Computational Linguistics and Chinese Language Processing
Vol. 16, No. 1-2, March/June 2011, pp. 15-26

Resolving Abstract Definite Anaphora in Chinese Texts

Tyne Liang* and Jyun-Hua Cheng*

Abstract

Anaphora is a rhetorical device commonly used in written texts. It denotes the use of terms referring to previously-mentioned entities, concepts, or events. In this paper, the definite anaphora in Chinese texts is addressed and empirical approaches to tackle abstract anaphors are presented. The resolution is built on the association between target anaphors and the corresponding referents in their multiple-type features extracted from different levels of discourse units. Experimental results show that features extracted from clauses are more useful than those extracted from sentences in referent identification. Besides, the presented salience-based model outperforms the SVM-based model no matter whether the best set of extracted features is employed or not.

Keywords: Anaphora Resolution, Chinese Text, Definite Anaphora, Feature Extraction

1. Introduction

1.1 Motivation

Anaphora is an instance of an expression referring to the preceding utterances. Effective anaphora resolution enhances understanding of a text and facilitates many applications of natural language processing. The resolution involves anaphor recognition and referent recognition. In Chinese texts, anaphors can be missing or be present as pronouns, demonstratives and definite descriptions. Common pronouns are like “他” (“he, him”), “她” (“she, her”), “它” (“it”), “我們” (“we, us”), “他們” (“they, them”); demonstratives are “這” (“this”), ”那” (“that”) and definite description are like the pattern “這+[quantifier]+noun phrase.” Without concerning zero anaphora, about 54% of the explicit anaphors are pronouns, 40% are definite descriptions, and 6% are demonstratives in a corpus containing 20 news articles.

Essentially, the challenges involved with Chinese anaphora resolution are attributed to the complexities of Chinese sentence structures. It is known that although a Chinese sentence

* College of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.
E-mail: tliang@cs.nctu.edu.tw; sunrise0406.iit97g@g2.nctu.edu.tw

features the subject-verb-object order, the sentence may be formed by a series of verbs or by pronoun or subject dropping, thus making sentence parsing difficult. Moreover, there is no blank space between adjacent words in Chinese sentences, making word or noun phrase identification difficult. Unlike most previous research projects focusing on non-abstract anaphora resolution, this paper addresses the definite anaphora in Chinese written texts and presents empirical parser-free approaches to resolve abstract anaphors, like ”這項方案” (“this plan”). The resolution is based on the linking between anaphors and their referents in multiple aspects of contextual, semantic and surface features. Among them, semantic features are extracted with the help of three outer resources, namely, Tongyici Cilin[1] (TYCC for short), CKIP Lexicon[2], and Google search results[3]. Additionally, the features extracted from different discourse units are investigated and the best feature set is verified at referent identification. In the experiments, both SVM-based and salience models are implemented for model comparison. Experimental results show that the features extracted from clauses are more useful than those extracted from sentences for anaphora resolution. Besides, the presented salience-based model outperforms the SVM-based model regardless of whether the best set of extracted features is employed or not.

1.2 Abstract Definite Anaphora

In Chinese texts, a definite anaphor contains a demonstrative (tagged as “Nep” by CKIP Chinese word segmentation system[4] (CKIP tagger for short)) followed with an optional quantifier (tagged as “Nf”) and a noun phrase. Lexicons with Nep-tag are like ”這, 此, 其, 那, 什麼, 其中, 個中, 甚, 啥, 哪, 斯, 甚麼”. Such anaphora is similar to the definite description anaphora in English texts in which the anaphors are composed of the definite article “the” followed by a noun phrase. In fact, there is no definite article in Chinese, so we may treat the definite noun “the+noun phrase” and demonstrative noun “this or that+noun phrase” to be the same in Chinese texts. In this paper, we focus on the “這+[quantifier]+ abstract-type noun phrase” anaphor since it is frequently expressed in Chinese texts. The abstract noun phrases are defined and categorized according to CKIP Lexicon. Table 1 shows some target anaphor instances we identified from our corpus.

Abstract definite anaphora can be expressed in two ways. One is direct anaphora in which both the referent and the anaphor contain the same head noun. For example, both anaphor ”這項方案” (this plan”) and its referent ”學生停車方案” (“student parking plan”)

[1] TongyiciCilin extended version: http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

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

[3] Google: http://www.google.com.tw

[4] CKIP Chinese word segmentation system: http://ckipsvr.iis.sinica.edu.tw/

contain the head noun “方案”. The other is indirect anaphora in which the anaphor “這項方案” and its referent “課後輔導” (“after-school assistance”) do not contain the same head noun and their resolution has to be done by considering their linking in contextual, syntactic and semantic structures. More challenges associated with indirect anaphora resolution come from the boundary identification for those referents crossing multiple clauses or sentences. For example, “這項方案” refers to “學校計劃將強制要求所有過重學生放學後都要留下來，參加二小時的體能訓練，直到學期結束。” (“The school will require all overweight students to remain after school to participate in two hours of physical training until the end of the term”). Besides, it is observed that Chinese texts are usually not written with accurate usage of punctuation marks; thus identifying such types of referents in Chinese texts is harder than in English texts.

Table 1. Some abstract instances and their CKIP Lexicon categories

Category	Example
特徵 (Characteristics)	想法 thought, 行爲 behavior
文明 (Enlightenment)	問題 problem, 決議 decision
法則 (Principles)	方式 way, 制度 system
社會活動 (Social_activities)	比賽 competition, 會議 meeting
法人 (Corporation)	社會 society, 學校 school
名稱 (Nomenclature)	職位 Job, 名字 name
狀況 (Situations)	情況 situation, 現象 phenomenome
社會關係 (Social_relation)	關係 relation, 情誼 frienship
財務關係 (Monetary_relation)	經費 funding, 收入 income
權力 (Authority)	政權 regime, 主權 sovereignty
疾病 (Illness)	病 disease, 病變 lesion
時間 (Temporal_relation)	期間 period, 階段 stage
事件 (Events)	行動 action, 過程 process
動名詞 (nominal verb)	調查 investigation, 會談 meeting

2. Related Work

In general, abstract anaphora can be resolved using pattern rules, statistical or hybrid approaches. For example, Byron (2002) presented PHORA to resolve pronouns referring to abstract entities in a dialogue corpus. The resolution is based on the semantic constraints imposed on verbs, predicate noun phrases and predicative adjectives in sentences. Later

Navarretta (2004) extended these semantic constraints with dialogue structures to resolve inter-sentential pronominal anaphors in Danish texts. Beside the rule-based approaches, Strube and Muller (2003) presented a decision tree resolution to identify both NP-type and non-NP antecedents with the employment of 23 features including NP and coreference features. The supervised learning method is also found in (Yang *et al.*, 2006) for pronoun resolution by taking into account the coreferential information of a candidate. In addition, some researches have tried to resolve indirect nominal anaphora via web search (Bunescu, 2003), WordNet (Poesio *et al.*, 2002), or statistical models like multi-layer perceptrons (Poesio *et al.*, 2004).

In contrast to the prevalent discussion on English anaphora, effective approaches to tackle Chinese abstract anaphora have not been widely discussed. Either parsing-tree based or machine based approaches have been primarily presented to resolve noun-phrase type references (Yeh & Chen, 2004; Zhao & Ng, 2007; Wu & Liang, 2008, 2009, 2011; Kong & Zhou, 2010). Nevertheless, it is observed that the average length of the referents includes more than one clause in a corpus like the news reports we extracted from Academia Sinica Balanced Corpus[5] (ASBC for short). Hence, this paper is motivated to present some feasible methods to facilitate such type referent identification.

3. Corpus Preparation

3.1 Corpus Tagging

The corpus we used for developing the resolution approach is extracted from ASBC, a corpus used for modern Chinese text processing research. For each extracted text, we manually tagged the target anaphors and filter out those "這" without a following noun phrase. For example, we would not use"這是" ("this is"), "這可能是" ("this might be"),...etc. We did not tag those "這" if that functioned as discourse-new or cataphor.

The corpus contained 885 texts and out of which there were 24062 sentences and 82783 clauses identified by any of punctuation marks ("。？！；") and ("。？！；，") respectively. Each clause was tagged with a sentence number s_i and a clause serial number c_j. Each clause was also manually tagged with < ana_i> or < ref_i> if an anaphor *i* or a referent for anaphor *i* was found in that clause. There were total 1538 definite abstract anaphor instances occurring in the corpus. Followings are three tagged examples in which referents were shown in italic form and anaphors were shown with underlines.

Example a：淡大自本學期開始<s_1, c_1>，實施*學生收費停車方案*<s_1, c_2><ref_1>。<u>這項收費停車方案</u><s_2, c_3> <ana_1>，…。

Example b：*天然氣*<u>這項乾淨的能源</u><s_1, c_1><ana_1 ><ref_1> …。

[5] Academia Sinica Balanced Corpus: http://www.sinica.edu.tw/SinicaCorpus/

Example c：*學校計劃將強制要求所有過重學生放學後都要留下來*<s_1, c_1><ref_1>，*參加二小時的體能訓練*<s_1, c_2><ref_1>，*直到學期結束*<s_1, c_3><ref_1>。這項方案已經校務會議通過<s_2, c_4> <ana_1>, 將在九十年學年開始實施。

Table 1 lists the 14 categories defined by CKIP Lexicon and some instances identified in our corpus. There are some observations from our tagged corpus. First, there might be multiple referents for a tagged abstract anaphor. In our corpus, 25% of the anaphor instances referred to more than one referent. 59% of the tagged referents contained more than one clause. 52% of the referents occurred in three clauses away from their anaphors and more referents occurred in the preceding sentences than the ones in the same sentences. Besides, 90% of the addressed anaphors and their referents were far away in three sentences. In this paper, the tagged referents were should be in consecutive clauses if they are referred to the same anaphor.

3.2 The Target Referent

Table 2 lists the statistical data of tagged anaphors and their referents. It is found that almost one-third of the tagged anaphors are either characteristics-type or enlightenment-type. The referents for situation-type anaphors contain more clauses than the ones for other types.

Table 2. Anaphors counts and referents lengths

Categories	Anaphors	Average referents length (in clause)
Characteristics	350	2.23
Enlightenment	205	2.79
Principles	231	2.44
Social_activities	106	1.21
Corporation	48	1.46
Nomenclature	12	1.09
Situations	130	3.34
Social_relation	9	2.7
Monetary_relation	28	2.71
Authority	5	1.4
Illness	11	1.73
Temporal_relation	65	2.25
Events	155	2.79
Nominal verb ([+nom])	148	2.07

4. The Proposed Resolution

Figure 1 is the presented resolution which involves POS tagging, anaphor recognition, feature extraction, and referent identification with the employment of three outer resources and processing tools, like a CKIP tagger and a well-designed NP-chunker. The three resources are CKIP Lexicon, TYCC and web search result which is a set of words extracted from Google's snippets. All of these resources will be used for semantic computation for identifying both anaphors and referent candidates.

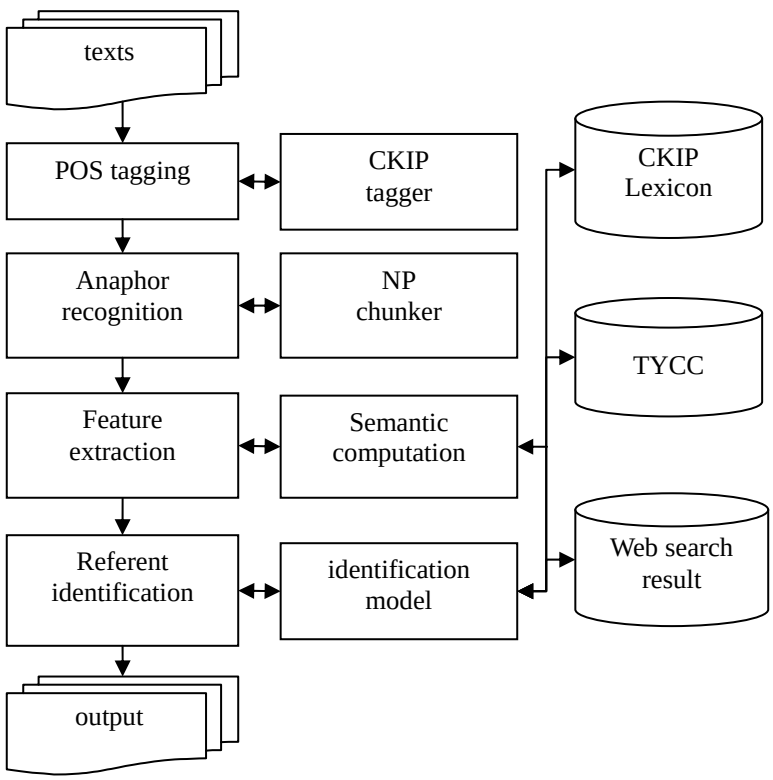

Figure 1. System architecture

4.1 Anaphor Recognition

The anaphor recognition is implemented by a finite-state-machine based NP-chunker (Yu, 2000). Following are some identified anaphors:

(1)這/this(Nep) 個(Nf) 國際/international(Nc) 金融/finance(Na) 中心/center(Nc)

(2)這/this(Nep) 種(Nf) 方式/way(Na) 進行/processing(A) 的(DE) 消費/consuming(Na) 商品/product(Na) 交易/trade(Na)

(3)這/this(Nep) 項(Nf) 國小/elementary school(Nc) 教師/teacher(Na) 心得/report(Na) 公開/publication(A) 發表會/meeting(Na)

(4)這/this(Nep) 個(Nf) 妥協/compromising(VA)

Afterwards, the last word of an anaphor will be extracted as the head noun and will be checked as to whether it is listed or not as an abstract object in CKIP Lexicon. The experimental results on 1538 anaphor instances show that the presented anaphor recognizer can yield 89.99% accuracy. The failures are summarized into three types as follows:

a. Verbal nominalization: our chunker extracts the anaphor "義工" only, rather than "義工培訓" out of "參加/participate(VC) 這(Nep) 次(Nf) 義工/volunteer(Na) 培訓/train(VC)[+nom]"

b. Complex sentence structure: for example, the correct anaphor is "超級省油車比賽", not "國際自動機工程" in the clause "這/this(Nep) 項(Nf)由/from(P)國際/international(Nc) 自動機/automobile(Na) 工程/engineering(Na) 學會/academic society(Nc) 中華民國/ROC(Nc) 分會/sub- academic society (Nc) 舉辦/hold(VC) 的(DE) 超級/super(A) 省/economic(VJ)油/oil(Na)車/car(Na)比賽/competition(Na)"

c. Inverted sentence: the correct anaphor should be "事" rather than "事研究院方面" in a sentence like "這(Nep) 事/thing(Na) 研究院/research institute(Nc) 方面(Na) 也/too(D) 漫無頭緒(VH)/with no idea about".

4.2 Feature Extraction

It is found that 90% of the referents in our training corpus are within the distances of three sentences away from their corresponding anaphors. So the clauses within this distance are considered as candidate referents. Candidates also include the clauses like "中東這個地區" (Middle-east this area). Here referent "中東" (Middle-east) is in the same clause of its anaphor "這個地區".

Table 3 lists the four types of the 10 features used in referent recognition. Among them, the thresholds for the distance and similarity features are measured by chi-square test so that each feature value is either one or zero. *Dice_Coefficient* is used to compute the semantic similarity between the words in candidate clause C and the words in anaphor A by measuring how many common nouns, proper nouns, location names, temporal lexicons and verbs are in common.

Table 3. The extracted features

Feature type	Feature description C: candidate clause, A: anaphor clause
Location	C and A are in the same sentence
	C and A are in the same clause
Distance	C and A are within the threshold distance in the terms of sentences
	C and A are within the threshold distance in the terms of clauses
Lexicon	C contains all words of A
	C contains some words of A
	C contains verbs occurring in A
Semantic	C and A are similar enough by computing *Dice_Coefficient*
	C and A contain the same sentential topic word
	C contains the words frequently occurring in text

$$\text{Dice\_Coefficient} = \frac{2|C \cap A|}{|C| + |A|}$$

$$|C \cap A| = \sum \text{Related}(c_i, a_j)$$

$$\text{Related}(c_i, a_j) = \begin{cases} 1, \text{ if } c_i = a_j \\ \quad \text{or CKIP}(c_i) = \text{CKIP}(a_j) \\ \quad \text{or TONYI}(c_i) = \text{TONYI}(a_j) \\ \quad \text{or } c_i \text{ in web(A)} \\ 0, \text{ otherwise} \end{cases}$$

C：set of words in candidate clause

A：set of words in anaphor

CKIP(x)：CKIP label for word x

TYCC (x)：TYCC label for word x

Web (C)：search result words of C

It is noticed that the computation is based on word expansion using the mentioned TYCC, CKIP Lexicon and the words extracted from web search results. TYCC contains 77270 words, each of them being represented with one code and tagged with five labels, representing five levels of word categorization. We use the chi-square test to select an appropriate category level of words for word expansion. CKIP Lexicon contains 14935 words as abstract-type

words and the words of the same type are treated as related words. For those words neither in TYCC nor in CKIP Lexicon, they will be expanded using search results. The expansion is built with the employment of anaphors and their sentential-topic words as queries to the search engine, Google. From the retrieved 100 snippets with respect to each query, we used chi-square test to find those words frequently co-occurring with the queries.

The topic word feature is employed by assuming that an anaphor and its referent may address the same topic in neighboring sentences. The sentential topic words are identified using the centering-theory based method (Pan, 2008) with which 76.59% F-score was yielded on an experiment containing 88 sentences. The employment of the feature of frequent words is based on the assumption that main concepts in an article may be mentioned repeatedly. In this paper, the main concept words are selected by evaluating the occurrence frequencies of those nouns and verbs in an article. The number of frequent words is also decided by chi-square test.

4.3 Referent Identification

We randomly selected 708 articles containing 1226 target anaphors as our training corpus and use the remaining 171 articles (containing 241 target anaphors) as the testing corpus. The candidate referents are those clauses in the distance of three sentences ahead of the anaphors. The referent identifier is implemented with a statistical model and a salience model for model comparison. All referents are searched backward from the target anaphors. For identification comparison, we implemented both SVM-based (LIBSVM[6]) and salience-based approaches on different discourse units, namely, single-clause, bi-clauses, and single sentence. Both models are incorporated with feature extraction which yields an optimal set of features and feature weights by running PyGene[7], a genetic algorithm tool in Python. The performance is measured in terms of accuracy which is the ratio of the number of referents tagged correctly at their sentential boundaries by the presented model to the number of referents tagged manually.

Table 4 lists the results of different identification models. It is observed that the salience-based model outperforms the statistical model in terms of higher accuracy. This is because the salience-based model is aimed at selecting the candidate that is the most relevant to the corresponding anaphor while the statistical model picks the relevant one only. On the other hand, clause-level approaches turn out to yield higher accuracy than sentence-level approaches. This is because there might be more irrelevant information acquired from larger discourse-units, like bi-clauses or sentences, thus affecting the selection of the right candidates.

---

[6] The LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/.

[7] PyGene: http://www.freenet.org.nz/python/pygene

Table 4. Results of different identification models

Model	SVM-based			Salience-based	
Feature extracted	Single clause	Bi-clause	Single sentence	Single clause	Single sentence
Same sentence	1	1	1	0.1	0.1
Same clause	0	0	0	0	0
Sentence distance	0	0	0	0.1	0.1
Clause distance	0	0	0	0	0
Full lexicon agreement	1	0	0	0.1	0.1
Partial lexicon agreement	1	1	1	0.2	0.2
Same verb	0	0	0	0.1	0.1
Clause similarity	1	1	1	0.1	0.1
Same topic word	0	0	0	0	0
Frequent words	1	0	1	0.2	0.2
accuracy	68.46	65.14	53.65	70.54	60.34

Some failures in identification are attributed to the errors in noun phrase chunking. For example, in the text "『天才是九十九分的努力加一分的才氣』這種話銘記在心…" ("Genius is ninety-nine-point hardship plus one-point talent" such saying should be memorized in mind")，the anaphor "這種話" ("such saying") refers the idiom "天才是九十九分的努力加一分的才氣". Such failure may be resolved by taking into account the punctuation marks as one useful feature at referent identification. One the other hand, the present resolution is unable to identify the semantic association between "大清帝國" ("Qing Empire") and "這個時代" (this era) in resolving the anaphor in the text like "在大清帝國這個時代中…" ("In this era of Qing Empire…"). How to improve the presented semantic computation should be concerned as the future work.

5. Conclusion and Future Work

In this paper, we describe definite anaphora in Chinese texts and present empirical methods to resolve the target abstract anaphors which are not widely addressed in previous research projects. In addition, we consider the factors of discourse levels from which the feature extraction is implemented. Without the help of a parser, our experimental results show that clause-level feature extraction is better than the sentence-level extraction in generating useful

identification features. Besides, the salience-based approach yields higher accuracy than the SVM-based model whether the best set of extracted features is selected or not.

In the future, we will take into account the Chinese discourse structure and discourse markers in order to improve the boundary identification especially for those referents containing multiple clauses or sentences. Besides, improvement of the semantic computation model should be made so as to enhance the semantic linking between anaphors and their corresponding referents.

References

Bunescu, R. (2003). Associative Anaphora Resolution: A Web-Based Approach. In *Proceedings of EACL 2003 workshop on The Computational Treatment of Aanphora*, 47-52, Budapest.

Byron, D. K. (2002). Resolving Pronominal Reference to Abstract Entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 80-87.

Kong, F., & Zhou, G. (2010). A Tree Kernel-based Unified Framework for Chinese Zero Anaphora Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 882-891.

Navarretta, C. (2004). Resolving Individual and Abstract Anaphora in Texts and Dialogues. In *Proceedings of the 20th International Conference of Computational Linguistics (COLING)*, 233-239, Switzerland.

Pan, S.-C. (2008). *Sentence-based Topic Identification and Its Applications in Chinese Texts*, Master thesis, National Chiao Tung University, Taiwan.

Poesio, M., Ishikawa, T., im Walde, S. S., & Vieira, R. (2002). Acquiring Lexical Knowledge for Anaphora Resolution. In *Proceedings of the 3rd Conference on Language Resource and Evaluation (LREC)*, Las Palmas.

Poesio, M., Mehta, R., Maroudas, A., & Hitzeman, J. (2004). Learning to Resolve Bridging References. In *Proceedings of Annual Conference for Association of Computational Linguistics*, 143-150.

Strube, M., & Müller, C. (2003). A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. In *Proceedings of the 41st Annual Meeting of Association for Computational Linguistics(ACL)*, 168-175.

Wu, D.-S., & Liang, T. (2008). Improving Chinese Pronominal Anaphora Resolution Using Lexical Knowledge and Entropy-based Weight. *Journal of the American Society for Information Science and Technology*, 59(13), 2138-2145.

Wu, D.-S., & Liang, T. (2009). Zero Anaphora Resolution by Case-based Reasoning and Pattern Conceptualization. *Expert Systems with Applications*, 36(4), 7544-7551.

Wu, D.-S., & Liang, T. (2011). Improving Definite Anaphora Resolution by Effective Weight Learning and Web-Based Knowledge Acquisition. *IEICE Transactions on Information and Systems*, E94-D(3), 535-541.

Yang, X., 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*, 41-48.

Yeh, C.-L., & Chen, Y.-C. (2004). Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing*, 17(1), 41-56.

Yu. C. H. (2000). *A study of Chinese information extraction construction and coreference*, Master 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*, 541-550.