[TMI 99: *Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, August 1999, Chester, UK]

Multiple Strategies for Automatic Disambiguation in Technical Translation

Teruko Mitamura, Eric Nyberg,
Enrique Torrejon and Robert Igo
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh PA 15213 USA
teruko@cs.emu.edu

Abstract

The use of knowledge-based machine translation with controlled technical text can produce high-quality translations. However, building and maintaining knowledge bases can require significant time and effort, since they typically involve hand-coding of semantic preferences. When a system can't disambiguate based on semantic preferences, it can initiate interactive disambiguation with the author to improve the likelihood of an accurate translation, but this decreases the productivity of text authoring. In this paper, we present an experimental evaluation of automatic disambiguation strategies which could eliminate the need for interactive structural disambiguation in the KANT machine translation system.

1 Introduction

Research and development has shown that knowledge-based machine translation, combined with the use of controlled language in well-defined technical domains, can achieve very high accuracy in translation (Nyberg & Mitamura 1992; Mitamura & Nyberg 1995; Kamprath et al. 1998). Detailed knowledge bases often include semantic preferences for disambiguating structural attachments (Baker et al. 1994). However, the efficacy of knowledge-based MT has often been questioned because of the significant time and effort required to build semantic knowledge bases (Hutchins & Somers 1992). The goal of this paper is to address this issue and demonstrate a method which reduces the time and effort to build high-quality KBMT systems.

A semantic model developed for a particular domain may not cover all of the structural attachments in sentences which the system will eventually encounter. Therefore, a system which relies only on a semantic model for accurate attachment will require constant update. Furthermore, it is often necessary to process new documents for new product lines not covered by the existing domain model, resulting in an ongoing need to update the domain model over time.

The KANT machine translation system (Mitamura et al. 1991) queries the author to disambiguate interactively if the domain model cannot disambiguate a structural attachment automatically. This solution is not always satisfactory - interactive disambiguation is not always accurate, and it is always a time-consuming task, and hence costly in terms of overall system productivity.

In this paper, we present the results of an experiment which combines domain-independent heuristics with a semantic knowledge base. We explore a multiple-strategy approach which preserves a high degree of translation quality, while reducing both the need for interactive disambiguation and the effort required to build and maintain a semantic domain model.

In Section 2, we describe in more detail the goals of the research. In Section 3, we explain how ambiguity is handled in the KANT system. In Section 4, we describe the experiment, which compared the accuracy of two translations of a sample corpus from English to Spanish: one using interactive disambiguation by the author, and one using automatic attachment heuristics. In Section 5, we present and discuss the results of the experiment, and in Section 6 we conclude with some remarks about the implications of our results and proposed future work.

2 Improving Automatic Disambiguation

There are several reasons why it is important to consider new methods for automatic structural disambiguation in KANT:

- **Ambiguity is pervasive.** In the corpus chosen for our experiment, a total of 11,607 PP attachments occurred in 12,000 sentences – an average of about 1 PP per sentence.
- **Unresolved ambiguity leads to higher translation costs.** Sentences which are not properly disambiguated are likely to be translated incorrectly, leading to a corresponding increase in the amount of postediting required.
- **Interactive disambiguation leads to higher authoring costs.** Ambiguity which is not resolved by the system can be resolved interactively with the author, thus improving the quality of the input text. In the chosen corpus, 29% of the PP attachments were not disambiguated automatically, and required author intervention, leading to a significant pre-editing task.
- **Authors don't always make the right choice during interactive disambiguation.** Since authors are often working under deadline pressure and don't always understand fine linguistic distinctions, they sometimes choose the wrong f-structure during interactive disambiguation. Hence a quality translation isn't guaranteed, even if the time is taken to disambiguate each input sentence interactively.

The goal of our experiment was to decrease interactive disambiguation to improve author productivity, while maintaining high-quality translation to minimize a potential increase in postediting. In the KANT system, this meant increasing the level of automatic disambiguation without relying on (expensive) hand-coding of additional semantic preferences in the domain model.

- **Bend** the *locks* **away from** bolts (7) .
- Buckets are **grouped** into two different *families* **by** capacity.
- *Check* the **connections between** the fuel tank and the fuel transfer pump.
- **Check** the *linkage* **for** smooth movement.
- Do not **expose** the *machine* **to** flames, burning brush, etc.
- This *is* an *indication* of the *need* for **repair to** the solenoid.

Figure 1: Example PP Attachment Ambiguities

3 Ambiguity Resolution in KANT

The experiment was conducted using the KANT machine translation system (English to Spanish) and a representative set of sentences drawn from technical texts in the domain of heavy equipment manuals. In this section, we provide some particulars regarding structural ambiguity in the domain, and discuss how KANT typically handles structural ambiguity.

3.1 Structural Ambiguity in Technical Text

The style of technical writing in our experimental domain is typical of instruction manuals in general: explanatory text (descriptive/declarative sentences) mixed with lists of procedural steps (commands/imperative sentences). There are two main sources of ambiguity in the domain: lexical ambiguity (words with more than one meaning for a given part of speech) and structural ambiguity (syntactic constituents which could conceivably modify (or "attach to") more than one word or phrase in the sentence). For the purposes of this experiment, we focused on structural ambiguity, specifically, the attachment of prepositional phrase modifiers[1].

Figure 1 contains some examples of ambiguous PP attachments found in the domain. The correct attachment site and the preposition are underlined; other potential attachment sites appear in italics. It should be clear from these examples that even simple sentences from this domain require careful attachment of PPs, since making the wrong choice of attachment site would most likely result in an unacceptable translation.

3.2 Disambiguation in KANT

A full description of the KANT software architecture is beyond the scope of this paper; the interested reader may refer to (Mitamura et al. 1991) for more detail. What

[1] For a full discussion of the types of ambiguity in technical text and how they are handled by the KANT system, see (Mitamura & Nyberg 1995).

follows is a more focused description of the mechanism used in KANT for resolution of structural (attachment) ambiguity.

During interactive grammar checking, KANT takes the following steps to analyze each sentence in the document:

1. Morphological analysis is performed, and the set of possible lexical entries for each input token is retrieved;
2. A unification grammar is used to produce the legal set of grammatical functional structures *(f-structures)* for the input tokens;
3. If there is more than one possible structure, the system uses a set of automatic disambiguation heuristics to prune less preferred readings of the input;
4. If there is more than one possible structure remaining after automatic disambiguation, then the author of the text is engaged in an interactive disambiguation dialog.

The most important method used to disambiguate automatically is the use of a semantic domain model. In KANT, the domain model encodes semantic attachment preferences in the form of *triples,* which are essentially (`<head> <semantic-role> <filler>`) tuples for preferred attachments. For example, the following triple encodes the notion that hoists are commonly used as the instrument in a lifting action:

- `(*A-LIFT INSTRUMENT *O-HOIST)`

  Lift the *engine* from the *chassis* **with** a hoist.

To prune less preferred f-structures, KANT uses the following algorithm:

1. Each PP attachment in a f-structure is checked against the triples in the domain model and assigned a score. Attachments which match a triple exactly receive a score of 0; attachments which match a triple under `IS-A` inheritance on the head or filler[2] receive a score of 1; and attachments which match a triple under inheritance on both head and filler receive a score of 2.
2. The attachment scores for the entire f-structure are summed.
3. The entire set off-structures is ranked in order of ascending aggregate score. All f-structures which receive scores (penalties) higher than the lowest score are pruned. The set of f-structures (equivalence class) with the lowest score is retained.
4. Hence, the f-structures which most closely match the specific domain knowledge encoded in the semantic model are preferred.

[2] Verbs and nouns in the lexicon are associated with semantic concepts in the domain model; e.g., ["lift",V] → `*A-LIFT`. An `IS-A` hierarchy is used to arrange the concepts into classes corresponding roughly to verb classes and object classes, e.g. `*A-REPAIR-ACTION`, `*O-LIFTING-TOOL-OR-ASSEMBLY`, etc.

Even after automatic disambiguation, there are many sentences which are truly ambiguous in the domain (the semantic model can't discriminate a single best f-structure). Other sentences cannot be disambiguated because there is no relevant semantic knowledge in the domain model. In these cases, the author is presented with a set of alternative f-structures with the attachment site and preposition highlighted. When a particular interpretation is chosen by the author, an SGML processing instruction is inserted into the source text, e.g.:

- `Do not expose the machine to<?CTE attach head='expose' head-pos='3' modi='to'> flames, burning brush, etc.`
- `This is an indication of the need for repair to<?CTE attach head='repair' head-pos='9' modi='to'> the solenoid.`

When the text is eventually translated, the information stored in the processing instruction is used to automatically select the desired prepositional attachment. The current production version of the KANT system frequently relies on interactive disambiguation, because it is costly to encode an exhaustive set of semantic attachment preferences for a domain of significant size.

4 The Experiment

The test corpus contains 12,000 sentences with 11,607 instances of structural disambiguation (prepositional phrase attachment). In actual production use, the KANT system disambiguated 8209 of these PPs by using a semantic domain model to select a particular attachment automatically (see Section 3.2). This corresponds to 9254 sentences (77% of the total) which were covered through automatic disambiguation by the domain model. On the other hand, 2748 sentences (23% of the total) were disambiguated interactively by the authors (see Table 1).

	Automatic (DM)	Interactive (by Hand)	Total
PP Attachments	8209 (71%)	3398 (29%)	11,607 (100%)
Full Sentences	9254 (77%)	748 (23%)	12,000 (100%)

Table 1: Ambiguity in the Test Corpus

In this experiment, we focused on just the 2748 sentences which were interactively disambiguated. Our task was to compare the performance of the system (translation acceptability) using interactive disambiguation with a second scenario where new attachment heuristics were used in the place of interactive disambiguation. In both cases, the Spanish output sentences were evaluated, allowing us to determine whether any variations in the output in the second scenario were due to improvements or regressions in translation quality.

4.1 Automatic Attachment Heuristics

We used the following automatic attachment heuristics:

- USE + something + FOR
  Do not **use** a *chain* **for** pulling.
  Use *shims* (27) **for** the steering clutches.
- INSTALL + something + OVER
  Install the *plastic cap* **over** the bolt.
  Install the *assembly* for the *access cover* **over** the rear of the drive.
- PROVIDE + something + FOR
  Air compressor (1) **provides** *pressure air* **for** the brake circuit.
  This force **provides** the *power* **for** the brake application.
- MOVE + something + AWAY-FROM
  Move the *axle* **away from** the machine.
  Piston (9) **moves** *valve* (13) **away from** the seat in valve body (14) .

Figure 2: Example VERB+Prep Attachment Patterns

1. **Disambiguation using Domain Semantics.** If the system can disambiguate an attachment using knowledge from the domain model, it attaches the PP accordingly (this effectively limits the scope of the experiment to the 2748 sentences in the corpus which were ambiguous and not disambiguated by the domain model).

2. **Syntactic VERB+Prep Attachment Patterns.** In this domain, a large proportion of the PP attachments to the main verb can be described by a small number of specific VERB+Prep patterns. If a PP can attach to the direct object of the verb or to the main verb syntactically, and the sentence matches one of these patterns, then the system chooses the f-structure where the PP is attached to the verb. Some common examples of VERB+Prep patterns are shown in Figure 2.

3. **Other Syntactic Attachment Patterns.** Other common patterns in the domain include the use of the -ing form following the preposition by, which almost always attaches to the main verb. Examples are shown in Figure 3.

4. **Default (Local) Attachment.** If none of the heuristic patterns match, then the system resolves PP attachment ambiguity by selecting the most local attachment site, i.e., the NP or Verb immediately to the left of the PP. Examples are shown in Figure 4.

5 Results and Discussion

First we examined the Spanish translation output for the 2748 sentences in the test corpus (the sentences in the original corpus where interactive disambiguation was used). Of these sentences, 2442 were translated correctly (89% of the total sentences used;

- VERB+something+BY+VERB-ing
  Release the *pressure* **by** loosening the filler cap.
  Measure the *pin* **by** prying between the bogies and the roller frame.
- VERB+something+FOR+VERB-ing
  Use a scraper to **rip** the *surface* **for** loading.
  We **recommend** the *C5 tool* **for** rough shaping.

Figure 3: Examples of Other Attachment Patterns

- (29) Shims for *adjusting* the **residual pressure** of the brake.
- Timing pin (1) *fits* into the **hole** **in** camshaft (4).
- *Raise* the *air pressure* in the **line** **to** the cylinders.
- The check valve *allows* **steering** **with** a dead engine.

Figure 4: Examples of Local Attachment

referred to below as Group A). On the other hand, 306 of the Spanish translations required some postediting (11% of the corpus; referred to below as Group B).

We then translated the same 2748 sentences, without interactive disambiguation and with the automatic attachment heuristics mentioned in Section 4.1. First we applied only the Local Attachment heuristic, and found that there were some regressions (sentences which were no longer translated properly). Among the 2442 sentences correctly translated following interactive disambiguation (Group A), 170 sentences received incorrect attachments when Local Attachment was applied. On the other hand, of the 306 sentences that were incorrectly translated following interactive disambiguation (Group B), 18 sentences received acceptable translations after Local Attachment was applied (39 sentences still exhibited incorrect attachments). Otherwise, the evaluation of translations in Group A and B did not change.

Therefore, we found 152 sentences (170 minus 18) out of 2748 sentences whose translations suffered due to the Local Attachment heuristic, resulting in a total of 458 sentences which required postediting. 2290 sentences were still translated correctly (about 83% of the test corpus).

We then introduced the Pattern Heuristics described above. The most common cases where the Local Attachment heuristic failed were instances of `VERB+Prep` patterns. We found that about 52% of the sentences from the set of 209 incorrect translations (170 plus 39) were matched by the `VERB+Prep` patterns. After applying all of the Pattern Heuristics described in Section 4.1, the system was able to attach an additional 159 cases correctly, and only 299 sentences were left which required some postediting. We

used a total of 83 different patterns for the experiment.

The results indicate that there was no significant difference in the quality of the Spanish translation output when automatic attachment heuristics were applied. On the other hand, a significant productivity increase was made on the authoring side, by eliminating interactive disambiguation entirely. A summary of the results is shown in Table 2.

Method Used	Correct Trans.	Incorrect Trans.	Total
Interactive Disambiguation	2442 (89%)	306 (11%)	2748 (100%)
Min. Attachment Only	2290 (83%)	458 (17%)	2748 (100%)
Min. Attachment + Patterns	2449 (89%)	299 (11%)	2748 (100%)

Table 2: Experimental Results: Summary

Our results are achieved by using a combination of semantic domain model, lexical attachment patterns, and a general heuristic (local attachment). This is in contrast with previous work. For example, (Whittemore et al. 1990) focused on a set of limited semantic and structural heuristics for disambiguation, while other approaches (notably, (Hindle & Rooth 1993)) have focused on deriving lexical attachment patterns from corpora.

Another distinguishing characteristic of the current approach is that it uses feedback from translation evaluation to determine the effectiveness of the attachment heuristics. The focus is not on handling all prepositional attachments equally well; rather, the focus is on handling the most frequent attachments accurately, to reduce post-editing cost.

Because this approach relies on the hand-coding of patterns rather than automatic extraction of preferences from corpora, it is important to consider the relationship between the cost to develop disambiguation heuristics, and the cost savings from reduced authoring time and improved translation quality. In the future, we hope to conduct a user study which relates the reduction in interactive disambiguation to the actual time saved during the authoring process.

6 Conclusion

It is clear from our results that combining the use of a semantic domain model with additional heuristics for automatic disambiguation (syntactic disambiguation patterns, local attachment) reduces the need for interactive disambiguation, with minimal impact on translation quality for the initial corpus of 12,000 sentences.

This research also suggests that the multiple strategies of semantic and syntactic disambiguation can work successfully when the system can use semantics to disambiguate about 3/4 of the cases. This study gives us an insight into the proportion of semantic disambiguation vs. syntactic disambiguation knowledge that might be required to build an application in a new domain. It is likely, however, that these results will not generalize directly from domain to domain or from language pair to language pair. Domains which are less narrow in focus will not exhibit such tight patterning

of prepositional usage, and syntactic `VERB+Prep` patterns might be less effective. It is also possible that a target language which diverges more greatly from English in its syntactic structure would be less forgiving about incorrect local attachments. Since English and Spanish have similar syntactic structures in many respects, some incorrect local attachments may not affect translation outputs because they are ambiguous with a non-local attachment (the correct attachment) in Spanish. We have also investigated attachments in English-Italian translation, and found that the same phenomenon applies[3].

For future applications of KANT, our experimental results suggest an alternative approach for developing domain coverage. Rather than creating a comprehensive semantic knowledge base initially, one might follow these steps instead:

- Translate a test corpus using only the local attachment heuristic for PP attachment disambiguation.
- Score the resulting translations, and identify sentences where an incorrect attachment led to a poor translation.
- Where possible, derive syntactic patterns like those presented in Section 4.1. This includes both general patterns and verb-specific patterns which appear frequently in the corpus and require non-local attachment (e.g., to the main verb).
- For the remaining high-frequency cases which cannot be described by general or verb-specific syntactic patterns, encode specific semantic knowledge (e.g., KANT triples) for the preferred attachment.
- Any remaining exceptional or low-frequency cases which cannot be disambiguated by the above methods are resolved automatically by the local attachment heuristic.

Based on our experimental results, we feel that this approach would yield a significant reduction in time and effort to develop new KANT applications, while maintaining a high degree of translation accuracy and eliminating the interactive disambiguation task for the author.

7 Acknowledgments

The authors thank their industrial sponsors for providing the test data used in the experiments. We also thank Michael Duggan for his hard work in setting up and running the translation experiments that were used to gather the data used in this paper. The authors would also like to acknowledge Joseph Giampapa, who collaborated with us on an earlier investigation of local attachment differences in translation.

[3] Joseph Giampapa, unpublished term paper.

References

Baker, Kathy, Alex Franz, Pam Jordan, Teruko Mitamura & Eric Nyberg: 1994, 'Coping with Ambiguity in a Large-Scale Machine Translation System', in *Proceedings of COLING-94,* Kyoto.

Hindle, Donald and Mats Rooth: 1993, 'Structural Ambiguity and Lexical Relations', *Computational Linguistics,* 19:1, pp. 103-120.

Hutchins, W. John and Harold L. Somers: 1992, *An Introduction to Machine Translation,* San Diego: Academic Press.

Kamprath, Christine, Eric Adolphson, Teruko Mitamura & Eric Nyberg: 1998, 'Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English', in *Proceedings of the Second International Workshop on Controlled Language Applications: CLAW-98,* Pittsburgh.

Mitamura, Teruko, Eric Nyberg & Jaime Carbonell: 1991, 'An Efficient Interlingua Translation System for Multi-lingual Document Production', in *Proceedings of the Third Machine Translation Summit,* Washington, D.C.

Mitamura, Teruko & Eric Nyberg: 1995, 'Controlled English for Knowledge-Based MT: Experience with the KANT System', in *Proceedings of TMI-95,* Leuven.

Nyberg, Eric and Teruko Mitamura: 1992, 'The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains', in *Proceedings of COLING-92,* Nantes.

Nyberg, Eric, Teruko Mitamura & Jaime Carbonell: 1994, 'Evaluation Metrics for Knowledge-Based Machine Translation', in *Proceedings of COLING-94,* Kyoto.

Whittemore, Greg; Ferrara, Kathleen; and Brunner, Hans: 1990, 'Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases', in *Proceedings of ACL-90,* Pittsburgh.