ROBUST METHOD OF PRONOUN RESOLUTION USING FULL-TEXT INFORMATION

Tetsuya Nasukawa

IBM Research, Tokyo Research Laboratory
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken 242, Japan

Abstract

A consistent text contains rich information for resolving ambiguities within its sentences. Even simple syntactic information such as word occurrence and collocation patterns, which can be extracted from the text without deep discourse analysis, improves the accuracy of sentence analysis. Pronoun resolution is a typical proceeding that utilizes this information. Through the use of this information, along with information on the syntactic position of each candidate, 93.8% of pronoun references were resolved correctly in an experiment on computer manuals.

1 Introduction

Resolving pronoun reference is a difficult task that requires consideration of both linguistic and cognitive aspects of a language. As a linguistic phenomenon, the use of pronouns is treated as a co-referential problem in which both the antecedent and the pronoun co-refer to some object. From this point of view, finding the object that is co-referred to by a pronoun is the main problem in pronoun resolution, and much research has been devoted to focusing on or inferring the referent object by considering the grammatical and semantic roles of each entity in the sentences [Sidner, 1983; Brennan, 1987; Kameyama, 1993]. This task is especially difficult when the referent object is not explicitly stated in a text, and common sense and deep inference are required in order to figure out the object, as in the classic problems described by Charniak [Charniak, 1973]. Since this approach of considering the grammatical and semantic roles of each entity depends heavily on accurate syntactic analysis and semantic analysis, it is not yet applicable to practical systems.

However, if we do not aim for perfect analysis, a simple syntactic-based heuristic rule for selecting a correct antecedent from several candidate noun phrases performs quite well, especially in technical documents such as computer manuals, in which we can usually expect an explicit antecedent within the same sentence or in a previous sentence. In this domain, a correct antecedent can be selected in almost 90% of all cases without any world knowledge other than simple semantic constraints [Hobbs, 1978; Walker, 1989; Lappin, 1990]. Moreover, several heuristic rules can be combined to improve the accuracy of the analysis [Rich, 1988; Carbonell, 1988].

This approach of resolving pronoun reference by applying simple heuristic rules seems to be adequate for a practical natural language processing system, yet in order to achieve a success ratio of over 90%, some kind of knowledge processing is required, such as the use of world knowledge or deep inference mechanisms for constructing and referring to a discourse structure. While the advantage of knowledge processing is widely recognized, this approach presupposes a large quantity of knowledge resources, and leads to a knowledge acquisition bottleneck. In order to solve this problem, various studies have been done on methods of using on-line text databases with less human intervention for word sense disambiguation and structural disambiguation [Jensen, 1987; Nagao, 1990; Uramoto, 1991; Hindle, 1993]. These methods can be applied to knowledge processing in pronoun resolution; however, no research has yet revealed sufficient world knowledge to cover general problems. In other words, methods of using world knowledge have not reached a level sufficiently mature for them to be used in broad-coverage systems.

This paper proposes a simple and robust approach that utilizes inter-sentential information, extracted from a source text by means of a simple algorithm, to improve the accuracy of pronoun resolution. For example, collocation patterns within a text offer information that corresponds to case frames in world knowledge, and word frequency also gives information relevant to the topic or focus of the subjects. Thus, instead of using outside knowledge resources, such information serves as world knowledge appropriate to the narrow domain of the source text. The effectiveness of each type of information extracted from a source text is evaluated in the light of the results of experiments on computer manuals.

In the next section, we introduce three effective factors in the selection of an antecedent from candidate noun phrases. Then, in the third section, we

specify the implementation of this method. Finally, in the fourth section, we evaluate the effectiveness of this approach on the basis of the results of an experiment.

2 Three factors for evaluating salience in candidates

In our approach, pronoun resolution basically consists of collecting candidate noun phrases and selecting the most preferable candidate as the antecedent of a pronoun by applying several rules to filter out inappropriate candidates and to attach preferences to appropriate candidates. Rules are divided into two types. One type represents grammatical constraints that must be satisfied, such as number and gender agreement. Since rules of this type can filter out inappropriate candidates, we apply them at an early stage of pronoun resolution. The remaining rules constitute the other type, which attaches a preference to each candidate noun phrase. After inappropriate candidates have been filtered out by the former rules, the latter rules determine the most appropriate candidate by measuring the salience of each remaining candidate noun phrase. Thus, the latter rules are important for selecting the exact antecedent from the remaining candidates and for improving the accuracy of pronoun resolution.

In this section, we describe three effective factors that utilize full-text information for measuring the salience of each candidate noun phrase. The reasons for their effectiveness are that they cover many aspects of linguistic phenomena and that their interpretation is simple enough to be used in a practical system.

2.1 Collocation patterns within a source text

In previous approaches, semantic constraints have been among the most basic factors for filtering out candidates that would be inappropriate as modifiers of the modifiee of a pronoun. However, in order to apply semantic constraints with broad coverage, a large amount of knowledge is required. For example, in processing a sample sentence provided by Hobbs [Hobbs, 1978],

> *The castle in Camelot remained the residence of the king until 536 when he moved it to London,*

the following knowledge must be supplied in order to filter out the candidates *536*, *castle*, and *Camelot*, and leave the correct antecedent, *residence*:

- Dates cannot move.
- Places cannot move.
- Large fixed objects cannot move.

In order to apply this knowledge, we also presuppose a correct analysis that categorizes each noun phrase as a date, place, large fixed object, and so on. Since many of the words in these noun phrases have word sense ambiguities, it is not practical to presuppose the correct application of such knowledge. Assembling a large body of knowledge poses another major problem.

Instead of such world knowledge, collocation patterns (namely, modifiee-modifier relationships) extracted from a discourse can be applied. Since word sense is usually unified within a discourse, and most words with the same lemma are frequently repeated [Gale, 1992; Nasukawa, 1993], the collocation patterns in the same discourse provide valuable data for determining whether a candidate can modify the modifiee of a pronoun. For example, if the sentence

> *He moved his residence*

is found in the discourse, this information indicates that the word *residence* can be the object of the verb *move*. Thus, the information works as a selectional constraint that the candidate can be an argument of a predicate that dominates the pronoun.

Moreover, since statements tend to be repeated in a discourse, the existence of an identical collocation pattern in a discourse may support selection of the candidate as the antecedent. In this sense, the preference for an identical collocation pattern also reflects *case role persistence* preference and *syntactic parallelism* preference, proposed by Carbonell and Brown [Carbonell, 1988]. The *case role persistence* rule prefers a candidate noun phrase that filled an identical case role in an earlier sentence. For example, after the sentence

> *Mary gave an apple to Susan,*

Susan is referred to by *her* in

> *John also gave her an apple,*

while *Mary* is referred to by *she* in

> *She also gave John an apple.*

The *syntactic parallelism* rule prefers a candidate noun phrase that preserves its surface syntactic role from the first of two or more coordinate clauses. For example, in

The girl scout leader paired Mary with Susan, but she had paired her with Nancy last time,

she refers to *leader*, and *her* refers to *Mary*, whereas in

The girl scout leader paired Mary with Susan, but she had paired Nancy with her last time,

she refers to *leader*, and *her* refers to *Susan*. By referring to the identical collocation patterns, we can resolve all the pronouns in the above examples correctly.

Since the identification of modifier-modifiee relationships is a basic feature of syntactic analysis, a procedure for identifying identical collocation patterns is not a hard task, as long as all of the sentences are parsed by a single system and share a single formalism for expressing modifier-modifiee relationships.

2.2 Frequency of repetition in preceding sentences

A characteristic of the pronominalization on which the centering approach [Sidner, 1983; Brennan, 1987; Kameyama, 1993] is based is that an object in focus is likely to be pronominalized. If this characteristic is expanded to all definite anaphoras, which include definite noun phrases as well as pronouns, a candidate noun phrase that is in focus may be repeated as a definite noun phrase before it is pronominalized. Thus, the frequency in preceding sentences of a noun phrase with the same lemma as a candidate noun phrase can be an index for the preference with which it is selected as the antecedent. The process for assigning this preference consists of a simple string match that checks words with the same lemma in preceding sentences.

In addition, when the source text is marked up with SGML or other such tags, the roles of some phrases such as titles and headings can be easily recognized, and words with such roles tend to represent the topics of the sentences following them. Thus, additional preference can be assigned by checking the tags of each word.

2.3 Syntactic position

As shown by Hobbs [Hobbs, 1978], a heuristic rule favoring subjects over objects performs remarkably well in English text. By traversing the surface parse tree of a sentence, a preference value can be provided for each candidate noun phrase according to its syntactic position. This factor has an advantage over other factors shown in previous subsections in the sense that it assigns a definite ranking for each candidate noun phrase, since each occupies a syntactic position in a text. Thus, this factor provides a default value for the preference of each candidate noun phrase when no other factor provides valid information, and it is adequate for a robust approach since it is basically assigned by traversing the surface parse trees of a sentence.

3 Implementation

In this section, we describe the actual implementation of the pronoun resolution procedure in an English-to-Japanese machine translation system, Shalt2 [Takeda, 1992].

The procedure consists of two steps:

1. Extraction of candidates for an antecedent
2. Selection of the correct antecedent from the candidates.

In order to achieve higher accuracy in pronoun resolution with robust processing, our strategy consists of

1. Extending a list of candidate noun phrases so that it does not exclude a correct antecedent
2. Referring to all information in the source text that can be syntactically extracted without referring to outside knowledge resources, in order to select the correct antecedent.

3.1 Extraction of candidates

To ensure that the correct antecedent is included in a list of candidate noun phrases, candidates are extracted from exactly two sentences with minimum filtering. First, the system checks whether any noun phrases earlier in the same sentence satisfy the number and gender constraints. It then checks the preceding sentences in order of proximity until candidates have been found in exactly two sentences.

During the extraction of the candidates, the system filters out noun phrases that do not satisfy the number and gender constraints, and also direct modifiees of the pronoun and its arguments, so that a non-reflexive pronoun and its antecedent may not occur in the same simplex sentence, as would be the case if *data* were the antecedent of *it* in the following sentence:

The device that writes onto a magnetic disk and reads data from it is called a disk drive.

3.2 Selection of an antecedent

In our implementation, the preference values provided by the algorithms described in the following paragraphs are combined into a single value, and the candidate noun phrase with the largest preference value is selected as the antecedent.

Preference according to the existence of identical collocation patterns in the text

As a preference value that indicates the satisfaction of selectional constraints and repetition of an identical statement, we assigned a constant value 3 for a candidate that has an identical collocation pattern with the modifiee of a pronoun within the source text. Furthermore, in order to extend the use of collocation patterns as knowledge on selectional constraints, an on-line synonym dictionary [3] is referred to, and thus a collocation pattern with a synonym can support candidates other than exactly identical collocation patterns.

Preference according to the frequency of repetition in preceding sentences

In order to provide a larger preference value for closer and more frequent occurrences of a lemma, the preference value is given by the total score calculated according to the following formula, for each appearance of a noun phrase with the same lemma as the candidate noun phrase that is found within the ten preceding sentences:

$$\frac{1}{(\textit{Number of sentences to the identical noun phrase})+1}.$$

Preference according to syntactic position

Among the candidate noun phrases, a candidate in a closer sentence, or the one nearest the beginning of the same sentence is preferred. Besides the left-to-right order within a sentence, a negative preference value is given for the distance (number of sentences) from the sentence that contains the pronoun to the sentence that contains the candidate. While the order of preference of candidates that is obtained in this manner is similar to that given by the naive algorithm proposed by Hobbs [Hobbs, 1978], our algorithm is much simpler, and does not even require the results of syntactic analysis.

3.3 Example

Figure 1 gives an example of system output that contains data on the preference of each candidate antecedent for a pronoun in a sample text extracted from the second chapter of a computer manual [2] in the manner described in the previous paragraphs.

In this figure, the number in brackets before each sentence indicates the sentence number in the text. As shown by these numbers, the output consists of eleven consecutive sentences, from the 104th to the 114th in the second chapter of the manual.[1] The order of candidates following the message

```
Candidates for the referent of CFRAME106579
("it") are:
```

reflects the order of preference values obtained by referring to the position of each candidate. As shown in this list, *key* is the most preferable candidate from the viewpoint of syntactic position. In this candidate list, CFRAME␣␣␣␣␣␣␣ indicates an instance of each content word in the discourse. Information on the position and on the whole sentence can be extracted from each of these CFRAMEs. A number in arrowhead brackets next to CFRAME␣␣␣␣␣␣␣, such as <113>, indicates the number of the sentence in which it occurs. A number in parentheses, such as 0.48571432 in key (0.48571432), indicates the preference value obtained by referring to the frequency of repetition.[2] Thus, from the viewpoint of the number of repetitions, *cursor* is the most preferable candidate for the antecedent. At the bottom of this figure, information on modifier-modifiee relationships that support candidates is shown. In this case, there is a collocation pattern such that *cursor* modifies the verb *reaches*, which is the modifiee of the pronoun *it*; thus, this information prefers *cursor* for the antecedent of *it*. Finally, after combination of all the preferences, *cursor* is selected as the most preferable antecedent of *it*.

4 Results

We examined 112 third-person pronouns in 1904 consecutive sentences from eight chapters of two different computer manuals. One [1] is a typical computer manual for computer experts such as programmers and system operators, and the other [2] is a primer for novice users of a computer.

In this experiment, we excluded instances in which *it* pronominalizes a sentence, as in

do it,

those in which *it* refers to a syntactically recoverable *that*-clause or *to*-infinitive clause, and those in which

[1] In the sentences contained in Figure 1, the underlining and the change of font for the target pronoun *it* were done by the author.

[2] Noun phrases with the same lemma referred to in the preceding sentences are indicated by underlines.

(104) *All four of the cursor movement keys are typematic;*
(105) *they keep repeating as long as they are held down.*
(106) *The Cursor Up key moves the cursor up one line.*
(107) *Like the other cursor movement keys, this key moves the cursor one line or many lines depending on how long you hold down the key.*
(108) *The Cursor Right key moves the cursor to the right.*
(109) *Hold the key down.*
(110) *When the cursor reaches the right end of the line, it goes off the screen and reappears on the left side, one line below the line it was on.*
(111) *If the cursor is on the bottom line of the screen and is run all the way to the right, it goes off the screen and reappears in the upper left corner.*
(112) *The Cursor Left key moves the cursor one position to the left.*
(113) *Hold this key down.*
(114) *When* **it** *reaches the left end of the line, it goes off the screen and reappears on the right, one line above the line it was on.*

```
Candidates for the referent of CFRAME106579("it") are:
     CFRAME106573<113> .... key (0.48571432)
     CFRAME106564<112> .... Cursor Left key (0)
     CFRAME106565<112> .... cursor (1.4337664)
     CFRAME106568<112> .... left (0)

>> With DIANA <<
  >> To support CFRAME106565(cursor) <<
   1 with SAME-ATTACHMENT-CAND-MODIFIEE in:
     "  When the cursor reaches the right end of the line,
        it goes off the screen and reappears on the left side,
        one line below the line it was on ."
      (reaches<CFRAME106454> in SENTENCE106453<No.110>)
```

Fig. 1: Sample data for resolving the first pronoun *it* in sentence 114

it occurs in a time or weather construction. When an identical pronoun is included in the candidate list, the system assumes that these pronouns share the same antecedent. For example, we assumed that all the instances of *it* in

> *When it reaches the left end of the line, it goes off the screen and reappears on the right, one line above the line it was on*

have the same antecedent.

As a result of our strategy of enlarging the scope for selection of candidates, the average number of candidate noun phrases was 4.1.

Our algorithm chose a correct antecedent in 105 cases, giving a success ratio of 93.8%. In 28 of those 112 cases, there was among the candidates an identical pronoun that referred to the same antecedent; thus, in 84 cases, antecedents were selected by evaluating the syntactic position, frequency of repetition, and collocation pattern of each candidate noun phrase.

As shown in Table 1, without any information on repetition or collocation patterns, the success rate of selection based only on syntactic position was 82.7%, while the success rate for selection based only on frequency of repetition was 60.7%. This result indicates that pronominalized noun phrases were actually repeated more than twice within ten consecutive sentences in over 60% of the cases. Thus, preference according to the frequency of repetition contributed to the selection of the correct antecedent. In 16 of the 22 cases in which this information preferred a wrong candidate noun phrase, the preference value was overridden by the negative preference value caused by a syntactic position far from the sentence of the pronoun, or by a larger preference assigned to some other candidate with an identical collocation pattern.

Identical collocation patterns were found within the same chapter in 22 of 84 cases in which the preference value was evaluated in selecting an antecedent. Although this is only 26.2% of the cases, collocation preference did not support any wrong candidates. Moreover, in 50.0% of the 22 cases, another preference value, either syntactic position or repetition, supported a wrong candidate. Therefore, preference according to collocation pattern contributed to the selection of the correct antecedent.

Table 2 shows the distances and directions of sen-

Table 1: Correlation between correct selection and selection in accordance with each type of preference

	Number of cases in which the correct antecedent was selected	Number of cases in which the wrong antecedent was selected	Number of cases without any valid information
Syntactic position	69 (82.1%)	15 (17.9%)	0
Frequency of repetition	51 (60.7%)	22 (26.2%)	11 (13.1%)
Existence of similar collocation	22 (26.2%)	0 (0%)	62 (73.8%)

Table 2: Distribution of information relative to a sentence that contains information for modification preference

Forward	Distance (number of sentences)	0 - 5	6 - 10	11 - 15	16 - 20	21 - 25	26 -
	Number of occurrences	10	4	1	4	2	0
Backward	Distance (number of sentences)	1 - 5	6 - 10	11 - 15	16 - 20	21 - 25	26 -
	Number of occurrences	12	1	0	2	0	0

tences in which a collocation pattern supporting a candidate noun phrase to modify the modifiee of the pronoun was found. The results indicate that such information was extracted from a relatively small area of a text. In addition, relative collocation patterns were extracted from both previous and following sentences.

Out of the 37 cases in which the identical collocation patterns were found, synonym relations were used in seven cases (18.9%).

5 Conclusion

We have proposed a robust method of pronoun resolution that refers to information within the source text in order to determine the preference value of each noun phrase that is a candidate for selection as the antecedent of a pronoun. This approach is practical in terms of the amount of knowledge it presupposes and the amount of computation it requires, since it basically relies only on the surface information in a text, and is free from the knowledge acquisition bottleneck.

In experiments on computer manuals, we achieved a success rate of 93.8%. A remarkable aspect of this result is that we achieved it without referring to any outside knowledge resource except for the synonym relations in an on-line synonym dictionary. By combining heuristic rules to utilize various information extracted from all the sentences in the source text, high accuracy can be achieved in pronoun resolution for a practical natural language processing system. The advantages of this approach are that a simple algorithm can extract information on syntactic position, repetition, and collocation patterns by referring to morphological information within a source text, and that it does not even assume a correct syntactic analysis or depend on the formalism of syntactic parse trees, since it does not rely on any grammatical information except for modifier-modifiee relationships. This approach is especially effective in technical documents such as computer manuals or patent documents in which words are used consistently in order to avoid ambiguity, and in which identical collocation patterns are frequently repeated in detailed descriptions of target objects or procedures.

Acknowledgements

I would like to thank Michael McDonald for invaluable help in proofreading this paper. I would also like to thank Taijiro Tsutsumi, Masayuki Morohashi, Koichi Takeda, Hiroshi Maruyama, Hiroshi Nomiyama, Hideo Watanabe, Shiho Ogino, Naohiko Uramoto, and the anonymous reviewers for their comments and suggestions.

References

[Brennan, 1987] Brennan, S. E., M. W. Friedman, and C. J. Pollard (1987). A Centering Approach to Pronouns. *In Proceedings of ACL-87.*

[Carbonell, 1988] Carbonell, J. G., and R. D. Brown (1988). Anaphora Resolution: A Multi-Strategy Approach. *In Proceedings of COLING-88.*

[Charniak, 1973] Charniak, E. (1973). Jack and Janet in Search of a Theory of Knowledge. *In Proceedings of IJCAI-73.*

[Gale, 1992] Gale, W. A., K. W. Church, and D. Yarowsky (1992). One Sense Per Discourse. *In Proceedings of the 4th DARPA Speech and Natural Language Workshop*

[Hindle, 1993] Hindle, D., and M. Rooth (1993). Structural Ambiguity and Lexical Relations. *Computational Linguistics, Vol. 19, No. 1.*

[Hobbs, 1978] Hobbs, J. R. (1978). Resolving Pronoun References. *Lingua, 44.*

[Jensen, 1987] Jensen, K., and J.-L. Binot (1987). Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions. *Computational Linguistics, Vol. 13, No. 3-4.*

[Lappin, 1990] Lappin, S., and M. McCord (1990). Anaphora Resolution in Slot Grammar. *Computational Linguistics, Vol. 16, No. 4.*

[Kameyama, 1993] Kameyama, M., R. Passonneau, and M. Poesio (1993). Temporal Centering. *In Proceedings of ACL-93.*

[Nagao, 1990] Nagao, K. (1990). Dependency Analyzer: A Knowledge-based Approach to Structural Disambiguation. *In Proceedings of COLING-90.*

[Nasukawa, 1993] Nasukawa, T. (1993). Discourse Constraint in Computer Manuals. *In Proceedings of TMI-93.*

[Rich, 1988] Rich, E.A. and S. LuperFoy (1988). An Architecture for Anaphora Resolution. *In Proceedings of ANLP-88.*

[Sidner, 1983] Sidner, C. L. (1983). Focusing in the Comprehension of Definite Anaphora. *In Computational Models of Discourse, M. Brady and R. Berwick, eds., Cambridge, Mass.: MIT Press.*

[Takeda, 1992] Takeda, K., N. Uramoto, T. Nasukawa, and T. Tsutsumi (1992). Shalt2 - a Symmetric Machine Translation System with Conceptual Transfer. *In Proceedings of COLING-92.*

[Uramoto, 1991] Uramoto, N. (1991). Lexical and Structural Disambiguation Using an Example-Base. *In Proceedings of the 2nd Japan-Australia Joint Symposium on Natural Language Processing.*

[Walker, 1989] Walker, M. (1989). Evaluating Discourse Processing Algorithms. *In Proceedings of ACL-89.*

[1] "IBM SAA ImagePlus Object Distribution Manager MVS/ESA High-Speed Capture Subsystem Guide Version 2 Release 1.1," IBM Corp. (1991)

[2] "IBM Application System/400 New User's Guide Version 2," IBM Corp. (1992)

[3] "The New Collins Thesaurus," Collins Publishers, Glasgow (1984)