The Anaphoric Expressions of Chinese Algebraic Word Problem

Yuan-Kai Wang, Wen-Lian Hsu, and Yi-Chiou Chen

Institute of Information Science, Academia Sinica,
115, Nankang, Taipei, Taiwan, R.O.C.
{ykwang, hsu, ysc}@iis.sinica.edu.tw

ABSTRACT

Discourse and anaphora analysis can be very important for intelligent human-computer interface. To study the discourse and anaphora issues of Chinese, Chinese algebraic word problem is a good test-bed. This paper classifies anaphoric expressions into four classes: zero anaphora, reflexives, personal pronouns, and pronominal noun phrases. We analyze algebraic word problem through the four classes and show that the problem can be so hard as a complex anaphora issue. We also present the distinguishing characteristics of Chinese algebraic word problem. Through the analysis of anaphora, this paper intends to show that Chinese algebraic word problem can be a typical problem of Chinese natural language understanding.

1. Introduction

Human-computer interactions can greatly assist people by incorporating natural language processing. Some researches on intelligent human-computer interface, like speech recognition, text-to-speech, and content-based information retrieval, are explicit examples that can benefit their performances by considering about linguistic information. One important common characteristic of these researches is that they need to process a sequence of words. Meanings of words are continued, expanded, and elaborated across sentences. The appropriate linguistic processing of the intelligent human-computer interactions thus should consider intersentential context, which is usually referred as discourse analysis [1].

In the last decade or so, discourse has been studied in pursuit of a general understanding of how language is used for the interactional needs of communication. One major issue in discourse analysis is to resolve the reference of object in discourse [2][3]. This problem is called anaphora. Anaphora essentially is the linguistic phenomenon in which an expression is used to relate to an antecedent that appears or is assumed to appear elsewhere in the discourse. It is such a fundamental issue that Charniak [4] said "In order to do pronoun resolution, one has to do everything else." Pronoun is one kind of anaphora and will be discussed in Sections 2 and 3.

Our study adopts a special problem which is called algebraic word problem to help research anaphoric expressions. One algebraic word problem includes several natural language sentences. Sentences of one problem form a coherent discourse that describes similar topics. The coherent discourse contains a question sentence that asks for an answer by performing some algebraic operations. While the word problem is normally simple and restricted in respect to its special domain on mathematical computation, it provides a wide range of linguistics phenomena and can be an excellent test-bed for anaphora. Consider the following example:

Example 1

約翰 ϕ 今年 7 歲, 13 年 後 他 是 幾歲?
Yuehan jin-nian 7 sui, 13 nian hou ta sh jisui?
John this-year 7 years-old, 13 years after he is how-old?
John (is) 7 years old this year. How old will he be after 13 years?

The answer can be obtained by a simple calculation 7+13 = 20. The first sentence in Chinese does not have a verb, which makes a classical parser difficult to generate a parse tree. The pronoun 他(he) in the second sentence is referred to the object 約翰(John) in the preceding sentence. Therefore, to obtain the answer 20, computer not only has to parse the two sentences, but also has to recover the lost verb and find the reference relationship between he and John. The addition of 7 and 13 can then be inferred after computer resolves the above two anaphors.

A question in algebraic word problem describes a unique topic. Sentences within a question thus have topic continuation. One question in algebraic word problem clearly form a discourse.

Algebraic word problem is a good test-bed for studying discourse analysis and anaphora due to the following three reasons. First, since algebraic word problem uses the most common description form of natural language, we can study most types of anaphora through it. Anaphora in algebraic word problem is notably an intersentential problem, but not an intrasentential one. That is, anaphora resolution of algebraic word problem needs to process across sentences. Second, its domain is restricted in algebraic math. This can help us focus on the linguistics issues of discourse and anaphora. Third, the performance of a discourse model and anaphora resolution approach in an empirical study on algebraic word problem can be easily verified through numerical answers. In this paper, we will discuss the anaphora types revealed in algebraic word problem. We also discover some characteristics of the problem. The next section will first discuss the classification of anaphora.

2. Anaphora in Chinese

According to the definition of Random House dictionary, anaphora is “the repetition of a word or words at the beginning of two or more successive phrases, verses, clauses, or sentences.” However, anaphora is more complicated than the above definition [5]. Anaphora is the phenomenon of the use of a *word* as a regular grammatical substitute for a referred *object* or group of objects. The referring expressions are called anaphor. Since the referred objects in anaphoric expressions usually precede anaphors, they are called antecedents.

In an anaphoric expression, there may not have an antecedent that is identified as the explicit object to be referred. Sometimes, an anaphor will refer to a group of objects, but not a group of words, around the expression. Many researchers of natural language processing considered anaphora resolution as a critical issue in natural language understanding [2].

Anaphora in English is studied and discussed in many works. Chinese is an isolating language with no morphological markings of case, gender, or number agreement in its syntactic structures. In [6], the major differences between English and Chinese is pointed out to lie in the fact that Chinese is characterized as basically a context dependent language, whereas English is a structural dependent language. Zero anaphora is arisen due to the difference. In English, even when the anaphoric pronouns can be easily understood out of context, English still requires the presence of those pronouns to complete a clause in a discourse. In fact, the structural completion of a sentence is so important in English that the language has to resort to a dummy pronoun it to fill in the slot of the grammatical subject.

There is some works in Mandarin Chinese, whereas most of them came from linguistics. Few provided simple computational models to resolve part of anaphora phenomena [7][8][9]. Among those works, there is no regular way to evaluate their results.

The following will discuss anaphora in Chinese based on the classification proposed by Huang [10]. We modify the classification into 4 classes that consists of zero anaphora, reflexives, personal pronouns, and pronominal noun phrases. Pronominal noun phrases include definite NP, possessive NP, and bare NP. This classification is also notably appropriate for English, while the usage frequency of each class in English may be different with in Chinese.

2.1 Zero anaphora

In most of discourses much of what is uttered is not explicitly stated. Object may be implicit in the sentence of the communication, and our computer needs some general rules of inference to find out. This is called ellipsis or zero anaphora which is a syntactic “hole” in the sentence where an antecedent is

understood but not explicitly mentioned. Zero anaphora is an important linguistic expression. It functions to make coherent and concise discourse segments and to prevent discourse from having redundant linguistic entity.

Zero anaphora is such a common linguistic device in Chinese that it may occur in almost any syntactic position in the sentence where a noun or a pronoun could appear. Zero anaphora is one of the most prominent features that distinguish Chinese from English [6]. Chinese exhibits heavy use of zero anaphora without any overt grammatical marking to indicate the missing nouns. Sentences can be ambiguous and are subject to different interpretation. However, in the discourse context, sentences with many uses of zero anaphora are as clear as can be to a native Chinese speaker. Zero anaphora appears in English in very limited cases. The distribution of zero anaphora in English is no comparison to that in Chinese, as has been demonstrated in [6].

Since zero anaphora itself is nothing but a "hole" in the sentence, it simply offers no clues as to which antecedent is to be understood for the hole if taken in isolation. It is a phenomenon that could be handled only in the perspective of the actual communication in which it is embedded.

2.2 Reflexives

Reflexives are a linguistic device that can be an intensification element, generic anaphor, or anaphoric element. For instance, in the sentence "**自己的功課自己做**(Do your home works by yourself.)" 自己(self) is a generic anaphor to represent persons. In the other sentence "**約翰喜歡自己買書**(John likes to buy books by himself.)." 自己(self) is an anaphor referring to 約翰(John). The uses of reflexive as a generic and anaphoric element exhibit in both Chinese and English, but the uses as an intensification element particularly exhibit in Chinese.

Chinese reflexives include 自己(self), 我自己(myself), 你自己(yourself), 他自己(himself), 她自己(herself), and 他/她們自己(themselves). The last five reflexives are called compound reflexives. The word 自己(self) is used to intensify the four pronoun prefixes 我(I), 你(you), 他(he), and 她(she). The use of compound reflexives, such as "**我自己買了十個蘋果**(I bought ten apples by myself.)," is called intensification and is one of the uses of Chinese reflexives [8]. A reflexive as an intensification element can easily be removed and usually do not change the meaning of sentences. The uses of generic anaphor or anaphoric element are utilized to refer generic object or antecedent. They cannot be removed.

2.3 Personal pronouns

Personal pronouns are the pronouns that refer to persons. They include first-person, second-person, and third-person pronoun. A complete list of these personal pronouns in Chinese includes: 你(you), 我(I, me), 他(he, him), 她(she, her), 它(it), 你們(you), 我們(we, us), 他們/她們/它們(they, them).

Among personal pronouns, third person pronouns are often seen as the 'prototypical' personal pronouns in contrast to the first and second person pronouns. Therefore most of the works on pronoun discuss only third person pronouns [15]. Antecedent of reflexives and personal pronouns can be found by an extension to Chomsky's binding theory [8]. However, this kind of theory treats reflexives and personal pronouns within isolated sentences. It cannot be a realistic theory in considering discourse.

Reflexives like 你們(you), 我們(we, us), 他們/她們/它們(they, them) usually refer to multiple antecedents. For example,

Example 2

約翰 有 10 元，瑪利 有 5 元，他們 共 有 幾元?
Yuehan you 10 yuan, Mali you 5 yuan, tamen gong you ji yuan?
John has 10 dollars, Mary has 5 dollars, they totally have how-much dollars?
John has 10 dollars. Mary has 5 dollars. How many dollars do they have?

Here the personal pronoun 他們(they) refers to two antecedents: 約翰(John) and 瑪利(Mary).

Personal pronouns in Chinese, like those in English, are usually marked for gender and number. The agreements between the pronoun and the referred noun or noun phrase is useful in resolution.

2.4 Pronominal noun phrases

Noun phrase is the most common and frequent anaphoric device in English. It also occurs frequently in Chinese. Pronominal noun phrases include definite NP, possessive NP, and bare NP.

Definite NP is the NP composed by three elements: a prefix in the set of "這(this)", "那(that)", "每(every)", a classifier such as "個", "套", "些", and a noun. For instance, "這列火車(this train)" is an example of definite NP. Number can be inserted between prefix and classifier, such as "這兩個人(the two persons)".

Possessive NP is the NP comprised of three elements: a personal pronoun/reflexive /proper name, a word "的", and a noun/NP. Bare NP is an NP without any syntax marker as its prefix or suffix.

Pronominal NP can also be an anaphoric expression referring to multiple objects. However, it occurs when there is a hierarchical relationship or containing relationship between the pronominal NP and its antecedents

Type of anaphor	Lexical realization
Zero anaphora	ϕ
Reflexives	自己(self), 我自己(myself), 你自己(yourself), 他自己(himself), 她自己(herself), and 他/她們自己(themselves), 約翰自己(John himself)…
Personal pronouns	你(you), 我(I, me), 他(he, him), 她(she, her), 它(it), 你們(you), 我們(we, us), 他們/她們/它們(they, them).
Pronominal noun phrases	這個(this), 這些(these), 那個(that), 那些(those), 每個(every, each), 我的(my), 你的(your), 他的(his), 她的(her), 他/她/它們的(their), 約翰的…

Table 1. Types of anaphor in Chinese

Table 1 summarizes different types of anaphor. There are two types of reference. According to Hirst [12], types of reference can be distinguished with *identity of reference anaphora* (IRA) and *identity of sense anaphora* (ISA). The IRA-type anaphor refers to the same object of the antecedent. The ISA-type anaphor denotes not the same object of the antecedent, but one of related objects. We can observe in the next section that both reference types exist in algebraic word problem.

3. Algebraic Word Problem

This section will give examples of algebraic word problem. These examples show that the problem does reveal much complex issues in Chinese language understanding.

3.1 Zero anaphora

By eliding the entity which appears before where the anaphor locates, one can construct more coherent and concise discourse. The first sentence in example 1 is found to elide a verb. Here is another example derived from example 1,

Example 2

約翰 今年 ϕ 7 歲, 13 年 後 ϕ 是 幾歲?
Yuehan jin-nian 7 sui, 13 nian hou sh jisui?
John this year 7 years-old, 13 years after is how-old?
John (is) 7 years old this year. How old will (he) be after 13 years?

The difference between examples 1 and 2 is that the third person pronoun in the second sentence of

example 1 is eliminated in example 2. Zero anaphora can be found almost in all of the following examples. But the two discourses have the same meaning. In example 2, the natural language system has to find a way to disambiguate the word senses of all words in the absence of a verb and a subject. It also have to detect and recover the elided subject in order to apply an addition operation due to that the two sentences describe the same property of the same person. However, since zero anaphor is just a “hole”, there is no obvious clue to resolve it. Literature usually applies pragmatic knowledge in the parsing stage [9] to do partial parsing, or use a coreference module to resolve zero anaphora after parsing [19]. Zero anaphora is such a prevalent phenomenon that we will see it in the following anaphora types.

3.2 Reflexives

Algebraic word problem has the issue of reflexive.

Example 3

約翰 幫 弟弟 買 糖果 花掉 100 元, 約翰 自己 也 花掉 20 元, 約翰 共用 多少 元?
Yuehan bang didi mai tangguo huadiao 100 yuan, Yuehan zji ye huadiao 20 yuan, Yuehan gong duoshao yuan?
John help brother buy candy pay 100 dollar, John himself also spend 20 dollar, John total spend how-many dollar?
John spends 100 dollars to buy candy for his brother. John also spends 20 dollars for himself. How many dollars does John spend?

In this example, the reflexive 自己(self) in the second sentence combines with the proper noun 約翰(John) and serves as an intensification element. This reflexive can be neglected. The proper noun can alone stand for the subject.

Example 4

約翰 幫 弟弟 買 糖果 花掉 100 元, 自己 也 花掉 20 元, 約翰 共用 多少 元?
Yuehan bang didi mai tangguo huadiao 100 yuan, zji ye huadiao 20 yuan, Yuehan gong duoshao yuan?
John help brother buy candy pay 100 dollar, himself also spend 20 dollar, John total spend how-many dollar?
John spends 100 dollars to buy candy for his brother. John also spends 20 dollars for himself. How many dollars does John spend?

In this example, the reflexive in the second sentence serves as an anaphoric element. The reflexive refers to the subject in the preceding sentence. Therefore, a natural language understanding system should resolve the reference relationship and then get the answer 100+20=120. There is no generic anaphor use of reflexive in algebraic word problem.

3.3 Personal pronouns

The following examples include not only the issues of personal pronouns, but also the issues of zero anaphora and reflexive. However, the reflexive in example 5 is an intensification element. System can neglect the reflexive and use resolution method with zero anaphora to resolve the pronoun. The reference distance (the distance between the anaphor and the antecedent) interestingly crosses two sentences. This indicates that the accessibility in algebraic word problem can be hard.

Example 5

約翰 原 有 100 元, ϕ 得到 20 元, 後來 ϕ 又 給 弟弟 10 元, 他 自己 還 有 幾 元?
Yuehan yuan you 100 yuan, dedao 20 yuan, houlai you ji didi 10 yuan, tai zji hai you ji yuan?
John original has 100 dollar, get 20 dollar, then again give 10 dollar, he himself still has how-many dollar?
John has 100 dollars. (He) gets 20 dollars. Then (he) give 10 dollars to his brother. How many dollars does he have?

Example 6

約翰 有 100 元, 包伯 有 10 元, 他們 共 有 幾 元?
Yuehan you 100 yuan, Baoba you 10 yuan, tamen gong you ji yuan?
John has 100 dollar, Bob has 10 dollar, they total has how-much dollar?
John has 100 dollars. Bob has 10 dollars. How much money do they have?

Example 6 exhibits an ISA-type anaphor. The personal pronoun in the third sentence refers to two preceding objects: 約翰(John) and 包伯(Bob).

3.4 Pronominal noun phrases

Here we give two examples of definite NP.

Example 7

約翰 上個月 得到 1000 元, ϕ 上個月 比 這個月 少 拿 100元, 這個月 ϕ 得 幾 元?
Yuehan shanggeyue dedao 1000 yuan, shanggeyue bi zhegeyue shao na 100 yuan, zhegeyue de ji yuan?
John last-month get 1000 dollar, last-month than this-month less get 100 dollar, this-month get how-many dollar?
John got 1000 dollars last month. (John) got less 100 dollars than last month. How many dollars did (John) get this month?

Example 8

瑪利 買 衣服, 上衣 ϕ 500 元, ϕ 比 裙子 便宜 150 元, 這 套 衣服 共 值 幾 元?
Mali mai yifu, shangyi 500 yuan, bi qunzi pianyi 150 yuan, zhe tao yifu gong zh ji yuan?
Mary buy dress, blouse 500 dollar, than skirt cheap 150 dollar, this CL dress total cost how-many dollar?
Mary (want to) buys dress. The blouse costs 500 dollars. The skirt is cheaper 150 dollars than the blouse. How many dollars does the dress cost?

The definite NP can be a simple IRA-type anaphor as in example 7. It can also be a complex ISA-type anaphor as in example 8. The following will give examples of possessive NP and bare NP.

Example 9

約翰 有 15 個 蘋果, 瑪利 是 他的 8 倍, 瑪利 有 幾 個ϕ ?
Yuehan you 15 ge pingguo,Mali sh tade 8 bei,Mali you ji ge?

John has 15 CL apple, Mary is his 8 time, Mary has how-many CL?
John has 15 apples. Mary has 8 times than John. How many apples does Mary have?

Example 10

農場 有 牛 5 隻，ϕ 羊 3 隻， 牲畜 共 有 幾 隻?
NongChang you niu 5 zhi， yang 3 zhi, shengxu gong you ji zhi
Farm has cow 5 CL， sheep 3 CL, animal total has how-many CL?
There are five cows in the farm. There are 3 sheep (in the farm). How many animals are there?

4. Computer Processing of Chinese Algebraic Word Problem

In this section, we will discuss the characteristics of the algebraic word problem. The corpus of Chinese algebraic word problem is texts obtained from mathematical word problem books from the third grade to the sixth grade of elementary school. The corpus consisted of a total of 3488 questions which include 12,620 sentences and 124,925 Chinese characters. Mean sentences numbers of each question is 3.62. Mean Chinese character numbers of each sentence is 9.9.

Table 1 gives the occurrences of the four anaphoric forms found in the corpus. We found that zero anaphora dominates the anaphora phenomenon in algebraic word problem.

	No.	%
Zero anaphora	5581	70.5
Reflexive	6	0.08
Personal pronoun	135	1.71
Pronominal noun phrase	2194	27.72

Table 2 Occurrences of anaphoric forms

A prototype is developed to study Chinese algebraic word problem [19]. It consists of user interfaces for knowledge acquisition, partial parsing, and discourse reasoning. The prototype system includes a lexicon with about 30,000 Chinese phrases, 20,000 semantic rules, 100 discourse procedures, and other linguistics information. It can achieve 90 correctness rates in solving the first 300 word problems of the corpus.

5. Conclusions

This paper analyzes the anaphora features of Chinese algebraic word problem. The linguistic characteristics of the algebraic word problem will be given in the following. First, a question in Chinese algebraic word problem is a coherent discourse. That is, there is no long-distance pronominalization that a noun phrase is first introduced into the discourse, and then, after the focus of the discussion shifts to some other antecedents, the first antecedent is mentioned by means of an anaphor. Second, there is no relative clause or complex NP, such as “他的朋友所養的貓(the cat that his friend raises).” Complex NP may frequently appears in English, but rarely in Chinese. However, complex NP is still far from being discussed even in English due to its remarkable complexity. Third, referents are always antecedents. Referents cannot occur after where the anaphor locates. This is, however, the assumption of almost all of the current research of anaphora. Fourth, studies based on Chinese algebraic word problem will focus on written utterances, but not spoken utterances. Fifth, there are narrative utterances in the problem, but no conversational utterances.

In past decades, many works on Chinese natural language pays attention on those phenomena within single sentence. This line of work cannot resolve many applications that consist of more than one sentence. We think that the research of Chinese natural language should pay more attention on discourse and anaphora analysis due to the very nature of Chinese. We also show that Chinese algebraic word problem can provide a good test-bed for researching these issues.

References

[1] Fox, B. A., *Discourse Structure and Anaphora: Written and Conversational English*, Cambridge University Press, 1987.

[2] Allen, J. *Natural Language Understanding*, 2nd, Benjamin/Cummings, 1995.

[3] Sidner, C. L., "Focusing in the comprehension of definite anaphora," *Computational Models of Discourse*, M. Brady and R. C. Berwick Eds., The MIT Press, pp. 267-330, 1983.

[4] Charniak, E., "Toward a model of children's story comprehension," *AI TR-266*, MIT AI Lab, 1972.

[5] Poesio, M. and R. Vieira, "A corpus-based investigation of definite description use," *Computational Linguistics*, pp. 183-216, 1998.

[6] Liang, T. "Zero anaphora in Chinese: cognitive strategies in discourse processing," Ph.D dissertation, University of Colorando at Boulder, 1993.

[7] Chen, B. L. and V. W. Soo, "An acquisition model for both choosing and resolving anaphora in conjoined Mandarin chinese sentences," *COLING-92*, pp. 274-280, August, 1992.

[8] Chen, H. H., "Anaphora resolution algorithms for Mandarin Chinese," *Communications of COLIPS*, vol. 3, no. 2, pp. 59-67, 1993.

[9] Lin, K. H. C. and V. W. Soo, "Toward discourse-guided theta-grid parsing for Mandarin Chinese -- a preliminary report," *Proceedings of ROCLING II*, pp. 259-270, 1993.

[10] Huang, S. "Getting to know referring expressions: anaphor and accessibility in Mandarin Chinese," *Proceedings of ROCLING V*, Taipei, R.O.C., pp. 27-51, 1992.

[11] Chen, P. "A discourse analysis of third person zero anaphora in Chinese," University of California, Los Angeles, 1984.

[12] Hirst, G. "Discourse-oriented anaphora resolution in natural language understanding: a review," *American Journal of Computational Linguistics*, vol. 7, no. 2, pp. 85-98, 1981.

[13] Lappin, S. and H. J. Leass, "An algorithm for pronominal anaphora resolution," *Computational Linguistics*, vol. 20, no. 4, pp. 535-561, 1994.

[14] Li, C. N. and S. A. Thompson, "Third-person pronouns and zero-anaphora in Chinese discourse," *Syntax and Semantics*, vol. 12, Academic Press, pp. 311-335, 1979.

[15] Wales, K., *Personal Pronouns in Present-day English*, Cambridge University Press, 1996.

[16] Grosz, B. J., A. K. Joshi, and S. Weinstein, "Centering: a framework for modeling the local coherence of discourse," *Computational Linguistics*, vol. 21, no. 2, pp. 203-225, 1995.

[17] Kamp, H. , "A theory of truth and semantic representation," *Formal Methods in the Study of Language*, MC TRACT 135, J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof (Eds.), Amsterdam, p. 277, 1981.

[18] Chen, B. L. and V. W. Soo, "An acquisition model for both choosing and resolving anaphora in conjoined Mandarin Chinese sentences," *COLING-92*, pp. 274-280, August, 1992.

[19] Wang, Y. K., Y. S. Chen, and W. L. Hsu, "Empirical study of Mandarin Chinese discourse analysis: an event-based approach," to appear in *10th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI'98), 1998.