ANAPHORA RESOLUTION IN MACHINE TRANSLATION

Ruslan Mitkov and Sung-Kwon Choi
IAI
Martin-Luther-Str. 14
D-66111 Saarbrücken
{ruslan, choi}@iai.uni-sb.de

Randall Sharp
DGSCA UNAM
Apdo. Postal 20-059
04510 Mexico, D.F.
randy@servidor.unam.rnx

Abstract

This work addresses the as yet insufficiently covered problem of Anaphora Resolution in Machine Translation. To start with, the paper discusses the translation of pronominal anaphors, the necessity of anaphora resolution in Machine Translation (MT) and previous work in the field. Next, it briefly presents the MT system CAT2 and reports on an anaphora resolution model developed as an extension to the system. Finally, some implementation issues of the model are outlined and its further improvement is discussed.

1 Introduction

Progress in MT has been observed at different levels, but discourse has yet to make a breakthrough. MT research and development has concentrated so far mostly on sentence translation (discourse analysis being a very complicated task) and the successful operation of most of the working MT systems does not usually go beyond the sentence level.

One of the most important aspects in successfully analyzing multisentential texts is the capacity to establish the anaphoric references to preceding discourse entities. The paper will discuss the problem of anaphora resolution from the perspective of MT (we have concentrated so far on pronominal anaphora) and will present an integrated model for anaphora resolution, developed for the needs of the MT system CAT2.

2 Anaphora Resolution and Machine Translation

Besides ambiguity resolution, often seen as the most important problem in MT [Hutchins & Somers 92], another major difficulty is the resolution of anaphora.

The identification of pronouns involves the identification of the earlier noun phrases to which they refer, called the pronoun's antecedent or referent (we will use these terms interchangeably in the paper). The establishment of the antecedents of anaphors is very often of crucial importance for correct translation. When translating into languages which mark the gender of pronouns for example, it is essential to resolve the anaphoric relation.

Furthermore, the translation of the predicates connected with the pronoun (verbs, nouns etc.) may change according to different antecedents.

Anaphora resolution reflects two essential topics in Machine Translation: ambiguity in a MT context and translation of discourse instead of isolated sentences. Anaphora can be viewed as a sort of ambiguity, in that the antecedent of a given pronoun might be uncertain and referential relations are one of the means that constitute coherence of texts.

2.1 Translation of Pronominal Anaphors

In the majority of language pairs and cases the pronouns in the source language are translated by target language pronouns which correspond to the referent of the anaphor. However, there are various exceptions. In some languages the pronoun is translated directly by its referent. In English to Malay translation for instance, there is a tendency to replace 'it' with its referent. Replacing a pronominal anaphor with its referent means, however, that the translator (program) must be able to identify the referent first.

Very often pronominal anaphors are simply omitted in the target language. For example, though the English personal pronouns have their equivalences in Spanish, they are frequently not translated because of the typical Spanish elliptical zero-subject, constructions.

Another interesting example is English-to-Korean translation. The English pronouns can be omitted elliptically, translated by a definite noun phrase, by their referent, or by one or two possible Korean pronouns, depending on the syntactic information and semantic class of the noun the anaphor refers to ([Mitkov et al. 94]).

2.2 The Necessity of Anaphora Resolution in Machine Translation

Whereas in most European language pairs anaphora resolution is "compulsory" (or else we risk rendering in certain cases quite unacceptable translations), there are certain language pairs and cases where anaphora resolution may seem "optional".
Consider the following sentences [Hutchins & Somers 92]:

(1) *The monkey ate the banana because it was hungry.*
(2) *The monkey ate the. banana because it was ripe.*
(3) *The monkey ate the banana because it was tea-time.*

In each case the pronoun *it* refers to something different: in (1) the monkey, in (2) the banana and in (3) to the abstract notion of time. If we have to translate the above sentences into German, then anaphora resolution is inevitable, since the pronouns take the gender of their antecedents, and since the German words *Affe* (masculine, "monkey"), *Banane* (feminine, "banana") and *es* (neutral, "it" for time notion) are of different gender.

Consider the translation of the sentences (1)-(3) from English to Korean and their literal descriptions [Mitkov et al. 94]:

(1') Pae.go.pa.so won.sung.yi.nun pa.na.na.rul mo.got.ta.
hungry-CAUSAL monkey-NOM banana-ACC eat-PAST,DECL
(2') I.go.so won.sung.yi.nun pa.na.na.rul mo.got.ta.
ripe-CAUSAL monkey-NOM banana-ACC eat-PAST,DECL

(3') Teatime.i.o.so won.sung.yi.nun pa.na.na.rul mo.got.ta.
tea time-CAUSAL monkey-NOM banana-ACC eat-PAST,DECL

Note that in the above Korean translations there are no pronouns. These examples might seem encouraging that we could translate from English to Korean, bypassing the tough problem of anaphora resolution. However, such a conclusion would be too misleading.

The assumption that anaphoric expressions in the source language can be easily mapped to the corresponding anaphors in the target language, or in many cases that they can be simply ignored in the transfer phase, is unfounded. It is not hard to find English sentences for which anaphora resolution is necessary in order to get their correct translation into Korean. Consider the sentences:

(4a) *Although programmers usually write good programs, they may still make a mistake.*

(4b) *Although programs are usually written by good programmers, they may still contain mistakes.*

In Korean, there are two types of pronominals corresponding to *they,* one for human beings and the other for non-humans. In order to assign the proper Korean pronominals to the English pronominal *they,* the system should be able to resolve *they* between the two possible referents, *programmers* and *programs.*

Anaphora resolution becomes a more serious business when we aim at achieving high-quality translation. The translation of (4a) and (4b) into Korean with the successful assignment of pronouns may still sound awkward to Koreans, because in Korean it is stylistically more natural not to explicitly mention anaphors in subordinate clauses that are coreferential with nominal expressions in the main clause. It is somewhat similar to English participle constructions whose subject is "understood." The best translation of (4b) in Korean could be described in English literally as:

(5) *Being usually written by good programmers, programs may still contain mistakes.*

Thus, if we are able to get the translation of (4a) and (4b) without overt pronominals, we are more likely to get better translation. This being so, anaphora resolution is very crucial in English-to-Korean MT because we must resolve the pronominal *they* to replace it by proper nominal expressions.

"Optional" anaphora resolution means preserving anaphoric ambiguity in case no anaphora resolution is undertaken. It may seem that carrying ambiguities over into translation is even more "authentic" from the point of view of having a mirror translation of the source text. Not resolving anaphoric ambiguity means that during the translation process text is not fully understood. Generally speaking, however, analysis is aimed at producing an unambiguous intermediate representation [Isabelle & Bourbeau 85].

Moreover, a system strongly relying on the "ambiguity preservation" method, in addition to offering no computational advantage when ambiguity-preserving situations must be identified dynamically, is extremely vulnerable in situations where (i) the lexicon is growing while the system is in use or (ii) when additional languages must be introduced [Nirenburg et al. 92]. Every new word sense added to the lexicon carries the potential of ruining the possibility of retaining ambiguity in translation for all previous entries. All this means that extra attention must be paid to the maintenance of the lexicons.

3 Previous Work on Anaphora Resolution in MT

Anaphora resolution is a complicated problem in natural language processing. Considerable research has been done by computational linguists (e.g. [Carbonell & Brown 88], [Hobbs 78], [Ingria & Stallard 89], [Rich & LuperFoy 88], [Sidner 86]), but no complete theory has emerged which offers a resolution procedure with guaranteed success. Most approaches developed - even if we restrict our attention to pronominal anaphora - from purely syntactic ones to highly semantic and pragmatic ones, only provide a partial treatment of the problem.

Though anaphora resolution has its own specific problems within the domain of MT, there has not been much work reported from this point of view.

Wada [Wada 90] reports on an implementation of Discourse Representation Theory in an LFG-based English-to-Japanese MT program. His anaphora resolution mechanism consists of three functional units: construction of discourse representation structure, storage of the salient element and search for the antecedent.

H. Saggion and A. Carvalho [Saggion & Carvalho 94] approach pronominal anaphora in a Portuguese-to-English MT system which translates scientific abstracts using syntactic agreement and c-command rules to solve intrasentential anaphora plus syntactic analysis of the immediately preceding sentence and a history list of previous referents to solve intersentential anaphora.

Preuß, Schmitz, Hauenschild and Umbach [Preuß et al. 94] describe work on anaphora resolution in an English-to-German MT system. Their approach uses two levels of text representation (structural and referential) and proximity, binding, themehood and conceptual consistency as factors.

Mitkov, Kim, Lee and Choi [Mitkov et al. 94] report on an extension of an English-to-Korean MT system to meet the needs of resolving pronominal anaphora.

4 The CAT2 Machine Translation System

CAT2 is a unification-based formalism designed for machine translation [Sharp 88, Sharp 91]. It was developed at IAI, Saarbrücken, as a sideline implementation to the Eurotra Project, and has been undergoing constant development and evolution since 1987. Experimental versions of numerous languages have been implemented, including English, German, Spanish, French. Portuguese, Italian, Dutch, Russian, Greek, Korean and Japanese. It is now being used in preindustrial projects for a number of commercial firms and academic institutions [Sharp & Streiter 95].

The translation strategy is based on tree-to-tree transduction, where an initial syntactico-semantic tree is parsed, then transduced to an abstract representation ("interface structure"), designed for simple transfer to a target language interface structure. This is then transduced to a syntactico-semantic tree in the target language, whose yield provides the actual translated text. The analysis of a source language, as well as the generation of a target language, is based on strictly monolingual rules, the transfer component being the only interface between two languages. Thus, an analysis in one language may be transferred to any number of target languages without requiring reanalysis. The various components

make use of common rules, much like subroutines, so that "universal" descriptions may be made applying equally to any number of languages, thereby significantly reducing the rule base, as well as simplifying the maintenance of grammars and the addition of new language components.

The formalism specifies two rule types for tree construction, and two rule types for tree transduction. Trees are built using "b-rules", which define a context-free backbone using at tribute-value pairs rather than simple category symbols. The following illustrates how the rule "S → NP VP" might be written in CAT2 notation:

```
{cat=s}.[ {cat=np}, {cat=vp} ].
```

The feature bundles may include any number of simple or complex feature descriptions; simple features have atomic values, e.g. **`cat=s`**, whereas complex features have feature bundles as values, e.g. **`agr={num=sing,per=3}.`** In addition, since the formalism is implemented in Prolog, a value may be a logical variable, bound to another variable with the same name within the rule; instantiation of one of the variables automatically instantiates the other. The implementation also allows for negative and disjunctive features, implemented in SICStus Prolog using the **`when/2`** construct for freezing goal evaluations.

The second rule type in tree construction is the "f-rule" for validating the feature content of partial trees. A simple f-rule for ensuring subject-verb agreement might be coded as follows:

```
{} .[ {cat=np}>>{agr=X}, {cat=vp}>>{agr=X} ].
```

This rule states that, in a tree configuration containing an NP as left daughter and a VP as right daughter, their agreement features must unify.

In practice, our grammars make use of a very small number of b-rules, based on X-bar syntax, (extended) head features [Streiter 94], and lexically-driven tree construction. The f-rules instantiate various universal and language-specific principles and properties, as well as supplying default values to lexical and phrasal constructions.

The tree transduction rules employ analogous rule types: t-rules transform tree structures, and tf-rules copy or transform selected features from source to target trees. The rule formats are similar to b- and f-rules, and again unification underlies the rule application. Since the rules for anaphora resolution in our model do not employ t- or tf-rules, they will not be further described here. See [Sharp 94] for a complete description of the formalism.

5 Our Anaphora Resolution Model

The current version of our anaphora resolution model implemented in CAT2 is exclusively based on syntactic and semantic constraints and preferences.

A. Syntactic Constraints

Various syntactic constraints for elimination of otherwise acceptable candidates for antecedents are used. The most obvious constraint is the agreement of the pronoun and its antecedent in number, person and gender. C-command constraints ([Ingria & Stallard 89]) are also used in the filtering process.

B. Syntactic preferences

- Syntactic parallelism preference: preference is given to antecedents with the same syntactic role as the pronoun.

The, programmer$_i$ successfully combined Prolog$_j$ with C, but he$_i$ had combined it$_j$ with Pascal last time.
The programmer$_i$ successfully combined Prolog with C$_j$, but he$_i$ had combined Pascal with it$_j$ last time.

- Topicalization preference: topicalized structures are searched first for possible anaphoric referents.
  It was Asia$_i$ who told Zahara$_j$ to go to Kuala Lumpur. Why did she$_i$ do it?

C. Semantic constraints and preferences

- Verb case role constraints: the case role semantics impose constraints on what can fill them. If filled by the anaphor, the verb case role constraints must also be satisfied by the referent of the anaphor.
  Vincent removed the diskette from the computer$_i$ and then disconnected it$_i$.
  Vincent removed the diskette$_i$ from the computer and then copied it$_i$.
- Semantic network constraints (semantic consistency): semantic networks indicate the possible links between concepts as well as concepts and their attributes.
  Do not run the program$_i$ on this computer because it$_i$ is encoded.
  Do not run the program on this computer$_i$ because it$_i$ is broken.
- Semantic parallelism preference: antecedents are favoured which have the same semantic role as the pronoun.
  Vincent gave the diskette to Sody$_i$. Kim also gave him$_i$ a letter.
  Vincent$_i$ gave the diskette to Sody. He$_i$ also gave Kim a letter.

6 Implementation in CAT2

We are interested primarily in intersentential anaphora resolution, but since CAT2, like most other MT systems, only operates on single sentences, we simulate the intersententiality by conjoining sentences, as in:

The. decision was adopted by the council; it published it.

Our task is to resolve the pronominal references in the second "sentence" with the antecedents in the first. It will be seen that our implementation, which successfully handles pronominal resolution in the context of English-to-German translation, can be carried over to multiple-sentence input.

The noun phrase features relevant for anaphora resolution are collected in the complex feature **anaph**, consisting of two additional features, **ref** (referential features) and `type`. The referential features include the noun's agreement features (person, number, gender), lexical semantic features (e.g. animacy), and its referential index; the type feature indicates whether the noun is a pronoun or not:

```
anaph={ref={agr=A,sem=S,index=I},type=T}
```

All non-pronominal nouns receive a unique index by a special index generator within CAT2; each pronoun takes its index value by way of unification with its antecedent, as outlined below.

Anaphora resolution in CAT2 occurs in two main steps. First, all **anaph** features within a sentence are collected in a `cand` (candidates) feature and percolated to the S node, so that anaphora resolution between sentences can take place locally under the topmost node.

(Intrasentential anaphora resolution will have already occurred.) Then, for each pronoun in the second sentence, its **ref** feature is resolved with those of the antecedents in the first sentence. Backtracking provides for all combinations, under the condition that the **ref** features agree, i.e. unify.

We assume that the verb is the head of the sentence, so that a verb's head feature will, by the standard notions of headedness and the percolation convention, be available at the S node. The anaphoric features of the verb's arguments are copied to the verb's head feature **cand**, which will therefore be available at the S node. This is accomplished by the following lexical f-rules:

```
f1 = {head={cat=v,cand={cand1=X}}, frame={arg1={head={anaph=X}}}}.[].
f2 = {head={cat=v,cand={cand2=X}}, frame={arg2={head={anaph=X}}}}.[].
f3 = {head={cat=v,cand={cand1=nil}}}.[].
f4 = {head={cat=v,cand={cand2=nil}}}.[].
```

The first rule, **f1**, copies the anaphoric features from the verb's first argument to the cand1 slot in the verb's cand feature; the rule **f2,** does the same for the verb's second argument. The rules **f3** and **f4** assign default values of nil in case the verb has no corresponding argument value, for example in the case of weather verbs and intransitive verbs. For simplicity, the above rules assume a 2-argument predicate structure, although this may be arbitrarily increased.

When a verb's argument is discharged during parsing, the anaphoric value of the argument is automatically bound within the verb's **cand** feature, so that by the time the sentence has been fully parsed, the **cand** feature contains the full complement of candidate anaphoric values occurring within the sentence.

The intersentential anaphora resolution process begins after two sentences have been combined under a single node, which has been constructed by the following b-rule (intricacies aside):

```
top = {}.[ {head={cat=v,vform=fin}},
           {head={cat=coord}},
           {head={cat=v,vform=fin}} ].
```

The anaphora resolution attempts to bind each pronominal reference in the second sentence with a candidate expression in the first sentence. The method of binding is effected by two f-rules, one for binding the first pronoun, if it exists, and one for binding the second, if it exists. The first f-rule is shown below:

```
anaph1 = {}.[ {head={cand={cand1={ref=R1},cand2={ref=R2}}}},
              {head={cat=coord}},
              {head={cand={cand1={type=pron}}}}
                  >>{head={cand={cand1={ref=(R1;R2)}}}} ].
```

In this rule, the first sentence's candidate referential features are bound to the logical variables R1 and R2. If the first candidate in the second sentence is a pronoun, then its referential features must unify with either R1 or R2. If unification fails, the sentence is deemed to be ill-formed. If unification succeeds for both R1 and R2, the disjunctive constraint is retained within the structure, pending resolution by subsequent resolution rules. A similar rule is provided for binding the second pronoun in the second sentence.

If the second sentence contains no pronouns, the f-rules do not apply. If it contains one pronoun, then it must unify with one of the candidates from the first sentence. It may

unify with either R1 or R2, providing ambiguous anaphoric readings. If the second sentence contains two pronouns, then both must be bound to the candidates in the first sentence. For example, the first pronoun's referential features may unify with R1, leaving R'2 to be bound to the second pronoun's features. If this unification fails, then backtracking will force the first pronoun to be bound to R2, leaving R1 to be bound to the second pronoun.

The rules above enforce strict binding constraints, based on morphological and semantic properties of the anaphora. Other rules based on semantic roles, not shown here, require the inclusion of the role values within the anaphoric candidate features, together with the appropriate f-rules for matching role values between the first and second sentences. Since these rules are preferential in nature, they follow the strict rules, and serve to resolve any remaining ambiguous anaphoric readings.

As an illustration, consider the following examples, translated correctly by our model:

- *The. council adopted the decisions; the commission published them.*
  Der Rat verabschiedete die Beschlüsse; die Kommission veröffentlichte sie.
- *The council adopted the law; it published it.*
  Der Rat verabschiedete das Gesetz; er veröffentlichte es.
- *The commission published the law; it was adopted by the council.*
  Die Kommission veröffentlichte das Gesetz; es wurde von dem Rat verabschiedet.
- *The decision was adopted by the council; it published it.*
  Der Beschluß wurde von dem Rat verabschiedet; er veröffentlichte ihn.

7 Further Improvement of the Model

As further improvement of the model, the integration of a center tracking engine is envisaged in CAT2. Center tracking is very important in anaphora resolution, because in case syntactic and semantic constraints are not sufficient to discriminate among a set of candidates, it is the centered noun phrase, which is considered as the referent.

The center tracking engine will be based on a statistical approach which we have developed to determine the probability of a noun phrase to be the center of a sentence ([Mitkov 94b]). Unlike the known approaches so far, our method is able to propose with high probability the center in every discourse sentence, including the first one. The approach uses an inference engine based on Bayes' formula which draws an inference in the light of some new piece of evidence.

In addition, we envisage the implementation of heuristic and domain modules, which together with a referential expression filter are expected to improve the performance of anaphora resolution in the CAT2 system.

REFERENCES

[Carbonell & Brown 88] J. Carbonell, R. Brown. Anaphora resolution: a multi-strategy approach. Proceedings of COLING'88, Budapest, 1988

[Hirst 81] G. Hirst. Anaphora in natural language understanding. Berlin Springer Verlag, 1981

[Hobbs 78] **J.** Hobbs. Resolving pronoun references. Lingua, Vol. 44, 1978

[Hutchins & Somers 92] J. Hutchins, H. Somers. An Introduction to Machine Translation, Academic Press, 1992

[Isabelle & Bourbeau 85] P. Isabelle, L. Bourbeau. TAUM-AVIATION: Its technical features and some experimental results, Computational Linguistics 11, 1985

[Ingria & Stallard 89] R. Ingria, D. Stallard. A computational mechanism for pronominal reference. Proceedings of the 27th Annual Meeting of the ACL, Vancouver, Canada, 1989

[Mitkov 94a] R. Mitkov. An integrated model for anaphora resolution. Proceedings of the 15th International Conference on Computational Linguistics COLING'94, Kyoto, Japan, 1994

[Mitkov 94b] R. Mitkov. A new approach for tracking center. Proceedings of the International Conference "New Methods in Language Processing", UMIST, Manchester, UK, 1994

[Mitkov et al. 94] R. Mitkov, H.K. Kim, H.K. Lee, K.S. Choi. The lexical transfer of pronominal anaphors in Machine Translation: the English-to-Korean case. Proceedings of the SEPLN'94 Conference, Cordoba, Spain, 1994

[Nirenburg et al. 92] S. Nirenburg, J. Carbonell, M. Tomita, K. Goodman. Machine Translation: a knowledge-based approach, Morgan Kaufmann Publishers, 1992

[Preuß et al. 94] S. Preuß, B. Schmitz, C. Hauenschild, C. Umbach. Anaphora Resolution in Machine Translation. In W. Ramm (ed): Studies in MT and NLP, Volume 6, EU Office, Luxembourg, 1994

[Rich & LuperFoy 88] E. Rich, S. LuperFoy. An architecture for anaphora resolution. Proceedings of the Second Conference on Applied Natural Language Processing, Austin, USA, 1988

[Saggion & Carvalho 94] H. Saggion, Ar. Carvalho. Anaphora resolution in a machine translation system. Proceedings of the International Conference "MT, 10 years on", Cranfield, UK, 1994

[Sharp 88] R. Sharp. CAT2: Implementing a Formalism for Multi-Lingual MT. Proceedings of the TMI'88 Conference, Pittsburgh, USA, 1988.

[Sharp 91] R. Sharp. CAT2: An Experimental Eurotra Alternative. Machine Translation 6, pp. 215-228.

[Sharp 94] R. Sharp. CAT2 Reference Manual, Version 3.6. IAI, Saarbrücken, 1994

[Sharp & Streiter 95] R. Sharp, O. Streiter. Applications in Multilingual Machine Translation. In: Proceedings of the Practical Applications of Prolog, Paris, 1995.

[Sidner 86] C. Sidner. Focusing in the comprehension of definite anaphora. In B. Grosz, K. Jones, B. Webber (Eds): Readings in Natural Language Processing, Morgan Kaufmann Publishers, 1986

[Streiter 94] O. Streiter. Komplexe Disjunktion und erweiterter Kopf: Ein Kontrollmechanismus für die MÜ. Proceedings of Konvens'94, 1994, Vienna

[Wada 90] H. Wada. Discourse processing in MT: problems in pronominal translation. Proceedings of COLING'90, Helsinki, 1990