AN EMPIRICAL STUDY ON THE GENERATION OF ZERO ANAPHORS IN CHINESE

Ching-Long Yeh[*]and Chris Mellish[†]
Department of Artificial Intelligence
University of Edinburgh
80 South Bridge
Edinburgh EH1 1HN
Britain

Abstract

In this paper, we describe the creation of rules for generating Chinese zero anaphors through a sequence of experiments in a stepwise enhanced manner. In the experiments, we basically examined the occurrence of zero anaphors in a real text and the ones generated by the algorithms employing the rules, assuming the same semantic and discourse structures as the text. The factors of locality, syntactic constraints, discourse structure and salience of objects were considered in the rules. The results of the experiment show that 93% of the zero anaphors in the text can be correctly generated by an algorithm using a rule involving all the above factors.

1 Introduction

Anaphoric expressions in Chinese can be classified as zero, pronominal and nominal forms, as exemplified in (1) by ϕ_1^i, ta^i (he) and nage ren^i (that person), respectively.

(1)a. Zhangsan^i jinghuang de paokai,
Zhangsan frightened NOM to run-away
Zhangsan was frightened and ran away.
b. ϕ_1^i zhuangdau yige dahan^j,
(he) bump-to a big-man
(He) ran into a big man.
c. ta^i kanqing le na ren ^{j} de zhangxiang,
he see-clear ASP that man GEN appearance
He watched clearly that man's appearance.
d. ϕ_2^i renchu na ren^j shi shei.
(he) recognize that man is who
(He) recognized who that man is.

In their paper [Li and Thompson 79], Li and Thompson have shown that zero anaphors in Chinese can occur in any grammatical slot with an antecedent that may occur in any grammatical slot, regardless of the distance between them. Although there is no clear rule to account for zero anaphora, nevertheless, as pointed out by Li and Thompson, zero anaphora commonly occur in the situation of a "topic chain," where a referent is referred to in the first clause, and then several more clauses follow talking about the same referent but with it omitted. In [Chen 87], Chen proposed the notion of "continuity" of referent in discourse to give a more specific account of zero anaphora.

In this paper, we aim at deciding when to generate zero anaphors from some internal semantic structure. Although there are no clear rules stated in previous linguistic work, we, nevertheless, can summarize a very simple rule, Rule 1 as shown below, for the generation of zero anaphors.

> **Rule 1:** If an entity, *e*, in the current utterance was referred to in the immediately preceding utterance, then a zero anaphor is used for *e*; otherwise a non-zero anaphor is used.

We performed an experiment by comparing the zero anaphors generated by the algorithm employing this rule and those occurring in real text to see how well it works. The initial result showed that zero anaphors were over-generated to a large extent in the text produced by employing Rule 1. Consequently, we considered other well-known factors namely, syntactic constraints [Li and Thompson 81], discourse structure [Grosz and Sidner 86] and the salience of objects in utterances [Sidner 83], to get better results.

2 Experiment

A number of articles written by different authors were selected as the linguistic sources with which the text produced by employing the generation algorithms can be compared. For the moment, the selected articles are restricted to the exposition type, namely, ones which explain an idea or discuss a problem. Two sets of data were selected; one consists of a number of scientific questions and answers for children and the other is a brief introduction to modern Chinese grammar. Basically the experiment was executed in three steps. First, zero anaphors within the selected articles were identified. Second, for each paragraph in the selected

[*] Also with the Department of Information Engineering at Tatung Institute of Technology, Taipei, Taiwan. Email address is chingyeh@aisb.ed.ac.uk.

[†] Email address is chrism@aisb.ed.ac.uk.

articles, we examined each utterance sequentially and recorded the occurrence of zero anaphors that would be obtained by applying the algorithm using a rule, like Rule 1. Third, we noted down the differences between the results of steps 1 and 2.

In step 3, we categorized the differences between the results as: *correct*, *false* and *missing* types. If a reference created by the algorithm is the same as the one in the real text, then it belongs to the *correct* type. If a zero anaphor is created by the algorithm, while the corresponding position in the real text is non-zero anaphor, then it belongs to the *false* type. Conversely, if a zero anaphor is found in some position in the real text, while a non-zero anaphor is created by the algorithm, then it belongs to the *missing* type. The task of step 3 is to count the number of cases in each type.

3 Results

Having done this, we carried out similar experiments with enhanced rules.

3.1 Effect of using Rule 1 and adding syntactic constraints

In Sets 1 and 2 of the testing data, there are 651 and 149 anaphors, respectively. By using the algorithm of Rule 1 on the data, the result is shown in Table 1. In the data, 7 and 1 long distance zero anaphors occur but the algorithm decides to use non-zero ones for the corresponding positions. Consequently, they belong to the *missing* type. From the result shown in Table 1, the performance of the algorithm is obviously unpromising.

There are certain syntactic constraints on zero anaphora, regardless of discourse factors, as shown in [Li and Thompson 79, Li and Thompson 81], . Therefore, we enhanced Rule 1 by adding the above syntactic constraints on zero anaphora, which becomes Rule 1a as below. Rule 1a can be alternatively be represented as a decision tree in Fig. 1, where internal nodes are conditions in the rule and leaf nodes are decisions about the anaphor type, either zero or non-zero.

> **Rule 1a:** If an entity, e, in the current utterance was referred to in the immediately preceding utterance and does not violate any syntactic constraint on zero anaphora, then a zero anaphor is used for e; otherwise a non-zero anaphor is used.

In Table 1, by using Rule 1a, the *correct* cases increase from 408 to 510 and 98 to 126 for Sets 1 and 2, respectively. Though Rule 1a improves its ancestor's performance, the result, however, still discourages us from using it for the generation of zero anaphors.

3.2 The effect of adding discourse structure

Grosz and Sidner suggest, that three structures can be identified within a discourse: *linguistic structure*, *intentional structure*, and *attentional state* [Grosz and Sidner 86]. An important idea in the theory is the mutual effect between the linguistic expressions in utterances constituting the discourse and the discourse segment structure. What concerns us here is the interrelationship between the forms of referring expressions and the discourse segment structures. In NL generation systems, the semantic structures of messages to be produced are usually organized according to hierarchical intentional structures; then, based on the structures, referring expressions are decided [Hovy 90, Dale 92]. Hence, in this subsection, we employ the idea of discourse structure to improve our algorithm for the generation of zero anaphors.

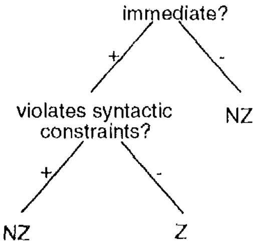

Figure 1: Decision tree for Rule 1a.

Table 1: The results of the algorithms of Rules 1 and 1a.

Set	Alg.	Cor.	Fal.	Mis.
S1	R1	408	236	7
	R1a	510	134	7
S2	R1	98	50	1
	R1a	126	22	1

In their study [Li and Thompson 79], Li and Thompson propose that "the degree of preference for the occurrence of pronominal anaphora in a clause inversely corresponds to the degree of connection with the preceding clause." They listed the following conditions of decreasing of connection: switching from background to foreground information, or vice versa, between two clauses, the second clause headed by an adverbial expression and two clauses spoken by two different participants.

In general, a zero anaphor used to refer to some entity in the previous utterance might be expected to indicate the continuation of a discourse segment, while a non-zero anaphor occurring in the same situation signals a boundary of discourse segment. From the generator's perspective, when the decision of the anaphoric form for a phrase referring to some entity in the previous utterance is to be made, the factor of discourse segment boundary must be taken into consideration. Therefore, based on this idea, we improve the previous rules for generation of zero anaphors, Rules 1 and 1a, to make the following rule. The decision tree for Rule 2 is shown in Fig. 2.

> **Rule 2:** If an entity, e, in the current utterance, u, was referred to in the immediately preceding utterance and does not violate

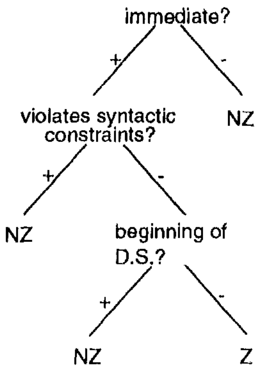

Figure 2: Decision tree for Rule 2.

> any syntactic constraints on zero anaphora, then if u is not the beginning of a discourse segment, then a zero anaphor is used for e; otherwise, a non-zero anaphor is used.

To perform the experiments for the new rules, we have to access the discourse segment structures of the testing data. Therefore, we annotated the boundaries between discourse segments in the testing data and the hierarchical discourse structures according to the discourse segment intentions. We further carried out a test by comparing our annotations with other native speakers of Chinese. In the test, four native speakers of Chinese were asked to do the same tasks we have done for five articles selected from the testing data. Comparing with the speakers' results, on average 76% of the speakers' annotations coincide with ours. According to the above comparison the annotations we made were reliable for the purpose of the experiment.

We then performed the experiment by employing the algorithm of Rule 2. As shown in Table 2, for the Set 1 data, 49 and 12 zero anaphors were over- and under-generated by the algorithm, respectively. For the other set of testing data, Rule 2 achieves an even better result.

Table 2: The results of the algorithm of Rule 2.

Set	Alg.	Cor.	Fal.	Mis.
S1	R2	590	49	12
S2	R2	145	3	1

3.3 The effect of topic

In this subsection, we use the feature of topic in Chinese to further refine the previous rules. The basic idea here is to investigate the positions of antecedent and anaphor in their respective utterances. In the following, we divided the position of anaphors in their respective utterances into topic and non-topic. For each anaphor, its antecedent's position is one of the following categories: topic, direct object or the NP following a presentative verb and others. We thus classify the following types, A to F, of antecedent-anaphor pairs: the antecedents of Types A and C are in topic position, of B and D are in direct object position or are the NP following a presentative verb, and of E and F are in other positions; the anaphors of Types A, B and E are in topic position, and of C, D and F are in non-topic positions.

Since in the new rule conditions on topic and non-topic will only be considered after the conditions in Rule 2, in investigating the antecedent-anaphor pairs, we have to exclude the ones with either their anaphors violating syntactic constraints on zero anaphor or at the beginning of discourse segments. In other words, the new condition will be attached under the Z-node in the decision tree of Fig. 2. In the Set 1 testing data, there are 239 such pairs, among which anaphors of 49 pairs are zeroed by the algorithm of Rule 2 but appear in non-zero forms in the text. In other words, the 49 anaphors were over-generated by our algorithm, which in our terms belong to the *false* type; the other 190 cases belong to the *correct* type. The number of each type of pairs for both *correct* and over-generated cases in the testing data are shown in Table 3.

Table 3: Occurrence of antecedent-anaphor pairs in the data.

Type		A	B	C	D	E	F	Total
S1	*correct*	158	27	1	0	0	4	190
	false	15	14	6	1	9	4	49
	total	173	41	7	1	9	8	239
S2	*correct*	44	1	0	0	1	0	46
	false	1	0	2	0	0	0	3
	total	45	1	2	0	1	0	49

For the Set 1 data in Table 3, the over-generated cases of both Types A and B, 15 and 14 out of 173 and 41, respectively, are the minorities of the respective types, while on the contrary, the number of over-generated cases of Types C and E are greater that their counterparts. Thus, if we let anaphors of Types A and B be zero and Types C and E non-zero, then there will be 29 (15+14) over-generated zero anaphors and 5 (1+4) under-generated ones for the Set 1 testing data. The numbers for Types D and F do not conclusively support either using zero or non-zero in this case. In Chen's study [Chen 87], he found a higher percentage of zero anaphors occurring in the topic position with their antecedent most frequently in the topic or object positions of the immediately previous utterance, which strongly supports the idea of letting anaphors of Types A and B be zero and others non-zero. We choose to generate non-zero anaphora for Types D and F. We thus obtain Rule 3 by adding the affect of topic into Rule 2. The decision tree for Rule 3 is shown in Fig. 3. The results of using the new algorithm are shown in Table 4.

> **Rule 3:** If an entity, e, in the current utterance, u, was referred to in the immediately preceding utterance, does not violate any syntactic constraints on zero anaphora, and u is not at the beginning of a discourse

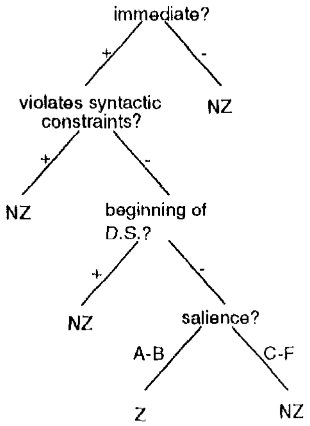

Figure 3: Decision tree for Rule 3.

Table 4: The results of the algorithm of Rule 3.

Set	Alg.	Cor.	Fal.	Mis.
S1	R3	605	29	17
S2	R3	146	1	2

segment, then if e is either a Type A or B pair, then a zero anaphor is used for e; otherwise, a non-zero anaphor is used.

As a short summary, the numbers of anaphors in the Set 1 testing data satisfying the conditions of Rule 3 are shown in Fig. 4, where Z, N and P represent zero, pronominal and nominal anaphors, respectively. Indicated in the root node are the total number of all kinds of anaphors in the data. The *correct* match is calculated by summing up the numbers of non-zero anaphors, pronouns and nominal anaphors, under non-zero leaf nodes and zero anaphors under zero leaf nodes. Non-zero anaphors under zero leaf nodes are the *false* matches. Conversely, zero anaphors under non-zero leaf nodes are the *missing* matches.

4 Future Work

In this paper, we focus on distinguishing zero and non-zero anaphors. To have a full account for anaphors in Chinese, two tasks remain to be done. The first is to distinguish pronouns and nominal anaphors for the non-zero cases, namely, to further add conditions under the non-zero nodes in the decision tree of Fig. 3. The second task is to develop an algorithm for the decision of an appropriate form for nominal anaphors [Dale 92, Reiter and Dale 92]. Afterwards, a Chinese NL generation system will be developeded to test the performance of the algorithms.

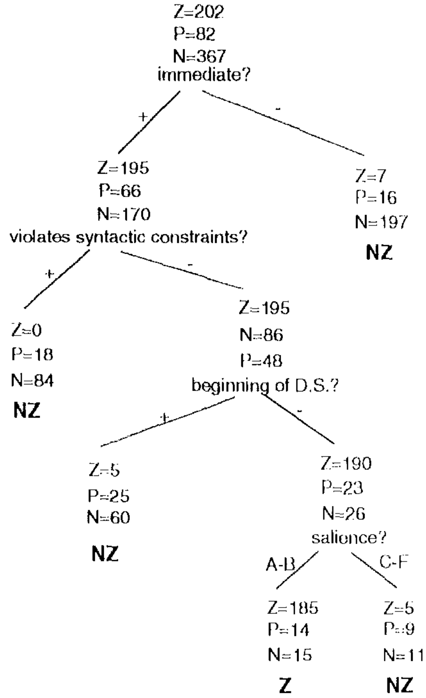

Figure 4: Anaphors in Set 1 testing data satisfying conditions of Rule 3.

5 Conclusion

A study on the generation of Chinese zero anaphors as opposed to the usual work from the comprehension side is presented. By doing experiments on a number of descriptive articles, we obtained a rule for the generation of zero anaphors, which incorporates the ideas of recency of occurrence, syntactic constraints, discourse segment structure and salience of objects in discourse. In the text generated by hand employing the algorithm of the above rule, assuming the same semantic structure and discourse segment structure as the real text, the use of zero anaphors is fairly close to those occurring in the real text. In the stepwise empirical study, the algorithms are improved through the test against real data, which in some sense provides the assessment for the effectiveness of the rule. The result of the assessment thus encourages us to employ the rule as a part of the referring expression component in the Chinese NL generation system we are developing.

References

[Chao 68] Chao, Y. R., *A Grammar of Spoken Chinese*, University of California Press, Berkeley, CA, 1968.

[Chen 87] Chen, P., "Hanyu lingxin huizhi de huayu fenxi (A discourse approach to zero anaphora in Chinese)" (in Chinese), *Zhongguo Yuwen (Chinese Linguistics)*, pp. 363-378, 1987.

[Dale 92] Dale, R., *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes,* The MIT Press, Cambridge, Massachusetts, 1992.

[Grosz and Sidner 86] Grosz, B. J. and Sidner, C. L., "Attention, intentions, and the structure of discourse," *Computational Linguistics,* 12(3), pp. 175-204, 1986.

[Hovy 90] Hovy, E., "Approaches to the planning of coherent text," in *Natural Language in Artificial Intelligence and Computational Linguistics,* Paris, C. L., Swartout, W. R., and Mann, W. C., (eds.), 1990.

[Li and Thompson 79] Li, C. N. and Thompson, S. A., "Third-person pronouns and zero-anaphora in Chinese Discourse," in Givon, T. (ed.), *Syntax and Semantics: Discourse and Syntax, Vol. 12*, pp. 311-335, Academic Press, 1979.

[Li and Thompson 81] Li, C. N. and Thompson, S. A., *Mandarin Chinese: a Functional Reference Grammar,* University of California Press, Berkeley, CA, 1981.

[Liu 84] Liu, Y. C., *Zhuowen de Fang Fa* (Approaches to Composition), Xuesheng Chubanshe, Taipei, Taiwan, 1984.

[Reiter and Dale 92] Reiter, E. and Dale, R., "A fast algorithm for the generation of referring expressions," *COLING-92*, pp. 232-238.

[Sidner 83] Sidner, C. L., "Focusing in the comprehension of definite anaphora," in Brady, M. and Berwick, R. C. (eds.),*Computational Models of Discourse,* pp. 267-330, MIT Press, 1983,