# INFORMATION TECHNOLOGY JOURNAL

# Discourse Anaphora Resolution Strategy Based on Syntactic and Semantic Analysis

[1]Yan Qiang, [1]Ruiliang Ma and [2]Wei Wei
[1]College of Computer Science and Technology,
Taiyuan University of Technology, Taiyuan 030024, China
[2]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

**Abstract:** With the dramatic development of computer network technique, different kinds of information have been enhancing quickly and the requirements for precisely located information gives a strong impetus to the NLP research. Natural Language Processing (NLP) research on the subject that how to make computer understand and generate unrestricted natural language text,such as Chinese, English,etc. It brings computer the understanding ability of natural language meaning and responding the questions persons asked in natural language. Its purpose is building a friendly relationship between human and machine; enabling computers communicate information in a high rate. As an important method for discourse connecting and consistency, the employ of anaphor makes discourse looks brief and anaphora reflects the semantic relations between sentences and makes a discourse be a discourse.

**Key words:** Natural language processing, anaphora resolution, framework for semantic annotation.

## INTRODUCTION

Early studies abroad anaphora resolution rely mainly confined to the simple rules of calculation revelation, 1970s and 1980s began to enter a variety of research and development programs or modes. These studies make full use of vocabulary, syntax, semantics, discourse and even help encyclopedic knowledge and other knowledge (Wang and Guan, 2005). These years of study tends to theoretical, researchers types and application of knowledge inquiry showed full interest and aspirations. The late 1990s, researchers consciousness anaphora resolution of high complexity and the importance of developing systems of work, More practical research, a number of landmark achievements also appeared, such as Lappin and Leass to refer to digestion procedures, Mitkov refer to eliminate system, Baldwin refer to the digestion process, high precision and so on. In recent years, as people refer to the natural language study enthusiasm rising, bilingual or multilingual refer to eliminate computing research showing a strong momentum (Wang, 2006). In addition, the other main achievements concentrated expression in three aspects: the probability and the application research of machine learning technology, the continuous research and application of the theory of "center" and to refer to the assessment of the related calculation method of resolution.

In China, the last ten years, anaphora resolution of NLP research has become a hot issue, has been widespread concern. Chinese anaphora resolution studies also began to be seriously. However, compared to many other languages, especially English anaphora resolution, the Chinese anaphora resolution research can only be counted as just getting started, there are many problems. Main research focus in Peking University, Beijing Normal University, Beijing University, Harbin Institute of technology, Xiamen University and other colleges and universities laboratory, appear based on decision tree, radial, based on maximum entropy model, based on the theory of text representation theory, based on the priority of selection strategies and so on, to achieve anaphora technical course (Fong, 2009). However, these studies beyond the experimental stage could not form a stable system.

Thus, foreign research anaphora resolution unusually rich domestic anaphora resolution research is still in its infancy. In view of this, the project's research will focus on two aspects: On the one hand, the development of the Chinese refer to tagging corpus study; On the other hand, aiming at the particularity of Chinese and Chinese discourse, proposes a set of specific implementation in Chinese discourse refers to digestion. This will play a domestic role in promoting the study of NLP, which is the applicant plans to carry out this study, the starting point.
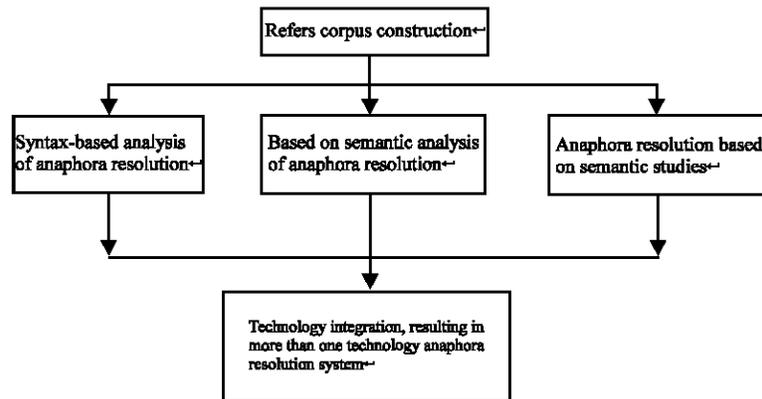
**Corresponding Author:** Yan Qiang, College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China

Fig. 1: Chinese discourse anaphora resolution overall block diagram

## BASED ON SYNTACTIC AND SEMANTIC ANALYSIS OF ANAPHORA RESOLUTION SYSTEM

Based on corpus refers to the labeling of construction, this study studies refer to eliminate algorithm based on syntactic, semantic; Then make a series of technical integration and functional integration, so as to realize a Chinese chapter refer to eliminate model (Hu, 2009). The overall block diagram as shown in Fig. 1.

## HNC CONCEPT SEMANTIC STRUCTURE

HNC theory, natural language understanding space from language to language concept space mapping process, the concept of human perception is the process of language to activate, extension, the whole process of enrichment, transformation and storage operations. HNC theory, the natural language statements basic types of mid-class expression is limited, is exhaustive, a total of 57 groups. This expression can be divided into 57 groups seven sentence and its subclasses, called basic sentence. The seven basic categories for the sentence: the role of sentence (X), during a period (P), the transfer period (T), a period effect (Y), between a period (R), the state sentence (S), determines a period (D). Each class has its own basic sentence late class expressions can be mixed between basic sentence form different statement expression, a total of 3192 group.

The semantic of semantic composition unit is a statement to the next level. This means that two points: first, semantics is the semantic composition unit of the sentence, rather than a syntactic constituent units; Second, semantic block is the sentence semantic composition unit directly, rather than indirect constituent units. The second point is necessary because the semantic meaning is the structural unit, the sentence is not pose a direct semantic unit

(Xu and Wang, 1999). To create the semantic structure of said type, the semantic concept of a block is required.

According to the sentence of the importance of basic semantic information. Semantic block divided into two categories based auxiliary, the main semantic block is the sentence semantics necessary backbone component, auxiliary semantic block is the sentence semantics optional minor ingredients, auxiliary block symbols collectively FK. The establishment of the semantic structure representation formula basically only need the main semantic blocks. Main semantic block has four kinds of primitive types, respectively is object semantics of semantic characteristics of block A, B, function semantics of semantic block C E and content. B, A and C are collectively referred to as generalized object semantic chunk, represented by symbol GBK.

Sentence category expression describes the basic semantics of the sentence constituted. Blocks from the main semantics, such as the role of sentence "XJ = A+X+B". In the sentence, semantic block can block expansion (semantic block extended sentence), mid slough (semantic block consists of sentences from degenerate), while block expansion (sentence slough) may contain a semantic block in righteousness, so that it will form a semantic block nesting hierarchy.

More about sentence category and semantic knowledge of the block of the personal pronoun anaphora resolution has a very important role in guiding, can determine a sentence within a block level and the semantic level, which according to statements hierarchy principle of parallelism and semantic block digestion pronouns (Zhang and Zhou, 2002).

## CORPORA

With other machine learning -based natural language processing techniques identical anaphora resolution also

need to mark a good corpus resources. Currently, the relationship is marked with the more commonly used to refer to a corpus of MUC and ACE. The following details the corpus used in this article and its label format.

## MUC CORPUS

MUC (Message Understanding Conference) is the U.S. government supported a meeting dedicated to understanding the real text of the conference from 1987 to 1998 held seven sessions, is responsible for the different units from around the world to understand the system message systematic evaluation. 1996 of the MUC-6 and 1997 MUC-7 meeting, anaphora resolution to become one of the important tasks MUC evaluation. In this study, MUC-6 corpus for training and testing, take the MUC. 6 of 30 documents as training documents, test with MUC. 6 of 30 Formal standard test documents.

MUC middle finger marked the introduction of two generations relations marked right, using SGML mark:

- 1) with (COREF ID = "I") in the left margin indicates an entity with (/COREF) indicates the right margin
- With (COREF ID = "J" REF = sample M") represents the left edge reference expression, the same (/COREF) indicates the right margin

Where I, J represents the sequence number, in one chapter, the serial number starting from 1 strictly increasing, J and M represents the number of antecedents. If M = I is the physical ID of J to point I of the noun phrase number. Only one entity ID no REF, the description for this noun phrase noun phrase does not point to anything else, but only pointed to by another noun phrase (Wang, 2000).

## ACE CORPUS

ACE evaluation brewing since 1997, started in December 2000. And from ACE2004 began, joined the Chinese corpus. ACE corpus mainly from the radio news (40%), newswires (40%) and network dialogue (20%), so the corpus subdivided into three sub- corpus : BNEWS, NStudy, NWIRE.

ACE corpus depending on the year, the study used has ACE2003, ACE2004 and ACE2005. ACE and MUC corpus annotation methods differ. The following describes the number of documents and ACE corpus annotation format. ACE2003 corpus document is divided into two parts, train and test, ACE2004, ACE2005 corpus without distinction. For ACE2004, ACE2005 corpus, whichever article in the experiment before 4 /5 document training, after 1/5 tests.

**The project source corpus size:** 121 full article, covering 14 areas, the article from the Foreign Language Teaching 80, online choice 41.

## SEMANTIC CATEGORY

1989 edition of "Ci Hai" semantics definition is: "The meaning of the expression language that is the meaning of the language contained in the content." Semantic information manifested in many forms, such as can be abstracted as the relationship between the two words, several word in the sentence syntactic relations, semantic category of a word, a sentence and recurrent semantic similarity between the two words (Li, 2005). If you can get this information correctly, the system performance improvement anaphora resolution is helpful. Researchers semantic information has been added to anaphora resolution, it is also achieved some results.

In September 1995, at the Sixth MUC meeting, introduced the named entities (Named Entity, NE) evaluation tasks. MUC will be named entity is defined as: proper nouns and people are interested in a particular quantifier, according to the MUC definition, someone named entity name (Person), place names (Location), organization name, date, time, percentage, currency (monetary value) these seven categories. Named entity is the text of the basic elements of information is the basis for correct understanding of the text. Narrowly speaking, named entity is a real-world concrete or abstract entities, such as people, organizations, companies, locations, etc., usually with a unique identifier (distinguished name) said that if the person's name, organization name, company name, place names. Broadly speaking, named entities can also include the time, the number of expressions and so on (Zhou, 2007). As to the exact meaning of the named entity, can only be determined according to the specific application. For example, in the specific application, you may need to address, e -mail address, telephone number, ship number, conference name, etc., as named entities.

ACE (Automatic Content Extraction) is the United States government responsible for the evaluation of an information extraction research projects. The evaluation for the named entities are defined as follows:

- **People (person):** People by name (Yao), position(butcher),family relationships (father), pronouns (he), etc. or by a combination thereof specified
- **Organizations or institutions (organization):** including the business sector, government departments, industry, sports teams
- **Geography/social/political entity (GPE):** A name used to represent geographic area can also be used to represent the government or the people there

(such as France), as a result of the three categories. Typical GPE is a national, state and city. If a buried region is not associated with a government, it is probably the premises entity

- **Premises (Location):** premises entity is limited to geographical entities such as geographic region, large tracts of land, water bodies, astronomical, geological structure and so on. For example, " solar ", " Jupiter ", " Europe ", " Middle East ", " Yangtze River. "
- **Facilities (facilities):** A large, mainly artificial functional structures, including houses, factories, stadiums, office buildings, a gym, prisons, museums, space station, warehouses, parking lots, hangars, streets, high-speed roads, ports, railway stations, bridges, tunnels and so on
- **Transportation (Vehicle):** A physical device is designed primarily used to move an object from one location to another location. Such as trains, bicycles, etc
- **Arms (Weapon):** a physical equipment, mainly used to harm or destroy the animal, human, buildings or other items of equipment. Such as bombs, daggers and so on

## CHINESE CHAPTER OF PRETREATMENT

Chinese across documents refer to eliminate the premise is to get as much information as possible from the corpus of the original. In order to obtain this information, as pretreatment process in Chinese across documents refer to eliminate the necessary condition of pretreatment process includes sentence recognition, the named entity recognition, participles labeling process, etc. In this study, we give the pretreatment process of Chinese across documents refer to eliminate system as shown in Fig. 2:

- Preprocessing stage is mainly obtained from the corpus anaphora resolution tasks grammar, syntax, semantics and other information. At this stage, the first born extracted from the corpus (plain text) and then after a clause, word, named entity access to information, tagging, noun phrase and get the word center, semantic role access to information and other steps, the final given after pretreatment results

## SENTENCE RECOGNITION AND SEGMENTATION

Chinese original for a given corpus, a pretreatment step sentence recognition. Correctly identify sentence boundaries, separating words and punctuation will help further application of various natural language processing tools to do their subsequent analysis and processing.

Sentence recognition is carried out the first step of the task anaphora resolution, extracted from a corpus of plain text message, sentences and between sentences are linked. At this stage, mainly on the extracted information as plain text clauses operation, the text appears in a period, question mark, exclamation mark as a sentence separator, a separate line for each word into text and then, on divided sentence text segmentation operation. Different Chinese and English, in English, between words are separated by spaces and in Chinese in words and words are linked together, no delimiters, so the need for a sentence for segmentation operation and the results of segmentation directly affect the digestion of items to be identified accuracy of anaphora resolution greater impact on performance, so this article is used in the experiment corpus annotation standard segmentation results.

## NAMED ENTITY RECOGNITION

Named Entity Recognition (NE) is the recognized text entities that have specific meanings, including names,
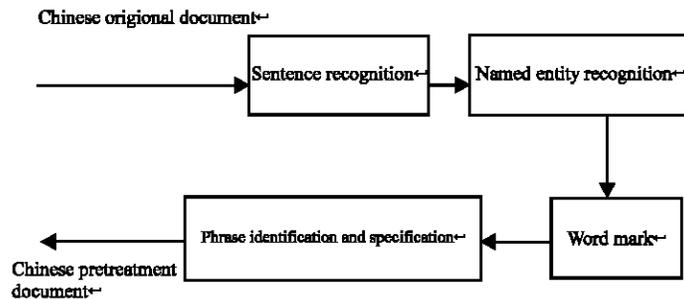


Fig. 2: Pretreatment processes

places, names and other institutions such as the entity name. Named entity recognition task is to recognize text refers to an entity that has a specific meaning, including place names, organization names and other proper nouns. In today's world, with the popularity of computers and the rapid development of the Internet, a lot of information presented in the form of electronic documents in front of people. In response to the serious challenges posed information explosion, there is an urgent need some automated tools to help them in the flood of information sources to quickly find the really important information (Wang, 2002).

Named Entity Recognition is information extraction, question answering system, syntactic analysis, machine translation and other applications important foundation tool in the natural language processing technology to practical use in the process occupies an important position.

In general, the named entity recognition task is to identify the text to be processed three categories (entity classes, time classes and digital classes), seven subcategories (names, organization names, place names, time, date, currency and percentage) named entity.

Named Entity Recognition process usually consists of two parts: (a) the entity boundary identification; (2) determine the entity (names, places, names, or other organization).

English named entities have more obvious forms of signs, so the physical boundary identification is relatively easy, the task is to determine the focus of the entity classes. And English compared to Chinese named entity recognition task is more complex and relative to the entity annotation subtasks, the entity boundary identification more difficult. Chinese named entity recognition difficulties mainly in:

- Chinese text does not like the English text, such as spaces explicitly marked word boundary identifier, named entity recognition the first step is to determine the boundaries of the word, that word
- Chinese word segmentation and named entity recognition affect each other
- Entities defined in addition to English, translation of foreign names and place names are present in the Chinese translation of two special entity type
- Modern Chinese texts, especially the network of Chinese text, often used interchangeably in English, then Chinese named entity recognition task which also includes the English named entity recognition
- Different naming entities have different internal characteristics, it is impossible to use a uniform model to describe all of the entity's internal

characteristics. In the named entities, the time words and quantifiers relatively easy to identify the existing general rule-based approach; entity name (names, places and organization name) recognition is the focus of research. This software tool is mainly for Chinese names (including the names referred to), place names (including names referred to) and launched special research organization name recognition, the use of large particle size characteristics (POS Feature) and small particle size characteristics of the combination of statistical models and expert knowledge, combined with Chinese named entity recognition model

## WORD MARK

Word mark refers to a character sequence cut into a one single word and identify parts of speech. Named entity anaphora resolution information is an important part, is a word very critical information elements, the semantic categories consistency in anaphora resolution filtering stage and feature selection stage has very important significance (Yang, 2008). MUC meeting in accordance with the definition of named entities can be divided into names, place names, organization name, date, time, percentage, money and other seven categories. Named entity recognition task is proper nouns in the text, including names, places, organization name and time expressions and numeric expressions identify.

As part of speech means the division of parts of speech based on the characteristics of the word. Modern Chinese word can be roughly divided into 12 categories.

Notional nouns, verbs, adjectives, numerals, quantifiers and pronouns.

Function words adverbs, prepositions, conjunctions, particles, onomatopoeia and interjections.

But this division for subsequent semantic analysis is not enough. Thus, different builders will be on the part of speech corpus finer division. Such as " noun" can be further divided into " general terms ", " terminology ", " abstract noun " and " Nouns " and so on. " Terminology" Available subdivided into "names", " name", " organization name " and so on.

Tagging of each word in a sentence is to mark the most appropriate part of speech. Chinese and English, an important difference is that Chinese do not change part of speech. Chinese have a variety of parts of speech for each word (such as " hope" that is a noun and a verb). However, the use of specific situations, such as a sentence, every word has uniquely identify the part of speech. In some parts of speech that candidates select an appropriate part of speech.

Tagging specification there are many, such as Chinese Treebank (CTB) standards. Such as the following example, "DT", "PN", "VV", "AS", "JJ", "DEG", "NN" and "PU" is defined CTB speech label.

For example: Last year / DT he / PN achieved / VV a / AS gratifying / JJ 's / DEG progress / NN.

In this system, the referenced fudanNLP segmentation tool, part of speech is divided into: verbs, Modal verbs, directional verb, the verb is a verb, predicate form, adjectives, adverbs, nouns, orientation words, names, places, organization name, time phrases, email, URL, model name, entity name, interrogative pronouns, demonstrative pronouns, personal pronouns, quantifiers, prepositions, numerals, CONVENTIONS, qualifier, conjunctions, interjections, ordinal, omitting the word, modal, structural particle, tense words, punctuation, onomatopoeia, words such as facial expressions.

Chinese words lack of morphological changes, not directly come from the word morphological changes discriminant word category. And most of the words are ambiguous and class is serious. Chinese tagging to rely more on semantics, the same words to express different meanings, its part of speech is often inconsistent. Therefore, through the dictionary and other simple tagging method effect will be relatively poor.

Currently, effective Chinese tagging methods can be divided into rule-based method based on statistical methods into two categories. Rule-based approach is limited in that the complexity of natural language, the establishment of the rule base requires a lot of expertise and high labor costs. Based on statistical learning methods rely heavily on the limitations of the data set in its scale and quality. In recent years, as people can be obtained at lower cost high-quality data sets, based on statistical learning of speech tagging method has achieved good results and become the mainstream method, commonly used learning algorithm has Hidden Markov Model (HMM), Maximum Entropy (ME), Conditional Random Fields (CRFs) and so on.

Part-of-speech tagging is to resolve the important part of it is the precondition for the identification of noun phrases and concludes that the noun phrase is the basis of the type, such as the part of speech is PN said the word is a pronoun, parts of speech is NR said is a proper noun, the part of speech is NT said is time kind of noun phrases, etc. The basic unit of the noun phrases are sure to eliminate, only to find out the noun phrase in the article, to refer to digestion. Noun phrases may be a single word, also can be some phrase structure. This study processing of noun phrases include not only common noun phrase, pronoun, proper nouns and noun phrases. Center words is the core of the noun phrases, is the main part of the noun phrases, the center word belongs to determine the type of phrases and its parts of speech, under normal circumstances, the central word and modifiers, prepositions etc. Constitute the noun phrases.

## K-NEAREST NEIGHBOR ALGORITHM (KNN ALGORITHM)

K Nearest Neighbor (K-on his Neighbor, kNN) algorithm, is one of the most simple and effective machine learning algorithms. Steps of the algorithm is that for a sample and found in the feature space and it is the most similar (i.e., in the feature space adjacent) of k samples, if most of k samples belong to one category, then the sample also belong to this category (Lang *et al.*, 2007). KNN algorithm, the choice of the neighbors have been the object of the correct classification. The method on the nominal decision making based on only the most adjacent categories to determine one or a few sample for sample points belong to the category of the. This also means that no real sense of KNN algorithm "training" stage., this is a very good judgment method to overcome the defects of the linear inseparable problem is also very suitable for the demand of the classification criteria are subject to change (just delete the old training sample, add new training sample, changed the classification criteria). When k = 1, also known as the Nearest Neighbor (NN) algorithm.

KNN method is equivalent to a nonparametric density estimation methods, only in decision-making related to the very small amounts of the adjacent samples. Because of the limited KNN method mainly by the surrounding adjacent samples, so for class field cross or overlap more nonlinear can score according to it, than other method is more suitable for KNN method.

A lack of KNN is a sample of the category, the need to comparing it with all known types of sample again, this is a considerable computational overhead. For example, a text classification system there are tens of thousands of classes, each class even only 20 training sample, a new sample to determine category, also want to do 200000 times of vector comparison. This problem can be indexed by means of the sample space to compensate.

KNN is another weakness is: when the sample is not balance, such as a class of the sample size is large and other types of sample size is very small, may cause when enter a new sample, the sample of the K neighbors large capacity samples of a class in the majority, lead to classification error.

## TEST RESULTS AND ANALYSIS

Respectively in the process of testing the performance test with different number of words in platform, in order to eliminate the effects of other factors

Table 1: Test results

| No. of articles | MUC (%) | | |
| --- | --- | --- | --- |
| | P | R | F |
| 5 | 67.27 | 73.66 | 70.32 |
| 10 | 66.66 | 73.04 | 69.71 |
| 15 | 66.69 | 74.28 | 70.28 |
| 20 | 65.83 | 74.56 | 69.92 |
| 25 | 65.18 | 74.52 | 69.54 |
| 30 | 65.91 | 74.03 | 69.74 |

on the experimental results experiments required information are obtained from the corpus, the experiment in the standard case.

As can be seen from the Table 1 the result of the experiment, based on the syntactic and semantic analysis of Chinese text to refer to eliminate platforms, the number of corpus to refer to eliminate the influence of the platform is not very big, with different number of corpora testing, platform performance fluctuation is not big. However, due to the pretreatment of the Chinese discourse in Chinese language after system of word segmentation, the named entity recognition, part-of-speech tagging and identification of noun phrases, noun phrases center word, semantic information access, etc. and pretreatment process may not entirely correct, so the results will also exist certain error.

## SUMMARY AND OUTLOOK

Study refer to eliminate problems, this study introduces the refers to the knowledge of dispelling reference commonly used to refer to digestion method based on machine learning, basic features extraction and implements a prototype system.

With the rapid development of computer technology and Internet, all kinds of information exploding, accurate positioning of information to the people's demand to promote the development of natural language processing technology. Refer to digestion is one of the key and difficult point for natural language processing, in this study, machine translation, multilingual text information processing and information extraction and many other applications are used. Refers to the use of pronouns, make the discourse expression do not show white, concise and clear, at the same time, to reflect the relationship between discourse in the semantic relations between each statement, is an important characteristic of discourse. Many of the existing natural language applications, such as text, information extraction, some adopted to directly extract sentences from the text, appeared many sentences without first language, which makes it difficult to understand. Refer to resolve similar problems can be solved without doubt. Also, refer to the phenomenon of processing in machine translation also

plays an important role, the context for subsequent processing, determine the attribute, plurality, determine the component ellipsis and eliminate ambiguity are helpful. Refer to, is to determine the anaphor and the interrelation between the first language, so as to clear the anaphor refer to what is the object. And refer to dissolve and MUC and ACE evaluation of important tasks, so more and more get the attention of the researchers at home and abroad.

Based on the national natural science fund "multilingual across text information extraction technology research" and national "863" plan project "based on machine learning, high-performance adaptive information extracting key technology research", mainly studies the Chinese refer to digestion technique based on syntactic and semantic and in-depth discussions on some related problems was made.

In this study, we obtain the following results:

- Basic realized based on the research of frame semantic annotation, to solve the problem of basic noun phrases to identify and refer to eliminate. The experimental results show that the accuracy of this method still remains to be improved
- The candidate antecedents to identify problems, sentence recognition and segmentation and named entity recognition, improve the precision of refer to eliminate pretreatment system
- Excavated, including syntax, semantic and syntactic information, refer to the digestion, 13 characteristics of each characteristic was analyzed through experiments in the role of refer to digestion and further analyzed the affect refers to eliminate all pretreatment processes

## ACKNOWLEDGMENTS

## REFERENCES

Fong, K., 2009. Anaphora resolution of key issues. Ph.D. Thesis, University of Suzhou, China.

Hu, N.Q., 2009. Eigenvector-based Chinese anaphora resolution research and system implementation. Ph.D. Thesis, University of Suzhou, China.

Lang, J., B. Qin and T. Liu, 2007. Intra-document coreference resolution: The state of the art. J. Chin. Language Comput., 17: 227-253.

Li G.C., 2005. Preferred strategy using personal pronouns Chinese anaphora resolution. Chin. Inform. Technol., 19: 24-30.

Wang, H.F., 2000. Chinese anaphora resolution and omitted restoration. Ph.D. Thesis, Institute of Acoustics, China.

Wang, H.F., 2002. Anaphora resolution of the basic methods and implementation techniques of Chinese. Inform. Process., 16: 9-17.

Wang, X.L. and Y. Guan, 2005. Computer Natural Language. Tsinghua University Press, Beijing.

Wang, Z.Q., 2006. Chinese anaphora resolution and related technology research. Ph.D. Thesis, Beijing University of Posts and Telecommunications.

Xu, M. and C.C. Wang, 1999. Chinese middle-generation study and discuss the problem. Southwest China Normal Univ.: Nat. Sci. Edn., 63: 337-376.

Yang, Y., 2008. Eigenvector-based research noun phrase anaphora resolution. Ph.D. Thesis, University of Suzhou, China.

Zhang, W. and C.L. Zhou, 2002. Chinese discourse understanding meta preliminary anaphora resolution. J. Software, 13: 732-737.

Zhou, J.S., 2007. Graph partitioning based unsupervised Chinese anaphora resolution algorithms. Chin. Inform. Technol., 21: 77-82.