PAPER Special Section on Data Engineering

Improving Definite Anaphora Resolution by Effective Weight Learning and Web-Based Knowledge Acquisition

Dian-Song WU[†a)], *Student Member and* **Tyne LIANG**[†b)], *Nonmember*

SUMMARY In this paper, effective Chinese definite anaphora resolution is addressed by using feature weight learning and Web-based knowledge acquisition. The presented salience measurement is based on entropy-based weighting on selecting antecedent candidates. The knowledge acquisition model is aimed to extract more semantic features, such as gender, number, and semantic compatibility by employing multiple resources and Web mining. The resolution is justified with a real corpus and compared with a classification-based model. Experimental results show that our approach yields 72.5% success rate on 426 anaphoric instances. In comparison with a general classification-based approach, the performance is improved by 4.7%.

key words: *definite anaphora resolution, feature weight learning, Web mining*

1. Introduction

In natural language, anaphora plays an essential role in the cohesion of discourses. Anaphora denotes the phenomenon of referring back to previously mentioned entities in a text. The referring expression is called an anaphor and the entity to which it refers is its antecedent. Anaphora resolution denotes the process of identifying the anaphoric relation between two expressions in a context. Effective anaphora resolution facilitates the task of natural language processing such as text summarization, information extraction, and machine translation. Different kinds of anaphoric expressions can be utilized in the context, such as personal pronouns, definite noun phrases, and ellipses. In this paper, we focus on the resolution of Chinese definite anaphora.

Traditionally, anaphora resolution is approached by hand-crafted rules concerning constraints like gender agreement, number agreement, syntactic parallelism, and proximity [1]–[4]. In the recent decade, the trend of anaphora resolution has been moved toward machine learning approaches [5]–[10]. Most of these learning-based approaches recast anaphora resolution as a binary classification problem. A classifier is trained in advance to determine whether a candidate and an anaphor are anaphoric or not. In addition, to deal with insufficient knowledge acquired from lexicons or given corpora, the World Wide Web has been widely used as a corpus [11]–[15]. For example, Modjeska et al. [14] utilized Web search and lexico-syntactic patterns to solve the out-of-vocabulary problem in hand-crafted lexicon.

In contrast to profound studies of English texts, efficient Chinese anaphora resolution has not been widely addressed [16]. Difficulties involved are mainly attributed to the following factors: First, morphological clues are rare for determining gender or number of Chinese nouns [17]. Second, no capitalization feature to identify proper nouns. Third, no sufficient ontology, such as WordNet, is available for identifying hypernymy or hyponymy relation between concepts. Not only morphological or syntactic knowledge but also information on lexical semantics and domain knowledge is required to enhance the resolution results. Moreover, there exist some drawbacks by adapting conventional approaches for anaphora resolution. For a rule-based approach, a salience score by manual weight assignment is usually adopted to select the antecedent. Errors may occur due to intuitive observations and subjective biases in selecting feature weight. On the other hand, the drawback of a classification-based approach is that it forces different candidates for the same anaphor to be considered independently [18]. Only a single candidate is evaluated at a time and the resolution proceeds in the reverse order of sentences until an antecedent is found. This may cause a real antecedent to be neglected once the classifier labels a candidate to be positive.

In this paper, a novel approach using two strategies is presented to resolve Chinese definite anaphors in written texts and avoid the drawbacks mentioned above. One is an adaptive weight salience measurement for antecedent identification. A weighted ranker is utilized to estimate the entire set of candidates simultaneously. Another is a Web-based knowledge acquisition model to extract useful lexical knowledge, such as gender, number, and semantic compatibility. The experimental results show that our proposed approach yields 72.5% success rate on 426 anaphoric instances, enhancing 4.7% improvement while compared with the result conducted by a conventional classifier.

The rest of the paper is organized as follows: Section 2 introduces the commonly-seen definite anaphora instances in Chinese texts. Section 3 describes the proposed method by using feature weight learning and lexical knowledge acquisition in detail. Section 4 describes the experimental results and analysis. Section 5 presents the final conclusions.

Manuscript received June 4, 2010.
Manuscript revised October 3, 2010.

[†]The authors are with the Department of Computer Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan.
a) E-mail: diansongwu@cs.nctu.edu.tw
b) E-mail: tliang@cs.nctu.edu.tw

DOI: 10.1587/transinf.E94.D.535

2. Chinese Definite Anaphora

2.1 Types of Chinese Definite Anaphora

Three common Chinese anaphoric phenomena are zero, pronominal, and definite anaphora. In our previous research, the distribution ratios of these three types are 57%, 29%, and 14%, respectively. In Chinese zero and pronominal anaphora resolution, it is showed that applying weight learning can significantly improve resolution performance in comparison with un-weighted methods [19]. Therefore, we address the problem of Chinese definite anaphora for overall comprehension of Chinese anaphora resolution tasks.

In Chinese definite anaphora (DA), an antecedent can be mentioned by a definite noun phrase preceded by demonstratives like "這" (this), "此" (this), "那" (that), "其" (that). Similarly, an English definite noun phrase is introduced by a definite article "the". In this paper, we tackle Chinese definite anaphor with the pattern like "[這 (this)] + [量詞 (quantifier)] + [實體名詞片語 (physical noun phrase)]". Grammatically, definiteness is a feature of noun phrases, indicating entities which are specific and identifiable in a given context. The type of DA may be partial overlap relation as in example (1), synonymous relation as in example (2), or hyponymy-hypernymy relation as in example (3).

(1) Partial overlap relation:
波灣戰爭$_1$於1991年由美國主導進軍伊拉克，
這場戰爭$_1$對國際政治與經濟造成巨大的影響。
Gulf War$_1$ was led by the United States to attack Iraq in 1991. The war$_1$ caused huge impact on international politics and economics.

(2) Synonymous relation:
上周鄰居家遭到小偷$_2$侵入，警方推測這竊賊$_2$可能是經由窗戶進入屋內。
Thieves$_2$ intruded into neighbor's house last week. The police thought that the burglars$_2$ probably enter the house through the windows.

(3) Hyponymy-hypernymy relation:
帕金森氏症$_3$是由於腦神經細胞退化所造成，在醫學上相信這種疾病$_3$與多巴胺有密切關係。
Parkinson's disease$_3$ is caused by the degradation of brain cells. Medically, it is believed that this disease$_3$ is closely related to dopamine.

Anaphoric relation in (1) can be resolved by matching the head nouns of noun phrases explicitly. As to the other two cases, surface features are no longer adequate to identify the correct antecedents. Most previous studies rely on pre-constructed lexicons as knowledge sources. However, it suffers from the problem of coverage. Besides, no sophisticated lexicon is available yet for identifying relation between Chinese expressions as shown in (3). Thus, we utilize a Web-based approach for exploiting semantic relationships such as hyponymy and hypernymy that are not included in lexicon resources.

2.2 Text Preprocessing

The training and testing data are selected from Academia Sinica Balanced Corpus[†] (ASBC). The named entity identification is done by applying the hybrid approach presented in [20]. 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 [21]; 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:

(1) 每(Nes)[††] 個(Nf)人(Na)的(DE)基本(A)權力(Na)
(the basic rights of everyone)

(2) 學生會(Na)主席(Na)陳文生(Nb)
(the student union chairman Chenwensheng)

(3) 非常(Dfa)著名(VH)的(DE)公司(Nc)
(a very famous company)

(4) 三月(Nd)三日(Nd)下午(Nd)
(the afternoon of March 3)

The presented chunker is also able to recognize verbal nominalization and transformation (as shown in Table 1) by utilizing heuristics discussed in [22]. 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.

Table 1 Verbal nominalization and transformation cases of word "努力" (work hard).

Type	Example
Verbal nominalization	所有(Neqa)的(DE)***努力***(VH)
	(all the ***efforts***)
Transformation of adjective case	***努力***(VH)的(DE)表情(Na)
	(the look of ***hard working***)
Transformation of adverb case	***努力***(VH)地(DE)工作(Na)
	(to work ***hard***)

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

[††]The symbol in brackets denotes the part-of-speech tag of a word. A detail description is available at http://ckipsvr.iis.sinica.edu.tw/papers/category_list.doc

Table 2 The positional distribution of anaphor-antecedent pairs.

Relative Position*	(a)	(b)	(c)	(d)
Number of pairs	223	433	575	585
Ratio	36.0%	70.0%	93.0%	94.6%

* Relative Position:

(a) Antecedents are in the same sentence.

(b) Antecedents are in the previous sentence.

(c) Antecedents are in the two previous sentences.

(d) Antecedents are in the same paragraph.

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.

In addition, we investigate the positional distribution of 618 anaphor-antecedent pairs in our training data. Table 2 shows that 93% of antecedents are in two sentences ahead of the definite anaphors.

2.3 Antecedent Candidate Identification

For each definite anaphor we extracted the set of all potential NP-antecedents in the two-sentence window. In addition, the following constraints are applied to filter out candidates with respect to a corresponding definite anaphor. *can* denotes an item in the candidate set preceding the definite anaphor *ana*. If *can* satisfies any of the following patterns, it is regarded as a non-antecedent instance.

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

3. The Approach

Figure 1 illustrates the presented definite anaphora resolution which is incorporated with three external resources, namely Web search results, CKIP lexicon†, and Tongyici Cilin††. The resolution is implemented in the training phase and the testing phase. The training phase involves feature weight learning and lexical knowledge acquisition. Three kinds of lexical knowledge are addressed, namely, gender, number, and semantic 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.

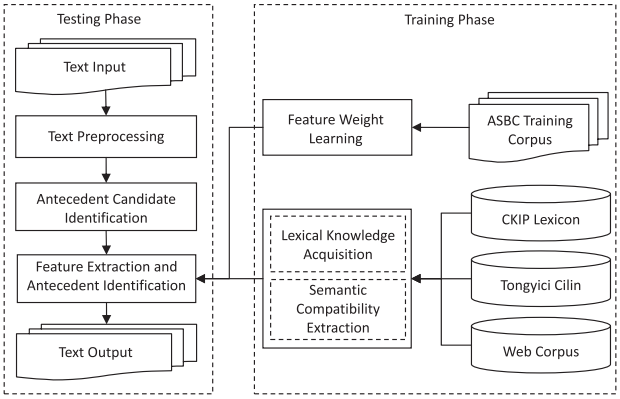

Fig. 1 The system architecture.

Table 3 A comparison of knowledge acquisition methods.

	[Bunescu, 2003]	[Markert et al., 2003]	[Shinzato and Torisawa, 2004]	[Garera and Yarowsky, 2006]	Our method
Language	English	English	Japanese	English	Chinese
Corpus	Web corpus	Web corpus	Web corpus	LDC Gigaword corpus	Web corpus
Outer resources	None	Gazetteers, IE software	None	WordNet	Name profile, CKIP lexicon, Tongyici Cilin
Method	Lexico-syntactic patterns	Query patterns	Itemization or listing clues	Co-occurrence model	Gender-indicating patterns, query patterns
Acquired knowledge	Identity and associative anaphora	Other-anaphora and bridging	Hyponymy	Hypernymy	Gender, semantic compatibility, definite anaphor

3.1 Lexical Knowledge Acquisition

To resolve Chinese definite anaphora more accurately, knowledge about gender, number, and semantic compatibility is essential. In order to acquire such knowledge, we utilize pre-constructed patterns, lexicon resources, and context information to extract lexical knowledge. However, these methods may suffer from the problem of data sparseness. To deal with this problem, web-based knowledge acquisition methods are then applied. Latent semantic relations which are not identified in local contexts can be acquired form web mining results. Table 3 shows the comparison of knowledge acquisition methods.

3.1.1 Gender Extraction

The gender extraction aims to classify each noun phrase to be male, female or unknown with the help of so-called gender-indicating pattern (GP) and Web mining results. All the gender modifiers are mined from the Web in advance by implementing the procedure as shown in Fig. 2. Moreover, there are six kinds of GPs (denoted as "GP_i" and $1 \le i \le 6$) and each GP is utilize to identify the occurrence of masculine pattern or feminine pattern as shown in Fig. 3.

Figure 4 illustrates the overall three-layer gender feature extraction for each N_i of an input document D_i and it is

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

††Tongyici Cilin extended version is available at http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

Algorithm 3.1. The gender modifier mining algorithm

Input: Randomly select 100 male name m_i and 100 female names f_i, respectively

Output: Top 5 clue words for male and female, respectively

Procedure Gmod():

Step 1: Submit each name to the search engine Google and acquire at most 50 snippets

Step 2: Retain nouns, verbs, adjectives, and adverbs in snippets as set $W=\{w_1,w_2,\ldots, w_n\}$

Step 3: For each $w_i \in W$ do

Calculate cnt_m: the frequency that w_i appears with male names

Calculate cnt_f: the frequency that w_i appears with female names

Step 4: Select the set $W_m=\{ w_1,w_2,\ldots, w_i \}$, where $\frac{cnt_m}{cnt_f + cnt_m} > 0.8$

Select the set $W_f=\{ w_1,w_2,\ldots, w_j \}$, where $\frac{cnt_f}{cnt_f + cnt_m} > 0.8$

Step 5: Use Bayesian Parameter Learning (Equation (1)) [23] 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

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

Step 6: Output top 5 clue words from W_m and W_f, respectively

Fig. 2 The gender modifier mining algorithm.

Algorithm 3.2. The gender-indicating pattern identification algorithm

Input: 1. A candidate noun phrase N_i

2. The count of masculine patterns C_m and feminine patterns C_f

Output: The number feature f_{gnd}, where $f_{gnd} \in$ {male, female, unknown}

Procedure Gender():

Step 1: Search *Attachment titles pattern* (GP_1): N_i is followed by a gender-marked title

(a). If GP_1 is N_i+[先生], then C_m++

(b). Else if GP_1 is N_i+ [女士|小姐|夫人], then C_f++

Step 2: Search *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 C_m ++

(b). Else if GP_2 is N_i+[的]+ [先生|丈夫|老公|男友|未婚夫], then C_f ++

Step 3: Search *Reflexives pattern* (GP_3): N_i is followed by a reflexive

(a). If GP_3 is N_i+[他自己], then C_m++

(b). Else if GP_3 is N_i+[她自己], then C_f++

Step 4: Search *Possessives pattern* (GP_4): N_i is followed by a possessive

(a). If GP_4 is N_i+[他的], then C_m++

(b). Else if GP_4 is N_i +[她的], then C_f++

Step 5: Search *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+[是]+Modifier+Male-noun, then C_m++

(b). Else if GP_5 is N_i+[是]+Modifier+Female-noun, then C_f++

Step 6: Search *Gender-modifier pattern* (GP_6): N_i is modified by a gender-modifier which is mined by **Gmod()** as shown in Figure 2

(a). If GP_6 is gender-modifier +N_i and gender-modifier like "英俊" (handsome), then C_m++

(b). Else if GP_6 is gender-modifier +N_i and gender-modifier like "溫柔" (tender), then C_f++

Step 7: Calculate the feature value f_{gnd} =*Gender*(N_i)

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

$$\rho_{male} = \frac{C_m}{C_m + C_f}$$

$$\rho_{female} = \frac{C_f}{C_m + C_f}$$

Step 8: Output f_{gnd}

Fig. 3 The gender-indicating pattern identification algorithm.

described as follows:

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

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

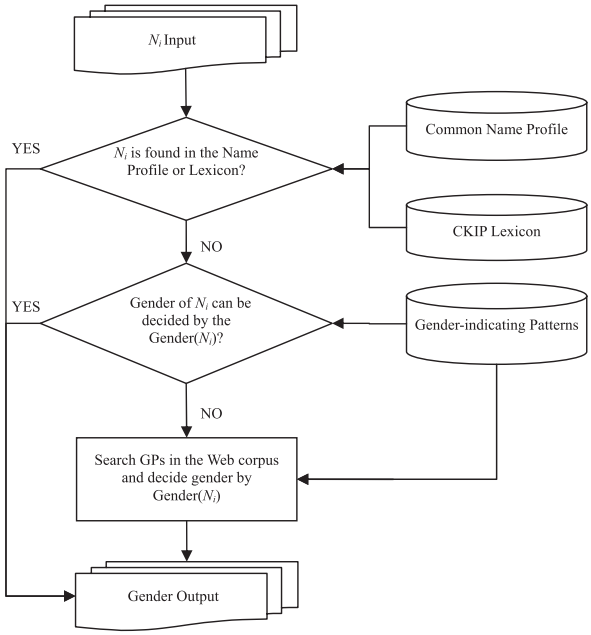

Fig. 4 The gender extraction procedure.

Step 3: Else transform N_i to queries according to each kind of GPs. For example, "N_i + [先生]", "N_i + [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, then return the corresponding gender.

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

3.1.2 Number Extraction

The number extraction is aimed to facilitate resolving plural anaphors. With the collection of numerical and quantitative clue words, the extraction is implemented as shown in Fig. 5.

3.1.3 Semantic Compatibility Extraction

To acquire semantic knowledge from the Web, we submit queries consisted of candidates and anaphors to the Google search engine. Queries are formed by patterns that structurally express the same semantic relationships. The co-occurrence statistics of such patterns can then be used as a mechanism for detecting the hypernymy-hyponymy relation between the definite anaphor and its potential antecedents. In the case of a candidate "蘋果 (apple)" and the definite anaphor "這種水果 (this kind of fruit)", queries like < "蘋

[†]Common Chinese person names are available at http://zh.wikipedia.org/w/index.php

Algorithm 3.3. The number extraction algorithm for assigning the number feature to each candidate noun phrase of input sentences

Input: 1. A candidate noun phrase *NP*
2. The set of quantifiers *Q*
3. The set of collective quantifiers *P*={群, 夥, 堆, 對,… 批}
4. The set of plural numerals *R*={全部, 所有, 數個, 許多,…, 諸多}

Output: The number feature f_{num}, where $f_{num} \in$ {singular, plural, unknown}

Procedure Number():

Step 1: Identify head noun *HNP* of the candidate noun phrase *NP*

Step 2: **If** *NP* satisfies any of the following conditions, **then** return f_{num} = 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 f_{num} = 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, return f_{mum} = unknown

Fig. 5 The number extraction procedure.

Algorithm 3.4. The semantic compatibility extraction algorithm for mining hypernymy and hyponymy relations

Input: A candidate noun phrase *can,* a definite anaphora *ana*

Output: The value of Sem_Com for pair *can* and *ana*

Procedure Sem_Com():

Step 1: Identify the head noun of *can* as *m*
Identify the head noun of *ana* as *n*

Step 2: Submit query [*m*]+[是一種]+[*n*] to Google
Calculate the number of pages cnt_{q1}

Step 3: Submit query [*m*]+[這種]+[*n*] to Google
Calculate the number of pages cnt_{q2}

Step 4: Submit query [*m*]+[和其他]+[*n*] to Google
Calculate the number of pages cnt_{q3}

Step 5: Acquire the numer of pages cnt_m by submitting *m* as query
Acquire the numer of pages cnt_n by submitting *n* as query

Step 6: Calculate $Sem\_Com(can, ana) = log \frac{p(cnt_{pair})}{p(cnt_m) \times p(cnt_n)}$ (3)

$$p(cnt_{pair}) = \frac{(cnt_{q1} + cnt_{q2} + cnt_{q3})}{cnt_{total}}$$

$$p(cnt_m) = \frac{cnt_m}{cnt_{total}}$$

$$p(cnt_n) = \frac{cnt_n}{cnt_{total}}$$

where cnt_{total} is the number of Google pages

Step 7: Output the value of $Sem\_Com(can, ana)$

Fig. 6 The semantic compatibility extraction algorithm.

果是一種 (apple is a kind of)" + "水果 (fruit)" >, < "蘋果這種 (apple this kind of)" + "水果 (fruit)" >, and < "蘋果和其他 (apple and other)" + "水果 (fruit)" > are concerned and the implementation is shown as Fig. 6.

3.2 Feature Set

There are fifteen features concerned as shown in Table 4. *can* denotes an antecedent candidate and *ana* denotes the definite 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.

Table 4 Summary of features.

Type	Feature	Description
Lexical	*Head_Match*	*can* and *ana* have the same head word.
	Str_Overlap	*can* and *ana* have overlapping words.
	Non_Emb	*can* is not an embedded noun phrase.
	Definite	*can* follows a determiner.
Grammatical	*Gender*	*can* and *ana* are the same gender.
	Number	*can* and *ana* are the same number.
	Role	*can* and *ana* are the same grammatical role.
Semantic	*Cilin_Syn*	*can* and *ana* are synonyms in Tongyici Cilin.
	Animate	*can* and *ana* are both animate entities.
	Sem_Com	The value of *Sem_Com*(*can, ana*) is maximum.
	Same_SC	*can* and *ana* are the same semantic class in CKIP lexicon.
Heuristic	*Coh_Cue*	*can* and *ana* are connected by coherence cue words.
	Repeat	*can* repeats more than once in the context.
	Sent_Lead	*can* is the first noun phrase in the sentence.
	Fwd_Cent	*can* is a forward looking center.

3.3 Feature Weight Learning

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. In the training phase, 318 news documents containing 618 positive and 1077 negative pairs are used as training data. The weight of each feature is calculated by Eq. (4). Figure 7 shows the entropy-based weight distribution of each feature.

$$weight_i = 1 - entropy_i(S)$$

$$entropy_i(S) = \sum_{j=1}^{v} \frac{|S_j|}{|S|} \times entropy(S_j) \quad (4)$$

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

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

3.4 Classification-Based Module

Support Vector Machine (SVM) is a useful technique for data classification. It is widely used in the research of natural language processing problems. In anaphora resolution, SVM-based classifiers are commonly applied for identifying

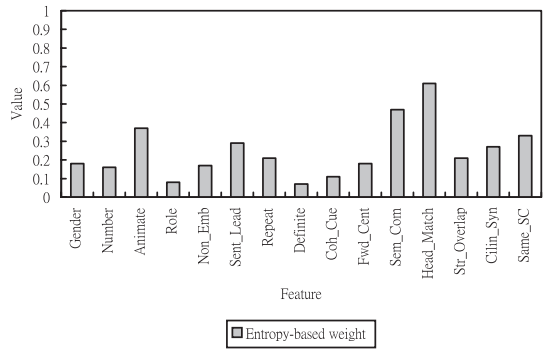

Fig. 7 Entropy-based weight distribution.

Algorithm 3.5. The antecedent identification algorithm

Input: A document D

Output: Anaphor-antecedent pairs $(ana, ant)_p$

Procedure Resolve():

Step 1: Build the internal representation data structure of input document D. For example, sentence offset, word offset, and word POS.

Step 2: **For each** sentence in D **do**

Identify noun phrases in each sentence by the NP chunker described above

Step 3: Identify all target definite anaphors ana_p throughout D

Step 4: **For each** ana_p **do**

Collect candidate set S. All antecedent candidates can_q in S are in a distance of two sentences ahead of ana_p

For each candidate $can_q \in S$ **do**

(i) Assign feature values to can_q

(ii) Rank pairs by Equation (5)

$$Rank(can, ana) = \frac{\sum_{i=1}^{n}(fval_i \times weight_i)}{\sum_{j=1}^{n}(\max(fval_j) \times weight_j)} \times \prod_{k=1}^{3} agreement_k \quad (5)$$

where

can: a candidate for a specified anaphor

ana: an anaphor to be resolved

$fval_i$: the i^{th} feature value

$max(fval_i)$: the maximum value of the i^{th} feature value

$agreement_k$: number, gender, and animate agreement

$weight_i$: the i^{th} feature weight is computed by Equation (4)

(iii) A candidate can_q with the highest Rank value is selected as antecedent ant_p for a definite anaphor

Step 5: Output $(ana, ant)_p$

Fig. 8 The antecedent identification algorithm.

potential antecedents [8], [9]. To compare with the performance of our proposed method, we used SVM as a baseline model and utilize LIBSVM[†] as a classification tool.

3.5 Antecedent Identification

The task of antecedent identification is to select the most likely candidate from the candidate set by Eq. (5). Each candidate is filtered by checking its gender, number, and animate agreement. "Agreement$_k$" is a binary function that has a value 0/1. It is noticed that the value of $Rank(can, ana)$ will be set to be zero if one of the three agreements is zero. A candidate with the highest value is selected as the antecedent for the target definite anaphor. The antecedent identification is implemented as shown in Fig. 8.

Table 5 Distribution of top 10 semantic classes.

Semantic Class	Ratio
mankind	22.0%
equipments	8.9%
place	6.1%
machines	4.4%
organizations	4.4%
buildings	3.8%
fine_arts	3.3%
nonhuman	3.0%
solid	3.0%
regions	2.8%

Table 6 Performance evaluation.

Models	Success rate
Equal-weighted	48.6%
Classification-based	67.8%
Our method	72.5%

Table 7 Performance of leave-group-out evaluation.

Type	Success rate
Lexical	62.6%
Grammatical	66.8%
Semantic	58.2%
Heuristic	65.5%

4. Experiments and Analysis

We extract 204 news documents from ASBC as our resolution corpus and from this corpus 426 anaphor-antecedent pairs are identified by experts. Table 5 lists the top 10 semantic class statistics in our corpus. The resolution performance is evaluated in terms of success rate defined by Eq. (6). To evaluate the performance of our proposed method, we implement three resolution strategies for comparison as shown in Table 6. The first model utilizes equal-weighted salience measures to identify antecedents. Namely, the weight of each feature is set to be 1. In the second model, a classification-based method is implemented by using SVM. In our proposed method, each feature is weighted by Eq. (4). It is found that features with top five weights are *Head_Match*, *Sem_Com*, *Animate*, *Same_SC*, and *Sent_Lead*, respectively. This result indicates that *Head_Match*, *Animate* and *Same_SC* features are three dominant features for the characteristic of semantic agreement. In addition, the *Sem_Com* feature shows the significance of collocate compatibility in selecting antecedents. *Sent_Lead* justifies the fact that Chinese is a topic prominent language.

Table 6 shows that our method yields 72.5% success rate on 426 anaphoric instances by employing entropy-based weight scheme and web-based lexical knowledge. It improves about 4.7% success rate while compared with a classification-based model. In addition, to find out the contribution of each type of features in our proposed method, we conduct a leave-group-out evaluation as shown in Ta-

[†]The LIBSVM tool is available at http://www.csie.ntu.edu.tw/~cjlin/.

ble 7. Four types of features are concerned, for example, lexical, grammatical, semantic, and heuristic. It shows that the type of semantic features plays the most important role since the success rate decreases significantly when this type of features is disable.

$$\text{success rate} = \frac{\text{number of correct resolution cases}}{\text{total number of anaphora cases identified}} \quad (6)$$

5. Conclusions

To our knowledge, our method represents the first attempt to use weight learning and Web-based knowledge acquisition for resolving definite anaphora in Chinese text. To overcome the drawback of common rule-based methods that employed manual weights, an effective measurement is constructed on the basis of entropy-based weight to estimate the likelihood of antecedent candidates. Moreover, to cope with the difficulty of feature extraction in Chinese texts, a Web-based knowledge acquisition model is proposed to extract gender, number, and semantic compatibility from contextual information and Web resources. Our experimental results show that the method can achieve a significant increase in the success rate of around 4.7% when lexical knowledge learning and entropy-based weighting are utilized.

References

[1] C. Kennedy and B. Boguraev, "Anaphor for everyone: Pronominal anaphora resolution without a parser," Proc. COLING, pp.113–118, 1996.
[2] S. Lappin and H. Leass, "An algorithm for pronominal anaphora resolution," Computational Linguistics, vol.20, no.4, pp.535–562, 1994.
[3] R. Mitkov, "Robust pronoun resolution with limited knowledge," Proc. COLING/ACL, pp.869–875, 1998.
[4] N. Wang, C.F. Yuan, K.F. Wong, and W.J. Li, "Anaphora resolution in Chinese financial news for information extraction," Proc. 4th World Congress on Intelligent Control and Automation, pp.2422–2426, 2002.
[5] V. Ng and C. Cardie, "Improving machine learning approaches to coreference resolution," Proc. 40th Annual Meeting of the Association for Computational Linguistics, pp.104–111, 2002.
[6] J. Lang, T. Liu, and B. Qin, "Decision trees-based Chinese noun phrase coreference resolution," Student Workshop of Computational Linguistics, 2004.
[7] V. Ng, "Machine learning for coreference resolution: From local classification to global ranking," Proc. 43rd Annual Meeting of the Association for Computational Linguistics, pp.157–164, 2005.
[8] S. Bergsma and D. Lin, "Bootstrapping path-based pronoun resolution," Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp.33–40, 2006.
[9] Y.C. Li, Y. Yang, G.D. Zhou, and Q.M. Zhu, "Anaphora resolution of noun phrase based on SVM," Computer Engineering, vol.35, no.3, pp.199–204, 2009.
[10] J. Peng and K. Araki, "Zero-anaphora resolution in Chinese using maximum entropy," IEICE Trans. Inf. & Syst., vol.E90-D, no.7, pp.1092–1102, July 2007.
[11] R. Bunescu, "Associative anaphora resolution: A web-based approach," Proc. EACL on the Computational Treatment of Anaphora, pp.47–52, 2003.
[12] K. Shinzato and K. Torisawa, "Acquiring hyponymy relations from web documents," Proc. HLT-NAACL, pp.73–80, 2004.
[13] K. Markert and M. Nissim, "Comparing knowledge sources for nominal anaphora resolution," Computational Linguistics, vol.31, no.3, pp.367–402, 2005.
[14] K. Markert, M. Nissim, and N. Modjeska, "Using the web for nominal anaphora resolution," Proc. EACL Workshop on the Computational Treatment of Anaphora, pp.39–46, 2003.
[15] N. Garera and D. Yarowsky, "Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora," Proc. 10th Conference on Computational Natural Language Learning, pp.37–44, 2006.
[16] S.P. Converse, "Resolving pronominal references in Chinese with the Hobbs algorithm," Proc. 4th SIGHAN Workshop on Chinese Language Processing, pp.116–122, 2005.
[17] H.F. Wang and Z. Mei, "Robust pronominal resolution within Chinese text," J. Software, vol.16, no.5, pp.700–707, 2005.
[18] P. Denis and J. Baldridge, "A ranking approach to pronoun resolution," Proc. IJCAI, pp.1588–1593, 2007.
[19] D.S. Wu and T. Liang, "Chinese pronominal anaphora resolution using lexical knowledge and entropy-based weight," J. American Society for Information Science and Technology, vol.59, no.13, pp.2138–2145, 2008.
[20] T. Liang, C.H. Yeh, and D.S. Wu, "A corpus-based categorization for Chinese proper nouns," Proc. National Computer Symposium, pp.434–443, 2003.
[21] C.H. Yu and H.H. Chen, A study of Chinese Information Extraction Construction and Coreference, Unpublished master's thesis, National Taiwan University, Taiwan, 2000.
[22] B.G. Ding, C.N. Huang, and D.G. Huang, "Chinese main verb identification: From specification to realization," International J. Computational Linguistics and Chinese Language Processing, vol.10, no.1, pp.53–94, 2005.
[23] S.J. Russell and P. Norvig, Artificial Intelligence: A modern approach, 2nd ed., Prentice Hall, 2003.

Dian-Song Wu received his M.S. degree in Computer Science from National Chiao Tung University, Hshichu, Taiwan in 2003. Now he is studying toward his Ph.D. degree at the graduate school of Computer Science, National Chiao Tung University, Taiwan. His research interests include machine learning, natural language processing, and web mining.

Tyne Liang received her Ph.D. Degree in Computer Science from National Chiao Tung University, Hshichu, Taiwan. Currently, she is an associate professor of the Dept. of Computer Science, National Chiao Tung University, Taiwan. Her research interests include information retrieval, natural language processing, web mining, and inter-connection network.