

Discourse Processing in MT:

Problems in Pronominal Translation

Hajime Wada : wadah@yvax.byu.edu

Department of Linguistics, The University of Texas at Austin
and

ECS, 455 North University, Suite 202, Provo, UT 84601, U.S.A.

Keywords: Machine Translation, Anaphora, Discourse Representation Theory, Focus

1. Introduction

Translation of anaphoric expressions has been problematic in most of the MT systems (Key 1986). One of the main reasons for the difficulties lies in the lack of discourse information representation in the MT systems. In this paper, we report an implementation of the Discourse Representation Theory in an LFG-based English-to-Japanese MT program, and discuss problems in translating anaphoric expressions in this system.

2. Problems in Translating Anaphora

Problems in translation of anaphoric expressions can be seen on three different but interactive levels of linguistic information: lexical, syntactic and pragmatic.

The main problem on the lexical level is due to a difference in the language specific parameters in the pronominal system such as F features (person, gender, number, etc.). Surface forms of pronominals depend on the F features of their antecedents, so that the translation of a particular pronominal form cannot be determined solely by the pronominal itself. For example, 'ship' in English is feminine, but its translation, 'hune', is neutral in Japanese. Thus, a proform 'she' for 'ship' should not be translated as 'kanojo' (3rd, sing, fem), but as 'sore' (3rd, sing, neut).

Problems on the syntactic level are mainly due to a difference in the distribution of anaphoric expressions. Mapping relations between English pronominals and their Japanese counterparts are shown below:

English	Japanese
Overt Pronoun	Overt Pronoun
Reflexives	zibun
	X-zishin
Empty Pronoun	Control-Zero
	Topic-Zero

The most crucial case is English overt pronouns. As Kuroda noted (1965:107), overt pronouns are not repeatable in Japanese, while they are in English.

(1)

- a. **He** does **his** work when **he** feels like doing it.
b. ***Kare-wa kare-ga** sitai toki-ni **kare-no**
he TOP he NOM like to do when his GEN
sigoto-wo suru.
work ACC do

Instead of repeating overt pronouns, zero pronominals and 'zibun' should be used though there are some variations in their use.

- c. **Kare-wa** \emptyset sitai toki-ni \emptyset sigoto-wo suru.
d. **Kare-wa zibun-ga** sitai toki-ni \emptyset sigoto-wo suru.

e. **Kare-wa** \emptyset sitai toki-ni **zibun-no** sigoto-wo suru.

Resumptive pronouns are also problematic. Possessive relative pronoun 'whose' in English must be translated as a resumptive pronoun whose antecedent is the head NP of a relative clause in Japanese.

The last set of problems is on the level of pragmatics. It has been known that in many languages certain types of pronominal expressions represent a focus in a discourse. For example, in English, overt pronouns play this role while zero pronouns (topic-zero) do the job in Japanese. In a certain discourse context, a discourse focus should be translated as a reflexive 'zibun' instead of a zero pronoun.

Some cases may be induced to lexical properties of predicates and pronominals such as functional and pseudo-functional control cases and reflexivity. Nonetheless, most of the above problems are not easily separable; they should better be solved in a unified model.

3. Treatment of Pronominal Translation

In this paper, we will focus on the problems of translating English overt non-reflexive pronominals into Japanese. These pronominals must be mapped into 1) overt pronominals, 2) 'zibun', 3) *pro*, or 4) discourse *pro* in Japanese.

3.1. Role of Discourse

To explain some exceptions on locality conditions and ambiguous cases in anaphora resolution, recent studies have focused on the role of discourse (Grosz 1981; Sidner 1983; Joshi & Weinstein 1981; Kameyama 1985; Roberts 1986) and world knowledge. Discourse factors affecting anaphora resolution are discourse structure and discourse focus/center. This report concerns only the latter; in particular, the selection of focus and its maintenance.

Following Grosz, we assume that a focus is the most salient NP in a sentence, and every sentence has one and only one focus¹. Often a focus is marked by the use of pronominal, stress, thematic role and particular syntactic constructions such as clefting, question, passivization, etc. However, there seems no clear generalization on the organization of the focus marking factors. We simply assume the most marked NP with respect to the number of found factors (Asher & Wada 1989; Alshawi 1987).

We have adopted the Discourse Representation Structure (Kamp 1981) to represent a current discourse structure which will be appended to an f-structure representing a sentence. DRS+f-structure gives a ground for examining logical conditions based

¹ We concern a local focus only, and multiple foci cases are excluded from our discussion.

on the accessibility condition, and locality conditions such as disjoint reference on each anaphoric expression (cf. Roberts 1986).

Now consider 4) discourse *pro*. Its characteristic is that it may not be bound in the current sentence, but must be bound by the focus in the previous discourse. In other words, when a pronominal cannot find its antecedent in the current DRS, we assume that it must be bound by the previous discourse focus unless the focus has been shifted, and in such a case we translate it as \emptyset , a lexical zero pronominal.

3.2. Proform Selection by FU

On the other hand, when an antecedent is found in the DRS, the anaphoric link will be noted not only in DRS, but also in relevant f-structures via features such as Ant(ecedent) and Desc(endent) with a common index. However, we still need to determine which form should be used: *pro*, 'zibun', or overt pronominals. The selection should be syntactically made on the Japanese side since no additional information is available for the selection in the English f-structures.

Let us discuss the syntactic mechanism for the proform selection. Suppose that a proform be lexically presented in each noun in the lexicon such as follows²:

PRED 'man'	PRED 'pro'
GENDER msc	GENDER msc
PROFORM 'he'	PROFORM 'he'

When the two shown above are anaphorically linked in a sentence, LFG offers a syntactic solution, namely, the Functional Uncertainty (Kaplan & Maxwell 1988)³. The link is represented by a functional-application equation (FU) such as follows: (2) $\langle \wedge /GF^* \backslash GF^* \text{PROFORM} \rangle = \langle \wedge \text{PROFORM} \rangle$. In the above FU, the second GF after the right slash represents an f-structure that dominates a possible antecedent, and the first GF* after the left slash, f-commanding path to the antecedent from the current f-structure⁴. That is, the FU imposes a constraint such that the right-hand side of the equation in (2) represents an f-structure containing a pronominal, and if there is an f-structure within the path represented by the left-hand side of the equation, then it must be the case that their PROFORMs are unifiable⁵.

For the generation of surface PROFORMs, we assume that a pronominal element possesses a lexically unspecified PROFORM, and it will be specified by its antecedent. In other words, given that a pronominal and its antecedent are coindexed, an FU such as follows should be satisfied.

$$(3) \quad \langle \wedge /GF^* \backslash GF^* \text{Ant} \rangle = \langle \wedge \text{Desc} \rangle$$

This equation can be solved only when the indices are identical. Once the equation (3) is solved, PROFORM will be unified between the antecedent and the pronominal⁶.

Selection on *pro* and 'zibun' can be considered as solving particularly restricted FU equations. For example, 'zibun' will be selected when the following FU equation is satisfied⁷.

$$(4) \quad \langle \wedge /GF(\text{ADJUNCT})GF^* \backslash \text{SUBJ Ant} \rangle = \langle \wedge \text{Desc} \rangle$$

This equation says that if there exists a SUBJ that f-commands GF(s) that may contain an adjunct which includes an f-structure in which the current *pro* resides, then it must be the case that the Ant value of that SUBJ and Desc value of the *pro* are identical. Each unspecified 'pro' will carry a set of equations, each of which selects a particular PROFORM disjunctively.

$$(5) \quad \{ [\langle \wedge /GF(\text{ADJUNCT})GF^* \backslash \text{SUBJ Ant} \rangle = \langle \wedge \text{Desc} \rangle \\ \langle \wedge \text{PROFORM} \rangle = \text{'zibun'}] \\ [\langle \wedge /GF^* \backslash \text{ADJUNCT} \backslash \text{SUBJ Ant} \rangle = \langle \wedge \text{Desc} \rangle \\ \langle \wedge \text{PROFORM} \rangle = \emptyset] \\ [\langle \wedge /GF^* \backslash GF^* \text{Ant} \rangle = \langle \wedge \text{Desc} \rangle \\ \langle \wedge \text{PROFORM} \rangle = \langle \wedge \text{PROFORM} \rangle] \}$$

4. Implementation of the Treatment

The anaphora resolution mechanism in our program consists of three functional units: the DRS construction (DRSCONS), the salient element storing (SLSTOR), and the antecedent searching (ANSR).

4.1. DRSCONS

This module constructs DRSS compositionally, following Montagovian tradition (Wada & Asher 1986, Reyle 1988). Thus, we assume that a semantic representation, SMS, is appended to each F-structure. SMS is a pair $\langle \text{SemType}, \text{DRS} \rangle$, where SemType is a semantic type.

4.2. SLSTOR

SLSTOR's main function is to store the most salient, focused element in the current discourse for the next sentence processing (Alshawi 1987; Asher & Wada 1989). In order to find the most salient NP, SLSTOR sets three kinds of filters among others such as grammatical function, use of pronominal, and syntactic construction (See Asher & Wada 1989), and check all of the NPs appeared in the current sentence with respect to the three filters.

After the most salient element has been selected, SLSTOR checks whether the current focus has been shifted or retained by comparing the current SLSTOR value and the newly selected one. If their gender, number, person, etc. are unifiable, the discourse focus is considered to retain and SLSTOR keeps the previous NP in the storage. If they are not unifiable, the focus has been shifted, and SLSTOR stores the newly selected element as a current focus.

4.3. ANSR

⁶Resumptive pronouns are treated in the same way shown above, except for the co-indexation. Since an antecedent for a resumptive pronoun is always the head NP of the relative clause and therefore, the co-indexation is not necessary.

⁷This is the condition that Kuroda (1965) proposed, and still has been a basically correct generalization.

²Plural PROFORMs should be assigned to NPs syntactically.

³The same mechanism has been used in other cases such as the noun classifier selection for numeral expressions and negative scope domain setting by 'wa' in Japanese (Wada forthcoming).

⁴This has been done by pointing the antecedent from the current f-structure.

⁵Note that the equation is far more powerful than this description.

This module consists of three functions. The first function is SEARCH. SEARCH searches antecedents by testing the accessibility on the DRS and morpho-syntactic constraints such as gender, number, and binding features (sub, nuclear, etc) on the current f-structure. If a search in the DRS fails, SEARCH further searches in SLSTOR, and check a focused element in SLSTOR. In effect, we distinguish pronominals into three classes according SEARCH's result; 1) an antecedent found in the current DRS, 2) an antecedent not found in the current DRS, but controlled by a discourse focus, and 3) an antecedent not found in either in the current DRS or in the previous focus.

The second function is LINK. In the case in 1) above, LINK will set a unique anaphoric index in both F-structures of the antecedent and its pronominal and assigns an unspecified 'pro' to the pronominal.

The last function is SETPRO, which assigns \emptyset to the pronominal in the case of 2), and assigns a default word-for-word translation in the case of 3).

5. Examples

Let us take some examples. First, consider the following donkey sentence.

(6-a)

Every man who loves a ship treats her tenderly.
'Her' will be linked with 'a ship' by SEARCH due to the accessibility of 'a ship' to 'her' on the DRS. Then, the same anaphoric index is assigned to the transfer f-structure of 'a ship' and of 'her' by LINK. 'Her' is translated as an unspecified 'pro' as shown below.

(6-b)

Hune wo aisuru otoko wa mina sore wo itukushimu.
ship ACC love man TOP every it ACC treat-tender
In this result, an overt pronominal, 'sore', is selected for 'pro' since the f-structure for 'hune' and 'pro' share an anaphoric index, and the last clause in the FU equation in (5) is satisfied.

Suppose that the next sentence appeared discourse-initially.

(7-a) John believes that she is a genius.

Since there is neither a sentence-internal antecedent nor a previous focus, 'she' gets a default translation, 'kanojo'. However, if the following sentence precedes the sentence in (7-a), a lexical \emptyset will be selected as the target for 'she' because 'Mary' is the focus of (7-b) since it is a SUBJ and appears in a passive construction.

(7-b) Mary was awarded as a distinguished artist.

Now, consider the next example.

(8-a) John works when he wants to do his work.

In the above example, 'his' should be linked to 'he' and 'he' to 'John'. Each anaphoric index has a different value, but both of the pronominals will be translated as unspecified 'pro's. However, the equation (5) gives them different PROFORMs, 'his' 'zibun' and 'he' \emptyset . This is indeed a desired result as shown below.

(8-b)

John-wa \emptyset zibun-no sigoto-wo sitai-toki-ni hataraku.
John TOP POSS work ACC want-to-do when work

6. Remaining Problems

A note on defects and shortcomings is in order.

It is no doubt that world knowledge plays an important role in anaphora resolution. We assume that in the near future, a world knowledge inferencing

module should become available, and that it will be incorporated into our system.

We assumed that an input text as a whole consists a discourse segment; however, this is due to the lack of a treatment in discourse segmentation in our program. One related problem is on the distinction of the global focus and local foci. Since our program considers only one segment at a time, there is no way to handle complex foci.

Another area which we need to handle but has been neglecting is number related expressions such as plural antecedents and split antecedents.

Also, the salient element selection mechanism in our program is still based on a crude stipulation. We need to conduct further research on this topic.

Finally, it is desirable to map a source DRS to the target DRS and handle language-specific discourse-based expressions. These will be our next task.

References

- Alshawi, H. (1987) *Memory and Context for Language Interpretation*, Cambridge University Press.
- Asher, Nicholas (1986) "Belief in Discourse Representation Theory", in *Journal of Philosophical Logic*, 5, pp.127-189.
- Asher, N. & H. Wada (1989). "Computational Account of Anaphora", ms, CGS, Univ. of Texas.
- Grosz, B. (1978). "Discourse Knowledge", in *Understanding Spoken Language*, ed. by Donald Walker, pp.228-345, .
- Grosz, B. (1981). "Focusing in Dialogue", in *TINLAP-2: Theoretical Issues in Natural Language Processing*. pp.96-103. ACM and ACL. N.Y.
- Joshi, A. & S. Weinstein (1981) "Control of Inference: Role of Some Aspects of Discourse Structure-Centering", in proceedings of 7th IJCAI, pp.385-387
- Kameyama, Megumi. (1985). *Zero Anaphora: The Case of Japanese*. Ph.D. dissertation. Stanford University.
- Kamp, H. (1981) "A Theory of Truth and Semantic Representation", in Groenendijk and Janssen (eds), *Formal Methods in the Study of Language*, Gronningen:Amsterdam, pp.1-41.
- Kamp, H. (1983) "Situation in Discourse", ms, CSLI.
- Kaplan, R. & J. Maxwell (1988). "An Algorithm for Functional Uncertainty", in the proceedings of COLING-88, pp.297-302.
- Kay, M (1986) "Machine Translation will not work", in the proceedings of ACL, p.268.
- Kuroda, S-Y. (1965) *Generative Grammatical Studies in the Japanese Language*, Ph.D. dissertation, MIT.
- Roberts, C. (1986) *Modal Subordination, Anaphora, and Distributivity*, Ph.D.dissertation, UMASS.
- Reyle, W. (1988) "Computational Semantics for LFG", in *Natural Language Parsing and Linguistics Theories*, eds by U. Reyle and C Rohrer, pp. 448-474.
- Sidner, Candice (1983). "Focusing in the Comprehension of Definite Anaphora", in *Computational Models of Discourse*, eds by R. Berwick and M. Brady, pp.267-330.
- Wada, Hajime (forthcoming). "Applications of Functional Uncertainty in ECS System", ms.
- Wada, Hajime & N. Asher (1986). "BUILDRS: An Implementation of DR Theory and LFG", in the proceedings of COLING-86, pp.540-546.