DETERMINING THE ANTECEDENT OF NOUN PHRASE CONTAINING THE DETERMINER *KONO* OR *SONO* IN JAPANESE

Toshimasa Koga, Haodong Wu, Teiji Furugori
Dept. of Computer Science
University of Electro-Communications
Chofu, Tokyo, Japan
Email: furugori@cs.uec.ac.jp

ABSTRACT

This paper offers a method for determining referent (antecedent) of the noun phrase containing determiner "kono (*this*)" or "sono (*that, its*)" in Japanese. It uses in the determination of the antecedent a statistical measure of conceptual similarities, taken from a corpus, between each candidate for the antecedent and the modeficand of *kono* or *sono*.

We describe our method and an algorithm. We then show an experiment that has given us an overall success rate of 85.2%. Our method is applicable to solve other semantic problems than that for finding the antecedent of noun phrase containing the determiner *kono* or *sono*.

1. INTRODUCTION

Reference problems are a central issue in natural language processing. For instance, we need to understand the antecedents of pronouns in translating one language into another. Consider:

> Someone killed Jim. The police have no suspect, but *they* think that *he* or *she* needed money and knew that *he* was a wealthy man. He walked in the house with a big suitcase and put the money in *it*.

We will not be able to translate the sentences into, say, Japanese when we are uncertain of what the pronouns like *he*, *she*, *they*, and *it* in this text are referring to.

A pronoun refers to a linguistic object in the preceding or succeeding sentences. In the sequence,

Clinton visited Japan. He gave a talk at a university.

the second sentence is connected to the first one by *He* (personal pronoun) referring to *Clinton* (its antecedent). The same relation holds in:

Clinton visited Japan. The president gave a talk at a university.

Here, the noun *president* refers to *Clinton.*

In this paper we try to devise a method for determining antecedent of the noun phrase containing determiner "kono (*this*)" or "sono (*that, its*)" in Japanese.

2. REFERENCE PROBLEMS AND COMPUTATION

2.1 Determiners *Kono* and *Sono* in Japanese

A pronoun may refer to an entity (noun) or an event (part of sentence or whole sentence). Similarly, the determiner like *this* or *that* together with its modeficand refers to an event or an entity. In the sequence of sentences:

I saw the latest work of Kurosawa. This movie was so good.

the noun phrase *this movie* refers to *the latest work of Kurosawa.*

We frequently use the noun phrase containing determiners *kono (this)* and *sono (that)* in Japanese. It refers to a linguistic object mentioned in the same or a preceding sentence:

(a) 都内の病院で肝臓移植が行なわれた。この手術は成功した。
(A *liver transplant* was performed in a hospital in Tokyo. *This surgery* was successful.)

(b) 岡の上に小さな家がある。その屋根は赤い。
(There is a *house* on the hill. *That (Its) roof* is red.)

(c) 一機の戦闘機が日本の領空をかすめた。政府はこの件を重くみた。
(*A fighter plane flew over the air of Japan.* The government considered *this invasion* very serious.)

(d)太郎は男を殴った。私は彼がどうしてその行動に出たか知っている。
(Taro *beat the man.* I know why he has come to take *that action.*)

This or 'この *(kono)*' together with its modeficand *surgery* (手術) in (a) refers to *liver transplant* (肝臓移植). And *that* or 'その *(sono)*' together with its modeficand *roof* (屋根) in (b) refers to *house* (家). Likewise, the noun phrase containing *kono* in (c) refers to the whole sentence, *A fighter plane flew over the air of Japan*, and the noun phrase containing *sono* in (d) refers to the verb phrase, *beat the man*, of the preceding sentence.

Table 1: Place of the Antecedent

Place of antecedent	*kono*	*sono*
same sentence	20 (10.1%)	153 (43.5%)
1st proceeding sentence	88 (44.4%)	165 (46.9%)
2nd proceeding sentence	46 (23.2%)	26 (7.4%)
3rd proceeding sentence	10 (5.1%)	5 (1.4%)
4th proceeding sentence	25 (12.6%)	3 (0.8%)
5th proceeding sentence	6 (3.0%)	
6th proceeding sentence	3 (1.5%)	

A survey in Table 1, taken from Mainichi Newspaper from January, 1994 to June, 1994 shows that there were 620 instances of *kono*'s and 740 instances of *sono*'s in the total of 331 editorials. Among them, 198 *kono*'s (32%) and 352 *sono*'s (48%) referred to entities and 422 (68%) and 388(52%) referred to events. It was also found that all the antecedents for *kono*'s were located within sixth preceding sentences and those of *sono*'s were located within fourth preceding sentences.

Another small survey in Table 2 shows the case markers or particles that are structurally associated with the entity antecedents. Over 70% of the antecedents appeared with the subject markers 'は(wa)' or 'が(ga)' and the object markers 'を(wo)' or 'に(ni)'.

Table 2: Antecedent and its Marker or Particle

	は(wa)	が(ga)	を(wo)	に(ni)	の(no)	で(de)	と(to)	も(mo)
kono (164)	29(17.7%)	26(15.9%)	28(17.1%)	20(12.2%)	26(15.9%)	19(11.6%)	13(7.9%)	3(1.8%)
sono (299)	62(20.7%)	62(20.7%)	63(21.1%)	38(12.7%)	29(9.7%)	21(7.0%)	19(6.4%)	5(0.2%)

2.2 Related Work

There are a number of studies on identifying the antecedent of pronouns. They can conveniently be classified into three methods depending on the means they use to find the antecedent: syntax-based, semantics-based, and discourse-based.

The syntax-based method uses certain fixed structures to determine the antecedent. Shimizu and Yokoo[1], for instance, identified the antecedent of an ellipsis or zero-pronoun according to the specific type of structures that appears in the predicate part of the sentence in which the antecedent is to be located. Thus, in

太郎は花子に宿題を手伝ってもらった。非常にうれしかった。
(*Taro* was given a help by Hanako for his homework. (Φ) was very grateful.)

the antecedent of the zero-pronoun (Φ) is *Taro* as the pattern 'もらった' (be given) in the predicate part of the preceding sentence suggests the antecedent to be its grammatical subject. There are a number of studies that employ similar measures in the determination of the antecedent of pronominal references[2,3,4,5].

The semantics-based method uses selectional restrictions[6] in the determination of the antecedent. In *Taro went to school and (φ) bought a textbook*, the antecedent of the zero-pronoun is *Taro* rather than *school* since the verb *buy* (bought) is required to take an animate subject. Many use selectional restrictions of a sort in identifying the antecedent of a pronoun[e.g. 7].

The discourse-based method uses theme, topic, focus, etc. and locational or distance information to find the antecedent. In an example, Kameyama[8] used the centering theory by Grosz and Sidner[9,10] to identify the antecedent of zero-pronoun. The antecedent is *Taro* rather than *Jiro* in the sentence:

Taro brought Jiro with him and (φ) was prized by his teacher.

This is so determined as the focus of the sentence is in *Taro*.

According to Kameyama, the antecedent of zero-pronoun tends to be the element for Topic, grammatical Subject, Object, and Others of a sentence in that preferential order. The use of topic, theme, focus, or similar concept to finding the antecedent is seen in other studies [11,12].

The syntax-based method is a rule-based and language-dependent device. The discourse-based one becomes language-dependent in the end, too, since there are no ways of finding the focus of a sentence without using syntactic information specific to the language being analyzed. The semantics-based one is language-independent, but it has a deficiency in its extensibility[13].

One reason or another, all of the studies concerning with the reference problems thus far mentioned have restricted test data to be a sentence or a pair of sentences. We notice also that the researchers have opted to use the data suitable only for their experiments.

3. METHOD OF DETERMINING THE ANTECEDENT

We try to make our method of determining the antecedent conceptual-based, rather than syntactic, semantic or discourse-based. We would like to make it language-independent also. To do so, we base our method for finding the antecedent on a statistical measure taken from a corpus.

3.1 Observations and Assumptions

It seems reasonable to think that the antecedent is related in the meaning to the head noun of noun phrase containing the determiner. The very example of this is the case where the antecedent is identical to the noun associated with a determiner.

わたしは心臓の手術をした。この手術は8時間以上かかった。
(I had a heart *surgery*. *This surgery* took well over 8 hours.)

Here, *this surgery* refers to (heart) *surgery* in the preceding sentence. In other cases, the antecedent may be a synonym of, part of, ..., or upper-class or subclass of the head noun containing the determiners. In an example, *engine* (エンジン) in the following sentence is a part of *car*.(車)

車のよしあしは、そのエンジンの性能でわかる。
(The quality of a *car* is found by the performance of *its* engine.)

When we use a pronoun or the determiner *kono* or *sono*, we make it refer to an entity that is close to it in distance. Otherwise, the discourse will lose coherency. Our survey in Table 1 in fact indicated that the place of antecedents would not go beyond the sixth preceding sentence. This is to say that the location or distance plays an important role in finding the antecedent.

Syntactic features may be important to determining the antecedent. Case markers in Japanese and modes (active or passive) in English, for instance, would give us a preferential information that can be usable to identify the antecedent.

3.2 Algorithm

Based on the assumptions in section 3.1, we devise an algorithm to find the antecedent of the noun

phrase containing the determiner *kono* or *sono*. We try to find the antecedent by measuring conceptual similarity, using mutual information[14], between each of probable antecedents and the noun associated with the determiner. If it does not work well, then we will use the case information and the distance information.[1]

Our algorithm for finding the antecedent is:

1. Read a text that contains the noun phrase with *kono* or *sono* pronoun in it. Let the modeficand of the determiner be N.
2. Find nouns C's (candidates for the antecedent) in the preceding part of the noun phrase.
3. If N is in C's, then choose the identical noun to N as the antecedent. Exit.
4. Calculate the mutual information(MI) between each of C's and N.
   - 4.1 If the MI values are zeros for all C's, then choose the antecedent by using the case and distance information.
   - 4.2 If MI is got for only one candidate, C1, of C's, then calculate MI values between the upper-class concept of the members of C's and N and then
     - if MI values are all zeros, then choose C1 as the antecedent. Exit.
     - if MI value for the upper-class of C1 is the highest, then choose C1 as the antecedent. Exit.
     - if MI value for the upper-class of C1 is not the highest, then select the two candidates whose upper-class concepts had higher MI values than others and make them to be C1 and C2. Goto step 6.2.
5. Select C1 and C2 from C's that are given higher MI values than others.
6. Select the upper-class concept of C1 and that of C2 and calculate MI between each of them and N.
   - 6.1 if the MI values are zeros, then choose C1 or C2 as the antecedent according to the original MI values.
   - 6.2 if the MI values are equal, then choose C1 or C2 as the antecedent by using the case and distance information. Exit.
   - 6.3 if the MI values are not equal, then choose C1 or C2 as the antecedent according to the MI values.

We note that this algorithm does not deal with the antecedent that is an event since it needs totally different strategies to find the antecedent. Among possibly many C's, we select only two of them, C1 and C2, in step 5 and use the two as the candidates of the antecedent. This choice is empirical. Our preliminary experiment indicated that two candidates with higher MI values were

---

[1] Mutual information between two concept is calculated as follows.

$$I(w1,w2) = \log_2\{N*f(w1,w2)/f(w1)f(w2)\}$$

Here N is the size of the corpus used in the calculation, f(w1,w2) is the frequency of the co-occurrence of the words w1 and w2, and f(w1) and f(w2) those of the individual words. We use the EDR Japanese Corpus [15], a 220,000 sentences corpus, to calculate the mutual information.

enough to get the best result: the success rate of finding the antecedent was about 42% when we used one candidate, 61% with two candidates, 57% with three candidates, and much less in other cases.

We use in step 6 the upper-class concepts[2], more general terms, of C1 and C2 to get mutual information between N and each of them. This use of the upper-class concepts is based on the following assumptions.

(1) C1 would be the antecedent of N when C1 and N has higher MI value than that of C2 and N since N and its antecedent are considered to be closely related semantically.

(2) When (1) holds, it would be generally true that the MI values between its upper-class and N would be higher than that of N and the upper-class concept of C2.

(3) C2 would be the antecedent of N when (2) would not hold, i.e., MI values between the upper-class concept of C2 and N gets higher than that of the upper-class concept of C1 and N, because the conceptual hierarchy of C1 may not have a semantic relevancy with N in this case.

Here, the assumption (2) is to reinforce (1). When the assumption (2) is false, we choose C2 to be the antecedent of N, using the assumption (3).

When the statistical measure would not provide us with the antecedent, we use the syntactic information to find the antecedent. Here, we give higher priority to the candidates that appear with the subject or object markers than to the ones that appear with other case markers. The distance information is a stopgap measure. We use it when the syntactic information fails to produce the antecedent. Here, the closer-the-better approach is employed, i.e., we choose as the antecedent the candidate closer to the determiner in distance. Other small details in steps 4 and 6 are all empirical.

4. EXPERIMENT AND RESULT

We now show how our algorithm works and an experiment that demonstrates its performance.

4.1 Examples

Consider the following text:

おそらく今後も解決が最も難しい課題の一つとして国際社会に重くのしかかるだろう。この難しさの一因は民族という概念がはらむ両面性をもっている。

The noun phrase with the determiner 'この ' is 'この難しさ(*this difficulty*)' and the candidates for its antecedent are '解決, 課題 and 国際社会'. The following are the mutual information between the head noun and each candidate and between the head noun and the upper-class concept of each candidate.

[2] The upper-class concepts, more general terms, are found by using a conceptual dictionary, Nippongo Goi-Taikei[16].

Candidates	MI	Upper Concepts	MI
解決 (settlement)	1.048108	解決・未決 (judgment)	1.048108
課題 (subject)	0.887529	問題 (problem)	0.762590
国際社会 (international society)	0.000000		

C1 and C2 chosen in this example are *settlement* and *subject*. The upper concepts of C1 and C2 are *judgment* for *solution* and *problem* for *subject*. C1 and its upper-class concept have higher MI values consistently and we thus choose *settlement* to be the antecedent, i.e., the noun phrase *this difficulty* refers to '解決の難しさ (difficulty of the settlement)'.

Let us see another example.

何らかの対策が求められる。メーカは、定められたテストに従った電化製品の検査を行い、そのデータの提出が義務づけられるであろう。

The antecedent candidates for 'そのデータ(*that data*)' in this text are '対策、メーカ、 電化製品、 テスト、and 検査'. Mutual information between 'データ' and each of '対策', 'メーカ', 'テスト', '電化製品', and '検査' is shown below. We select '電化製品' and '検査' as C1 and C2. The MI values between each of them and N are 1.49 and 0.93. However, those of their upper-class concepts, '商品' and '調査', and N are 0.09 and 0.59. Thus, we choose we choose C2 instead of C1 as the antecedent.

Candidate	MI	Upper Concept	MI
対策	0.000000		
メーカ	0.342484		
テスト	0.860056		
電化製品	1.486846	商品	0.086740
商品	0.927538	調査	0.594604

4.2 Result and Evaluation

We have tested 54 instances of *kono*'s and 101 instances of *sono*'s in the editorials in Mainichi Newspaper. Table 3 shows the results from our experiment. As is indicated here, the overall success rate of finding the antecedents is 85.2% (132/155) against human judgments.

The use of conceptual information is effective in statistics-based studies of natural language processing. For our purpose, hierarchical concepts such as upper and lower classes work much better than horizontal concepts such as synonyms and antonyms. We have found that the success rate of finding antecedent was far lower when we used the candidate words, C's, whose overall success rate was about 53.5% to determine the antecedent.

We have tested other methods to identify the antecedents. For the data we used, the success rate was only 48.4% when we employed only the information from distance and case markers. We can not compare our results with those of other studies directly since the conditions of experiments and types of reference problems dealt with vary. However, it seems that the overall performance of our experiment is better than those of others.

Table 3: Experimental Result

	kono		*sono*	
Main Algorithm	49/52	(94.2%)	69/78	(88.5%)
Case and Distance	0/2	(0.0%)	14/23	(60.9%)
Total	49/54	(90.7%)	83/101	(82.2%)

5. CONCLUDING REMARK

We presented a method for identifying the antecedent of noun phrase containing the determiner *kono* or *sono*. We then tried to prove the effectiveness of our method in a computational experiment. The result is promising.

The computational studies done so far for identifying antecedents of pronouns or the like are language-dependent, regardless of the methods they employed. Furthermore, all of them restrict test data to be a sentence or a pair of sentences with the input suitable for their experiments. Ours is statistics-based, language independent, though we have enhanced the algorithm by using the syntactic and locational information, and text-independent in that we have no restrictions as to the kinds of input we take.

Our method is not problem free, however. One computational problem lies in that we used and made up a small conceptual dictionary of our own since Nihongo Goi-Taikei is neither a machine-readable nor complete. Another is that we faced the problem of dividing a compound noun into a series of simple nouns to get the candidates of antecedent. This is an old problem where a solution is necessary in many application areas of Japanese language prccessing.

References

[1] Kazukiyo Shimizu and Hidetoshi Yokoo, "Empathy Extraction and Anaphoric Disambiguation in Japanese Language Understanding System, " Transactions of Information Processing Society of Japan, Vol.36, No.2, pp.236-246, 1995.

[2] Hiromi Nakaiwa and Satoru Ikehara, "Intrasentential Resolution of Japanese Zero Pronouns using Pragmatic and Semantic Constraints," Journal of Natural Language Processing, Vol.3, No.4, pp.49-65, 1996. (in Japanese)

[3] Shin'ichiro Nisizawa, Keichi Kimura and Hiroshi Nakagawa, "The Resolution of Coreferential Relation in Japanese Complex Sentences Based on Semantics of Verbs and Adjectives, " Transactions of Information Processing Society of Japan, Vol.38, No.3, pp.472-481, 1997.

[4] Takehiko Yoshimi and Jiri Jelinek, "Anaphora Resolution of Sentences and Noun Phrases by Matching Dependency Structures, " Journal of Natural Language Processing, Vol.4, No.1, pp.111-123, 1997. (in Japanese)

[5] Tsuyoshi Yamamura, Noboru Ohnishi and Noboru Sugie, "A Method of Anapora Resolution Based on Concept of Observation, " The Transactions of the Institute of Electronics, Information and Communication Engineers D-II, Vol.J77, No.1, pp.162-169, Jan.1994.

[6] Msaaki Murata, Makoto Nagao, "An Estimate of Referents of Pronouns in Japanese Sentences using Examples and Surface Expressions, " Journal of Natural Language Processing, Vol.4, No.1, pp.87-109, 1997. (in Japanese)

[7] Msaaki Murata, Makoto Nagao, "Indirect Anapora Resolution in Japanese Nouns using Semantic Constraint, " Journal of Natural Language Processing, Vol.4, No.2, pp.41-56, 1997. (in Japanese)

[8] Megumi Kameyama, "A Property-Sharing Constraint in Centering, " Proc.of 24th Annual Meeting of Association for Computational Linguistics, pp.200-206, 1986.

[9] Barbara J.Grosz, Aravind K.joshi, Scott Weinatein, "Providing a Unified Account of Definite Noun Phrases in Discourse, " Proc.of 21st Annual Meeting of Association for Computational Linguistics, pp.44-50, 1983.

[10] B.Grosz and C.Sidner, "Attention, Intensions, and the Structure of Discourse, " Computational Linguistics, Vol.12, No.3, pp.175-204, 1986.

[11] Shingo Takeda and Norihisa Doi, "Focusing and Zero-Pronouns from the Viewpoint of Paragraphs, " The Transactions of the Institute of Electronics, Information and Communication Engineers D-II, Vol.J79, No.6, pp.1089-1097, 1996.

[12] Shingo Takeda and Norihisa Doi, "Centering in Japanese: A Step Towards Better Interpretation of Pronouns and Zero-Pronouns, " Proc.of COLING '94, pp.1151-1156, 1994.

[13] Bolinger. Dwight L, "The Atomization of Meaning, " Language 41, pp.555-573, 1965.

[14] Kenneth Ward Church, Patrick Hanks, "Word Association Norms, Mutual Information, and Lexicography, " Computational Linguistics, Vol.16, pp.22-29, 1990.

[15] Japan Electronic Dictionary Research Institute, EDR Japanese Corpus, Version 1.5, 1995.

[16] Nippongo Goi-Taikei, Iwanami Shoten, Tokyo, 1997.