Pronominal Anaphora Resolution in the KANTOO Multilingual Machine Translation System

Teruko Mitamura, Eric Nyberg
Enrique Torrejon, Dave Svoboda
Annelen Brunner, Kathryn Baker
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
teruko@cs.cmu.edu

Abstract

We present an approach to pronominal anaphora resolution using KANT Controlled Language and the KANTOO multilingual MT system. Our algorithm is based on a robust, syntax-based approach that applies a set of restrictions and preferences to select the correct antecedent. We report a success rate of 93.3% on a training corpus with 286 anaphors, and 88.8% on held-out data with 144 anaphors. Our approach translates anaphors to Spanish with 97.9% accuracy and to German with 94.4% accuracy on held-out data.

1 Introduction

Since encoding all the linguistic and domain knowledge for pronominal anaphora resolution is thought to be a difficult and time-consuming task, recent research in anaphora resolution seeks a simple and rapid approach which does not require extensive syntactic and semantic knowledge (Mitkov 1998; Mitkov 1999; Nasukawa 1994). We also follow the trend toward a simple and rapid approach, but our approach utilizes syntactic knowledge for resolution. We present an algorithm that effectively resolves anaphors while preserving the accuracy of the translated text. In contrast with statistical methods, our approach does not require a large bilingual aligned corpus for training.

Our approach draws from linguistic approaches explored in earlier work (Carbonell and Brown 1988; Lappin and McCord 1990; Lappin and Leass 1994; Ferrandez et al. 1998). These approaches require a full parse or partial parse of the input sentence. Our approach is a sequential, rule-based, domain-independent procedure, which has been implemented and integrated into the KANTOO Multilingual MT system. The success rate for anaphora resolution is 93.3% on our training corpus and 88.8% on held-out data. The accuracy rate for translation of pronouns is 97.9% for Spanish and 94.4% for German on held-out data.

Our algorithm utilizes the syntactic f-structure that results from a full parse using an analysis grammar. The algorithm applies a set of well-known heuristics (constraints and preferences) used in "knowledge-poor" systems. However, it differs from previous approaches in that it does not calculate weights for the heuristics in order to choose the right antecedent; rather it applies heuristics in a sequential manner until one candidate antecedent remains. Since our evaluation indicates performance comparable to that of score-based, knowledge-poor systems, it can be inferred that adding more linguistic knowledge reduces the need for scoring procedures to prune incorrect antecedents. If necessary, semantic knowledge can be used once syntactic rules have been exhausted.

In the next sections, we discuss the details of our resolution algorithm, and the results of an evaluation on technical texts which were translated to Spanish and German. We conclude with a discussion of some implications for current and future work.

2 KANTOO Multilingual MT System

The KANT System (Mitamura et al. 1991) is a knowledge-based, interlingual machine translation system, developed for multilingual translations of technical documents in various domains. Application domains include heavy equipment documentation, computer manuals, automotive documentation, and medical records written in controlled language (Mitamura and Nyberg 1995).

KANTOO is the reimplementation of the original KANT MT system, and also accepts Controlled English as input. The current input specification is referred to as KANT Controlled English (KCE). KCE places some restrictions on vocabulary and grammar. Although some of the sentences in this study were rewritten to conform to KCE, we did not edit pronominal anaphors or any other constituents relevant to the anaphor resolution process in our initial study. We later experimented with the introduction of a few explicit rules regarding the use of pronominal anaphors in controlled English, with the goal of improving resolution accuracy (see Section 5).

2.1 Anaphora Resolution in KANTOO Analyzer

The KANTOO Analyzer parses sentences using a lexicon and a unification grammar, and produces a set of LFG-style f-structures. Then, the Interpreter maps each f-structure to produce an interlingua representation, as illustrated in Figure 1.

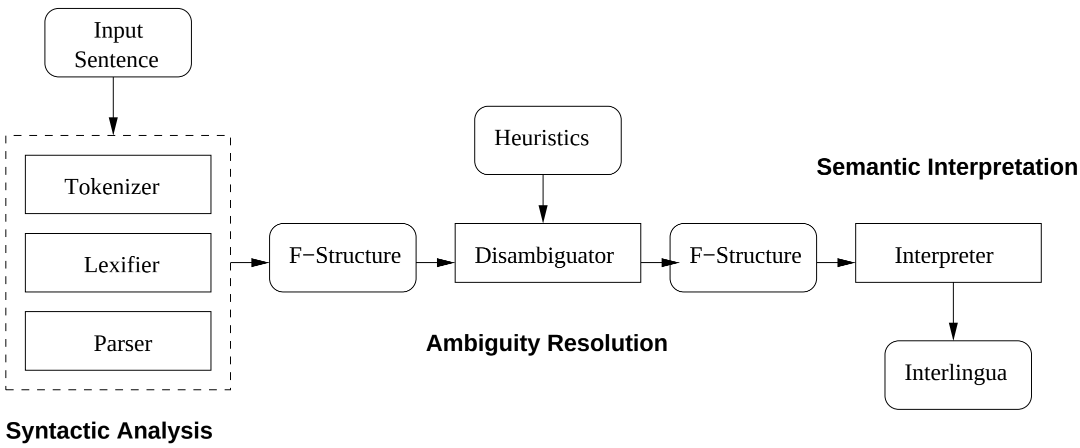

Figure 1: KANTOO Analyzer Architecture

The KANTOO Analyzer takes an assembly-line approach to sentence parsing. A sentence is first broken up into tokens, where a token can represent a word, number, SGML tag, or punctuation. In the second step, lexical meanings are assigned to tokens, including definitions for single words and multi-word phrases. Words and phrases may have multiple meanings. In the third step, tokens and their meanings are given as input to a non-deterministic parser, which generates every possible valid syntactic parse for the sentence. Lexical and syntactic ambiguities are preserved. The fourth step in analysis involves resolution of any ambiguities that are detected; anaphor resolution takes place during this final step.

During the ambiguity resolution process, the system creates several ambiguity tables, which provide direct access to the different parses which are determined by particular types of ambiguity. For example, there is a table that shows all of the valid constituents associated with a token. Most tokens will have only one constituent associated with them; if a token has two or more, then we have a lexical ambiguity. There is also a table that shows, for a particular constituent, all of the other constituents that it attaches to in the different parses found. Most constituents will have only one attachment site; if a constituent has two or more attachment sites, then we have a syntactic ambiguity.

The anaphor resolution algorithm is the last operation performed by the Analyzer. It traverses through the sentence looking for anaphora, and it runs the algorithm on each anaphor it finds. It starts by identifying all the possible antecedents that occur in the sentence before the anaphor, and then it employs the anaphor heuristics to prune out all but the 'correct' antecedent.

Each heuristic takes as input the ambiguity tables, the pronoun, and the set of candidate antecedents, and returns a (possibly modified) set of candidate antecedents. A heuristic may prune the set of antecedents according to the particular strategy it implements. If a heuristic is inapplicable, it may leave the antecedent set unchanged. Some heuristics might eliminate all the candidate antecedents; other heuristics may only eliminate some candidate antecedents.

If a heuristic eliminates every possible antecedent, the anaphor resolution algorithm then applies the heuristic on the same anaphor, but using candidate antecedents from the sentence that was previously processed by the Analyzer. If the previous sentence also does not contain a valid antecedent, the anaphor resolution algorithm then gives up, and does not assign an antecedent to the anaphor[1].

3 Anaphora Resolution Algorithm

In order to identify an antecedent for a given pronoun, we developed a two-stage approach. The first step is to identify possible antecedents by applying a set of pre-defined constraints. The second step is to eliminate candidates by applying a set of selection rules (heuristics) in a particular order.

3.1 Identification of Possible Antecedents

Possible antecedents for a given pronoun are identified according to a set of pre-defined constraints:

[1]We determined in an initial empirical investigation that the context for resolving anaphors in technical documentation rarely extends beyond the current sentence and the prior sentence.

1. The candidate antecedent must be a proper noun, noun, unit, SGML-tagged constituent, or conjoined NP.
2. If the antecedent is in the same sentence as the pronoun, it must precede the pronoun.
3. If the antecedent is a conjoined NP, it must conjoin NPs with *and* or *or*.
4. Prune any antecedent that is a part of a coordination.
5. The pronoun and candidate antecedent must agree in number (a coordination is implicitly considered plural).
6. If the pronoun is a verb argument, the antecedent may not be an argument of the same verb (note: we do not consider reflexives such as 'itself'.)
7. If the pronoun is the object of a prepositional phrase or relative clause modifying a noun, then that noun may not be a valid antecedent.

3.2 Antecedent Selection

After identifying the set of valid candidates, we apply the heuristics in the following order, to select the preferred candidate. After each heuristic is applied, if the set of valid candidates contains only a single antecedent, it is selected, otherwise the next heuristic is applied. If there is no candidate found in the same sentence, then the candidates from a previous sentence will be examined.

It is important to note that not every heuristic is tried for each anaphor, and sequential ordering is used to rank the heuristics. This is in direct contrast with approaches that try all heuristics on every anaphor, and use a weighted-sum scoring technique to make the final selection.

1. Prefer an antecedent that is also an anaphor.
2. Prefer an antecedent that is not a SGML-tagged constituent.
3. If two antecedents occur in this form: `<np1>` of `<np2>`, prefer `<np1>`. But if `<np1>` is one of "type/length/size/part", prefer `<np2>`.
4. Collocation: Prefer antecedents that attach to the same syntactic constituent as the pronoun.
5. Syntactic Parallelism: Prefer antecedents that attach to the same part of speech as the pronoun.
6. Syntactic Parallelism: Prefer antecedents that fill the same grammatical function as the pronoun.
7. Prefer antecedents that are conjunctions.
8. Definiteness-1: Prefer nominal antecedents that have a determiner, quantifier, or possessor, or are the value of a tag.

Pronouns	Correct	Total	Success Rate
IT	194	219	88.5%
THEY	24	24	100%
THEM	41	46	89.1%
Total	259	289	89.6%

Table 1: **Training Corpus Test Results (A)**

Location of Antecedent	Correct	Total	Success Rate
Intra-sentential	187	210	89%
Inter-sentential	72	79	91.1%

Table 2: **Training Corpus Test Results (B)**

9. Definiteness-2: Prefer nominal antecedents that have a definite determiner.

10. Closeness: Prefer the last (most recent) antecedent.

4 First Experiment

In order to tune the algorithm, we first selected a corpus from electronic product manuals. The corpus consists of 221 sentences containing 289 third person pronouns (*it, they*, and *them*) with inter- and intra-sentential antecedents. Roughly 27% of the pronouns are inter-sentential and 73% intra-sentential. The average number of candidate antecedents (noun, noun phrases, and pronouns) for each anaphor is 3. The algorithm resolved the anaphors in the training corpus with a success rate of 89.6%, as shown in Table 1. The distribution and success rate of intra- and inter-sentential is shown in Table 2.

Although the results from the training corpus were promising, we investigated the 30 problematic cases to determine why the algorithm failed. We partitioned these cases into the following categories. In some cases, a particular failure belongs to more than one category.

- **Lack of Domain Knowledge**: In some cases, domain knowledge would help to resolve the anaphoric reference. For example,

  "The contrast setting affects the lightness or darkness of an outgoing **fax** as **it** is being sent."

  KANTOO selects *contrast setting* instead of *fax* as an antecedent, because of rule 6, Syntactic Parallelism. If the system contains the knowledge that *fax* denotes an object that can be sent, but *contrast setting* and *lightness or darkness* do not, then *fax* will be selected. However, we note the existence of other cases where this type of knowledge is insufficient. For example,

  "If you have **another parallel port** on the computer, try plugging the cable into **it** instead."

KANTOO selects *computer*, instead of *another parallel port*, due to the syntactic parallelism rule. The system can't use simple semantics to make a distinction, because both a computer and a parallel port can receive plug connections. To select the correct antecedent, the system might have to prefer the most specific antecedent, in semantic terms, using partonomic information (e.g., *parallel port* is a part of *computer*).

We found that about 33% of the 30 problematic cases require this level of semantic treatment in order for the system to select the correct antecedent.

- **Ordering Problems**: Since antecedent selection is performed sequentially, there are some cases where the chosen ordering of the heuristics selects an incorrect antecedent, although a later heuristic might have made the correct choice. In the following sentence,

  "If you use the first procedure, then the **setting** remains on for all jobs until you change **it** again."

  KANTOO selects *procedure* as an antecedent, instead of *setting*, because the syntactic parallelism condition comes earlier than the closeness condition.

- **Lack of CL Rules for Anaphora Resolution**: Even though the heuristics for antecedent selection are syntactic-based, we did not employ any controlled language rules for the use of pronouns. In some cases, it is not practical to enforce restrictions on writing style, due to decreased author productivity and system utility. Nevertheless, the use of CL rules can improve the accuracy of anaphor resolution; possible constraints are discussed in the following section.

- **Miscellaneous Problems**: We found that there are a few miscellaneous problems which prevented the system from finding the correct antecedent, for example: a) ungrammatical writing on the part of the author, e.g. sentences like "place a semicolon in front of each line in order to remark them out", where plural *them* is used in place of the correct anaphor *it* when the antecedent is singular (e.g., *each line*); b) Proper nouns which weren't recognized as possible antecedents because they appeared inside SGML tags (e.g. `trademark`). The system was extended to address the latter case by explicitly searching for proper nouns inside SGML tags.

5 Controlled English for Anaphora Resolution

Although it is generally preferable to reduce the restrictions placed on the author, rules or restrictions that are easily learned can be followed without too much difficulty. We formulated two possible authoring restrictions which improve the effectiveness of anaphor resolution, and used then to rewrite some sentences from the test corpus.

- If you have more than one pronoun in a sentence, the pronouns must have the same antecedent. Otherwise, split the sentence into two sentences.

- If the antecedent is in the previous sentence, introduce the pronoun as soon as possible. For example, in the following sentences,

Pronouns	Correct	Total	Success Rate
IT	202	217	93.0%
THEY	24	24	100%
THEM	41	45	91.1%
Total	267	286	93.3%

Table 3: **Revised Results, Training Corpus**

"The laser printer 1234 product operates as a regular fax machine."

"You are not required to install the software or turn on the computer in order for *it* to work."

it refers to the noun *product* in the previous sentence. In this case, the author will be encouraged to rewrite the sentence as:

"In order for *it* to work, you are not required to install the software or turn on the computer."

so that *it* is introduced in the beginning of the sentence.

We found about 20% of the 30 problem cases in the training corpus could be addressed with these restrictions.

6 Evaluation of Antecedent Selection

After we fixed some miscellaneous problems and rewrote some sentences based on the CL rules, we ran the same corpus again. We did not add specific domain knowledge for antecedent selection. The results are shown in Table 3.

The total number of cases was reduced slightly because we found that in some cases, the use of pronouns in the original sentences was inappropriate, since there was no obvious antecedent. Such examples were rewritten without pronouns.

We need to point out that rewriting sentences according to controlled language guidelines dose not guarantee correct selection of an antecedent when there is more than one choice. The rules are intended to help authors make their sentences easier to understand, and sometimes improve the accuracy of anaphora resolution.

6.1 Results from Held-Out Data

Since the training corpus was used for determining the ordering of the heuristics for antecedent selection, we tested the algorithm on held-out data in a corpus containing 134 sentences with 144 pronouns. The distribution of intersentential and intrasentential pronouns was similar to distribution in the training corpus (25% and 75% respectively). The sentences were rewritten in KCE where necessary, taking care that this editing did not affect any pronoun or candidate antecedent in the evaluation. There is only one sentence which was rewritten in order to follow the guidelines created for anaphora resolution. The average number of candidate antecedents for each pronoun in this corpus was also 3, just like in the training corpus. The algorithm resolved the anaphors in this corpus with a success rate of 88.8%, as shown in the Table 4.

Pronouns	Correct	Total	Success Rate
IT	73	84	86.9%
THEY	22	24	91.6%
THEM	33	36	91.6%
Total	128	144	88.8%

Table 4: **Results on Held-Out Data**

7 Evaluation of Translation Results

We tested our Spanish and German translation systems on the held-out data to determine the translation accuracy rate when the anaphora resolution heuristics were used. In the following sections, we will discuss target language-specific issues concerning the results.

7.1 Spanish Translation Results

For the held-out data in the test corpus, the Spanish translation accuracy rate was 97.9%. There are 13 anaphors that are correctly translated, even though an incorrect antecedent was chosen by the resolution algorithm. This occurs when the selected antecedent happens to have the same gender in Spanish as the correct one, or when the anaphor is in subject position and gender is irrelevant when the translated pronoun is dropped (although the gender information must be correct if the verb phrase contains an attribute which must agree with the subject). For example, in the sentence below, the pronoun *they* is wrongly associated with *spots and smudges* as the antecedent.

Spots and smudges appear in the background areas of transparencies when **they** *are projected on the screen.*

The correct antecedent is *transparencies*. However, both antecedents are feminine, and the anaphor is correctly translated. Although it is dropped in the Spanish translation, the participle *projected* must agree in gender and number with the subject *they*. This is the case in the translation rendered by KANTOO: where *proyectadas*, feminine plural, agrees with the dropped subject of the verb *son*.

Las marcas y las manchas aparecen en las áreas de fondo de transparencias cuando son proyectadas en la pantalla.

Many commercial Spanish MT systems translate every third person pronoun as masculine, without trying to identify the correct antecedent. If we consider the nouns listed in the Spanish Thesaurus compiled by Julio Casares (1996), we see that almost 40% of them are feminine. The implication is that translating always to a default choice (masculine) will achieve a probable average success rate of 60% in similar cases.

The 60% (masculine) vs. 40% (feminine) gender distribution of the Spanish language is almost exactly the distribution found in the held-out data. We found that 40.15% of the correct antecedents in the corpus are feminine, and 59.85% of them are masculine. Since the success rate of KANTOO's algorithm is 97.9% when used with KANTOO Spanish MT, the postediting effort required to fix wrong translations of anaphors is significantly less.

7.2 German Translation Results

We also tested the resolution results for German translation of the same held-out data. In the German language there are three genders: masculine, feminine and neuter. The translation of third person plural pronouns into German is comparatively easy, as there is only one form, *sie*, for all genders. Because of this we have a 100% success rate in the translation of *them* and *they*. Third person singular pronouns, however, appear in three different forms. The translation of the word *it* in nominative position could be *er*, *sie* or *es*, depending on the gender of its antecedent. This distinction also exists if the pronoun is a direct object. Only in indirect object position the same form, *ihm*, is used in the case of masculine as well as neutral antecedents.

The percentage of all anaphors correctly translated in German is 94.4%. Compared to the percentage of anaphors correctly resolved by the Analyzer, we gain 5.6%. This is because German (like English) has only one form for plural pronouns, and because of coincidences similar to those in Spanish translation, where a wrongly chosen antecedent happens to have the same gender as the correct one.

If we utilized a default translation of third person pronouns as masculine, we would still get the correct translation for all *them* and *they* forms. However, if we utilized a default strategy for translation of *it*, performance would decrease significantly, as the distribution of gender in German is relatively balanced. In our test corpus, the pronoun *it* has to be translated as masculine in 37.6% of the cases, as feminine in 31.7% and as neuter in 30.5%. This would imply a success rate of 37.6% maximum using masculine as a default, compared to 90.5% correct translation of *it* using our algorithm. The total percentage of anaphors correctly translated with a masculine default would be 63.4% for the whole corpus, similar to Spanish, but we can achieve 94.4% with the KANTOO algorithm. Thus for German, as in Spanish the postediting effort is significantly reduced when the new approach is adopted.

8 Discussion

Although the translation accuracy of Spanish and German anaphors is very high, there is still a room for improvement. The weakness of our approach is that since antecedent selection is performed sequentially, we find some cases where the chosen ordering selects an incorrect antecedent. Since the advantage of this approach is simple and fast development of anaphora resolution using syntactic knowledge, we have not yet encoded the domain knowledge, which will be a time-consuming task.

Encoding the world knowledge and deep inference mechanisms required for selecting the right antecedent is a bottleneck in reaching 100% coverage in unrestricted texts for both "knowledge-poor" approaches (Dagan and Itai 1990; Kennedy and Boguraev 1996; Mitkov 1998; Nasukawa 1994; Mitkov 1999) and linguistic approaches. Nasukawa pointed out the difficulty of achieving a success rate of over 90% without this type of knowledge.

Since the KANTOO MT system already uses domain knowledge for disambiguation (Mitamura and Nyberg 1995), it could be extended to include the domain knowledge for anaphora resolution. However, our current translation results into Spanish and German show that only a few percent of the anaphoric references require post-editing, even though KANT Controlled English has been extended to allow pronouns without using domain knowledge.

References

Carbonell, J. and R. Brown (1988). *Anaphora Resolution: A Multi-Strategy Approach.* In Proceedings of the 12th International Conference on Computational Linguistics (COLING'88). Budapest, Hungary.

Dagan, Ido and A. Itai (1990). *Automatic Processing of Large Corpora for the Resolution of Anaphora References.* In Proceedings of the 13th International Conference on Computational Linguistics (COLING'90). Vol. 3, 330-332, Helsinki, Finland.

Ferrández, A., M. Palomar, and L. Moreno (1998). *Anaphora Resolution in Unrestricted Texts with Partial Parsing.* In Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL'98), Montreal, Canada.

Diccionario Ideológico de la Lengua Española de Julio Casares (1996) Editorial Gustavo Gili, Barcelona.

Kennedy, C. and B. Boguraev (1996). *Anaphora for Everyone : Pronominal Anaphora Resolution without a Parser.* In Proceedings of the 16th International Conference on Computational Linguistics (COLING'96). 113-118. Copenhagen, Denmark.

Lappin, S. and H.J. Leass (1994). *An Algorithm for Pronominal Anaphora Resolution.* Computational Linguistics, 20(4), 535-561.

Lappin, S. and M. McCord (1990). *Anaphora Resolution in Slot Grammar.* Computational Linguistics, 16 :4, 197-212.

Nasukawa, T. (1994). *Robust Method of Pronoun Resolution Using Full-Text Information.* In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), 1157-1163, Kyoto, Japan.

Mitkov, R. (1998). *Robust Pronoun Resolution with Limited Knowledge.* In Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL'98), Montreal, Canada.

Mitkov, R. (1999). *Multilingual Anaphora Resolution*, in Mitkov, R. (ed.), Special Issue on Anaphora Resolution in Machine Translation and Multilingual NLP, *Computational Linguistics*, 14:3-4, December.

Mitamura, T., E. Nyberg, and J. Carbonell (1991). *An Efficient Interlingua Translation System for Multi-lingual Document Production.* In Proceedings of the Third Machine Translation Summit, Washington, D.C.

Mitamura, T. and E. Nyberg (1995). *Controlled English for Knowledge-Based MT : Experience with the KANT system.* In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation. Leuven, Belgium.