How Much Semantics is Necessary for MT Systems?

Winfield S. Bennett
Siemens Communications Systems, Inc.
and
The Linguistics Research Center
University of Texas at Austin

ABSTRACT

Semantics has been an issue in machine translation since the beginning of research in the field. Inherent in the issue are questions of the centrality of semantics to MT systems and of the form and extent for semantic components. This paper examines these questions in light of experiences in research, proposing that the standard views may need reassessment. Additionally I discuss the use of semantics features in dealing with several issues.

1. Introduction

Semantics has been an issue in machine translation research from the very beginning. Warren Weaver (1949) recognized the necessity of dealing with semantics in his famous memorandum, although his views of just how systems could deal with semantic problems appear rather naive to us today. Bar-Hillel (1960) and Yngve (1964) both clearly stated their belief in the need for MT systems to handle semantics. There has been little, if any, change in the views of the importance of the treatment of semantics in machine translation in recent years. It seems to me that the current views are summarized by Lytinen (1987, 302): 'It has long been realized by MT researchers that semantics must be used to resolve many of the lexical and structural ambiguities that occur in natural language.'

If one assumes that semantics is necessary for any machine translation system, one must then decide just what form that semantics must take. As Lehrberger and Bourbeau (1988, 103) put it: 'The need for semantic analysis is generally recognized; the big question is how to do it.'

My purpose in this paper is to examine the question of the necessity of deep semantic analysis in an MT system and to explore the form of any semantics. This paper is a continuation of Bennett (1989), in which I first suggested that perhaps our conventional wisdom in this matter should be examined further.

2. Is Deep Semantic Analysis Necessary for Machine Translation?

As I stated in Bennett (1989): 'It is probably uncontroversial to state that semantics would generally enhance any machine translation system.' However, the issue is often not viewed as a matter of system enhancement, but of the centrality of semantics in such systems. From my viewpoint the basic assumption is that one must have some sort of full-blown semantic analysis to have a viable MT system. I readily concede that some semantic analysis will, in principle, improve the quality of any system, but, at the same time, I question the notion that deep semantic analysis is essential to machine translation.

The assumption that a deep semantic component is essential to viable machine translation stems from the need for systems to handle all the idiosyncrasies of a given source language, including the very real problem of ambiguity. Ambiguity certainly tops the list of problems for MT systems. There are any number of well-known examples; I will confine myself to three:

(1) The box was in the pen. (Bar-Hillel 1960, 158)
(2) While driving down route 72, John swerved and hit a tree. (Carbonell 1987)
(3) The cleaners dry-cleaned the coat that Mary found at the rummage sale for $10. (Lytinen 1987, 303)

Two points can be made about these and similar examples with respect to machine translation. First, based on my experience, the sort of ambiguity represented by these examples is not really typical for the kinds of texts to which machine translation is commonly applied. This is not to say

that such examples should not be considered in MT research and system design, but their importance must not be overstated.

Second, the three examples represent instances of ambiguity which are, in some sense, impenetrable. In example (1) the underlying assumption is that *pen* can only refer to an enclosure large enough to accommodate a four-sided container, an assumption which does not take into account that *pen* can refer to a penal institution or that *box* can refer to tiny objects embedded in the barrels of writing implements. Examples (2) and (3), on the other hand, may be disambiguated only by reliance on knowledge which is more in the realm of pragmatics than semantics.

The issue of semantics for the resolution of ambiguity in machine translation texts should focus on the major examples found in the sorts of texts to which machine translation is routinely applied. Based on my observations, the most common types of ambiguity encountered by MT systems involves the prepositional phrase attachment of the following sort:

(4) The computer outputs the data in the file.
(5) The system uses the information to print the rule lines in the footer.
(6) Write the name on the piece of paper.

Examples (4)-(6) are not easily disambiguated by any sort of semantics, since attaching the PP to the NP is possible in any of them. In fact resolution of the ambiguity in any of these is a matter of pragmatics (as in (2) and (3), above), unless one wants to resorts to heuristics for choosing one representation over another.

None of my points is meant to dismiss the notion of a semantic component in machine translation systems. On the contrary, I believe we must pursue any course of research which can lead to semantic invariance, defined as: 'preserving invariant the meaning of the source text as it is transformed into the target text' (Carbonell and Tomita 1987, 68). I take issue, however, with the notion that a deep semantic analysis is an essential component in any viable MT system, given the fact that the bulk of source text can be translated without relying on such a semantic compo-

nent. Clearly some sort of a semantic analysis is a desirable part of any MT system.

3. How Much Semantics?

If one wants a semantic component in an MT system, the issue then is one of 'how to do it', as Lehrberger and Bourbeau (1988, 103) put it.

Obviously, specific answers to this question lie in the architecture of the systems themselves. I do not intend to prescribe just what should be in any particular system. Nor do I intend to avoid the question by facilely stating that there should be just enough not to cost too much but to handle the problems.

Approaches to semantics in machine translation systems can be quite varied, as one can easily see in reading Hutchins (1986 and 1988). While realizing the dangers of overgeneralization, I believe that approaches to MT semantics may be characterized as points on a continuum ranging from simple semantic feature systems to full-blown semantic analyses. Views on just what approach is necessary vary from researcher to researcher. It is noncontroversial to assume that the more powerful the semantic component the more thorough the analysis, all things being equal. However, it is likewise noncontroversial to assume that any component of a system has its cost. Such cost may be counted in any number of ways; in looking at the issue of semantics I will consider: efficiency, lexicon or knowledge base acquisition, and cost effectiveness.

By efficiency I mean the relationship of speed to effectiveness. I assume that any semantic component will reduce the speed of a system by its presence, just as any other component would. I would further assume that the more sophisticated the semantic component the greater the loss of speed. The trick, then, is to have a semantic component which is maximally effective while producing the least amount of 'drag' on the system.

The issue of lexicon or knowledge base acquisition and maintenance is one which is generally ignored in theoretical considerations for machine

translation systems. I think this is an error, since the time it takes to build a lexicon or knowledge base is a real factor in the cost of a system. My assumption is that a more elaborate semantic analysis will require more time for the acquisition and maintenance of the lexicon or knowledge base than a more modest one, because such a system would need more detailed semantic information to be effective.

Cost effectiveness I take to be the relationship of expense to the capability of the system to perform as required. This is certainly an issue for production systems, but may be something which should be considered in experimental ones. In some ways cost effectiveness involves the previous two points, since efficiency and the time to acquire and maintain the lexicon or knowledge base must be factors in the assessment of the real cost of a system.

In considering cost in relation to semantic components I will look at two extremes of the continuum I alluded to earlier in this section: a semantic feature system and a deep semantic analysis. By semantic feature system I mean a system in which semantic information is coded on lexical entries and handled in analysis in the same way as morphological or syntactic information. In such a system semantic information is simply manipulated as features on any given node in the tree. In referring to deep semantic analysis I envision a system in which the semantic component is a distinct and fundamental part of the system. While the full-blown semantic analysis I envision is somewhat hypothetical, the feature system is not; its operation in relation to PP attachment is described in some detail in Meya (1990).

I assume that generally the computational cost of a feature system is considerably less than that of a full-blown semantic analysis, since such systems treat semantic information in the same way as syntactic or morphological information, i.e., as features to be accessed in analysis. A semantic analysis system requires additional computation for the semantic component. The issue, then, is efficiency: is the additional computational cost for a full-blown system justified by the results? I claim that in the case of PP attachment, at least, it does not, since even the cleverest of se-

mantic analyses can only resolve a portion of these ambiguities. This is an instance where the investment of computation would gain little.

From the descriptions in the literature it is apparent that any full-blown semantic analysis requires considerably more investment in time in lexicon or knowledge base acquisition than a simple feature set. I admire the efforts at knowledge base acquisition at Carnegie Mellon (see, e.g., Nirenburg, et al. 1988) but there is no doubt that acquisition of such a knowledge base is a costly endeavor. A semantic feature system, on the other hand, involves little investment above that of coding the syntactic and morphological information for lexical items.

Finally, considering the matter of cost effectiveness, one must concede that, given its complexities, a full-blown semantic analysis is a much bigger investment in money than a semantic feature system. As I indicated above, this is not particularly a matter for theoretical systems, but is a very real one for any MT system which hopes to be a production system. The financial cost of developing, running, and maintaining a full-blown semantic analysis versus the gains in analysis might argue against such investment. Semantic feature systems have virtually no additional expenses for their development or operation.

My point here is not to argue against deep semantic analyses, but to point out that the view that such systems represent the only valid approach to semantics in machine translation is problematic at best. In the three cost areas I chose, using a full-blown analysis is not necessarily the best approach. Certainly, a semantic feature system offers real solutions at considerably less cost.

4. Possibilities for Semantic Feature Systems

In my earlier paper (Bennett 1989) I alluded to some possible uses for semantic feature systems without devoting much space to the issues. Here I want to explore the matter somewhat more fully. The work described here is the result of almost two years of study of computational semantics

at the LRC, funded by the State of Texas* . A work-in-progress paper on this project was published last year in *Machine Translation* (Bennett, et al. 1989); my comments here represent a cursory and updated overview of the project. I must note that this work has not been fully implemented in any production MT system, but has been implemented on an experimental basis.

The treatment of aspect is a problem for any machine translation system. While any number of possible approaches may be envisioned, the use of features seems to be the most economical from all three of the cost standpoints I discussed earlier. The semantic feature system uses four semantic primitives, coded as values for a single lexical feature on each verb. In analysis these values are used to calculate the aspect of the predicate at each relevant juncture. For example, given the verb *paint,* which is coded as [dynamic progressive] lexically, aspect is calculated as ACCOMPLISHMENT for the verb phrase *painted the picture,* but EXTENDED ACTIVITY for *painted the picture all week,* as a result of the adverbial phrase. The calculus is simply another operation in the overall analysis of the sentence; the cost in terms of computation is no more than that of any other operation. The sentential aspect, then, is the result of a series of calculations. Thus, *she paints pictures* would be STATIVE aspect, while *she painted pictures* would be EXTENDED ACTIVITY. This approach is fundamentally a synthesis of the approaches described in Meya and Vidal (1988) and Bennett, et al. (1989).

Verb semantics is an issue which has received little attention from either theoretical or computational linguists. Certainly the issue is a difficult one, but verb semantics can potentially solve a number of analysis problems for MT. The approach we have taken is to code a constellation of values for a single feature for each verb. A given verb may have several sets of values depending on its various meanings. For example, *misinterpret* has [non-action change] as its semantic type, while *lie* has [action no-change communication] (= 'tell an untruth') and [non-action no-change po-

* Texas Higher Education Coordinating Board Advanced Research Programs, Grant No. 1631.

sition] (= 'recline'). The verb semantic values will be used in the resolution of several analysis problems, two of which are: differentiating verb meanings and eliminating erroneous verbal complements in the course of analysis.

A third issue for the project is that of anaphora resolution from a semantic standpoint. The approach is to use the existing noun semantic features to resolve problems of anaphora. For example, *she made a lot of cookies for the children; they are very happy* and *she made a lot of cookies for the children; they are delicious* are syntactically similar and unresolvable using only syntax; however, using the semantic information inherent in the second clause in each sentence it is not difficult for the system to find an antecedent capable of happiness in the first instance and one which is edible in the second. Our approach is simply to use the values for the noun semantic feature to get the necessary semantic information from the second clause.

The feature system, outlined above, is able to contribute to semantic analysis accurately with a minimum of cost. Since it requires no additional mechanism for its operation, it does not have any noticeable effect on efficiency. Lexicon acquisition at present is simply a matter of coding verb semantic and lexical aspect features on verb entries and and noun semantic features on noun entries. Such coding requires an insignificant amount of time beyond that required for coding in general. This approach is cost effective, since there is really no additional financial outlay for it.

5. Conclusion

The desirability of a semantic component in a machine translation system is non-controversial. At issue is the form and extent of such a semantic component. While the conventional wisdom of our field leans toward deep semantic analysis as the best way to proceed, I question this view. For a number of reasons, we should consider alternatives to full-blown semantics, notably semantic feature systems, which can serve well to meet the semantic needs of MT systems without the costs of more powerful mechanisms.

REFERENCES

Bar-Hillel, Yehoshua. 1960. "The Present Status of Automatic Translation of Languages", in Franz L. Alt (ed.), *Advances in Computers,* 1, 91-163. New York: Academic Press.

Bennett, Winfield S. 1989. "The Place of Semantics in MT Systems". *Literary and Linguistic Computing,* 4, 200-202.

Bennett, Winfield S., Tanya Herlick, Katherine Hoyt, Joseph Liro and Ana Santisteban. 1989. "Toward a Computational Model of Aspect and Verb Semantics". *Machine Translation*, 4, 247-280.

Carbonell, Jaime G. 1987. "Knowledge-based Machine Translation". Presented as a tutorial at the 25th Annual Meeting of the Association for Computational Linguistics, 6 July 1987, Stanford University, Handout.

Carbonell, Jaime G., and Masura Tomita. 1987. "Knowledge-based Machine Translation, the CMU Approach", in Sergei Nirenburg (ed.), *Machine Translation,* 68-89. London: Cambridge University Press.

Hutchins, W. John. 1986. *Machine Translation: Past, Present, Future.* Chichester, England: Ellis Horwood Ltd.

Hutchins, W. John. 1988. "Recent Developments in Machine Translation", in Dan Maxwell, Klaus Schubert, and Toon Witkam (eds.), *New Directions in Machine Translation,* 7-63. Dordrecht: Foris.

Lehrberger, John and Laurent Bourbeau. 1988. *Machine Translation. Linguistic Characteristics of MT Systems and General Methodology of Evaluation. Studies in French and General Linguistics,* 15. Amsterdam: John Benjamins.

Lytinen, Steven L. 1987. "Integrating Syntax and Semantics", in Sergei Nirenburg (ed.), *Machine Translation,* 302-316. London: Cambridge University Press.

Meya, Montserrat. 1990. "Semantic Disambiguation for PP-Attachment" *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.* Austin: Linguistics Research Center, University of Texas at Austin.

Meya, Montserrat, and Jesus Vidal. 1988. "An Integrated Model for the Treatment of Time in MT-systems", *Proceedings of the 12th International Conference on Computational Linguistics, 2,* 437-441. Budapest: John von Neumann Society for Computational Sciences.

Nirenburg, Sergei, Ira Monarch, Todd Kaufmann, Irene Nirenburg and Jaime Carbonell. 1988. *Acquisition of Very Large Knowledge Bases: Methodology, Tools and Applications.* CMU-CMT-88-108. Pittsburgh: Center for Machine Translation, Carnegie Mellon University.

Weaver, Warren. 1949. "Translation", reprinted in W. N. Locke and A. D. Booth (eds.), *Machine Translation of Languages,* 15-23. New York: John Wiley and Sons, 1955.

Yngve, Victor H. 1964. "Implications of Mechanical Translation Research", *Proceedings of the American Philosophical Society,* 108, 279-281.