominal references in chinese with the hobbs algorithm. 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.

Lappin, S. & Leass, H. (1994). An algorithm for pronominal anaphora resolution. Computational Linguistics, 20(4), 535–561.

Liang, T., Yeh, C.H., & Wu, D.S. (2003). A Corpus-based categorization for Chinese proper nouns. Proceedings of the National Computer Symposium (pp. 434–443).

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

Mitchell, T.M. (1997). Machine learning. McGraw-Hall companies.

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

Modjeska, N.N., Markert, K., & Nissim, M. (2003). Using the Web in machine learning for other-anaphora resolution. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 176–183).

Ng, V. (2005). Machine learning for coreference resolution: From local classification to global ranking. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (pp. 157–164).

Ng, V. & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 104–111).

Russell, S.J., & Norvig, P. (2003). Artificial intelligence: A modern approach (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Strube, M., & Muller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (pp. 168–175).

Stuckardt, R. (2002). Machine-learning-based vs. manually designed approaches to anaphor resolution: The best of two worlds. Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (pp. 211–216).

Wang, H.F., & Mei, Z. (2005). Robust pronominal resolution within Chinese text. Journal of Software, 16, 700–707.

Wang, N., Yuan, C.F., Wang, K.F., & Li, W.J. (2002). Anaphora resolution in Chinese financial news for information extraction. Proceedings of the 4th World Congress on Intelligent Control and Automation (pp. 2422–2426).

Yang, X.F., Su, J., & Tan, C.L. (2006). Kernel-based pronoun resolution with structured syntactic knowledge. 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. (2005). Zero anaphora resolution in Chinese with shallow parsing. Journal of Chinese Language and Computing. (to appear)

Yu, C.H., & Chen, H.H. (2000). A study of Chinese information extraction construction and coreference. Unpublished master's thesis, National Taiwan University, Taiwan.