THE CHALLENGE OF DOMAIN-INDEPENDENT SPEECH UNDERSTANDING

Robert C. Moore

SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025, USA

ABSTRACT

To achieve widespread acceptance, speech understanding technology needs to be domain independent. Deep understanding, however, appears to require knowledge that is domain specific. Speech understanding technology, therefore, must be partitioned into domain-independent and domain-specific components. Development of domain-independent components could be promoted by creation of semantically annotated corpora. Any such corpus, however, would be difficult to produce and would necessarily be controversial because of lack of widespread agreement on principles of semantic analysis. The use of such a corpus for performance evaluation should therefore be left largely up to the research community rather than being imposed by funding agencies.

1. INTRODUCTION

To realize its full potential, speech recognition must be integrated with natural-language understanding to create speech understanding systems. The limits of recognition without understanding are demonstrated by the fact that, despite dramatic performance improvements over the last ten years, there is still only one large-vocabulary application of speech recognition in widespread use: dictation. As important an accomplishment as that may be, it is a far cry from the "Star Trek" vision of interacting with computers simply by talking to them in ordinary natural language.

In order for enabling technologies like speech recognition or speech understanding to be put into widespread use, the technology must be substantially domain independent. Speech recognition comes close to achieving this goal (although high performance may require domain-specific training data, particularly for language modeling). Speech understanding technology, in so far as it exists at all, is far from domain independent.

In the 1990s, a number of impressive demonstrations of speech understanding were produced, but only for restricted domains. In particular, from 1990 to 1994, DARPA sponsored a series of benchmark tests of speech understanding on the Air Travel Information Service (ATIS) task, in which several systems achieved high performance on unseen test data. All of these systems, however, incorporated crucial components that were highly domain-specific. ATIS cannot, therefore, be claimed to have produced much in the way of domain-independent speech understanding technology.

2. THE DIFFICULTIES OF DOMAIN-INDEPENDENT UNDERSTANDING

Why is domain-independent speech understanding so difficult? Perhaps the greatest problem is that it is very difficult to separate understanding language from understanding the subject matter that the language is about. To illustrate this point, let us consider the task of automated document retrieval. In the age of the World Wide Web, perhaps the foremost potential application of language understanding (whether speech or text) would be to have *real* content-based information retrieval. Imagine being able to ask an Internet search engine a specific question and get back *only* the small number of pages containing the information actually desired rather than thousands or millions of pages containing some subset of the key words in the query, most of which are irrelevant to the request.

Consider what would really be required to have this capability. Suppose a user makes the request, *Show me news stories about natural disasters in the United States in the spring of 1997.* A story about a flood in North Dakota in April 1997 would clearly be relevant. Making this connection, however, requires knowing that April is in the spring, North Dakota is part of the United States, and that floods are natural disasters. But is knowing these facts part of language understanding? Is someone who doesn't know that North Dakota is in the United States ignorant of a fact about the English language, or a fact about geography?

One might concede that language understanding does require such general knowledge of the world, but hope that something like simple taxonomic knowledge might suffice, since the cases we have just cited (April being in the spring, North Dakota being part of the United States, floods being natural disasters) seem to be more or less of this type. Things are not so simple, however. For instance, not all floods are natural disasters. Some floods are not natural

(e.g., when terrorists blow up a dam), and not all floods are disasters (e.g., if the only thing that floods is an empty flood plain). Precise judgments about when floods are natural disasters potentially requires complex reasoning about natural vs. man-made causes and whether enough damage is caused to constitute a disaster.

For another type of example, consider the following pair of sentences from Terry Winograd's [1, p. 295] landmark thesis on natural-language understanding:

> *The city councilmen refused the demonstrators a permit because they feared violence.*
>
> *The city councilmen refused the demonstrators a permit because they advocated revolution.*

In the first sentence, *they* clearly refers to the city councilmen, while in the second, *they* equally clearly refers to the demonstrators. To understand who *they* refers to in these sentences requires more than simple taxonomic knowledge. It seems to require knowledge of typical (or stereotypical) behavior of city councilmen and demonstrators. That surely is not to be counted as knowledge of language, but just as surely, these sentences cannot be fully understood without it.

These examples are intended to show that deep understanding of an utterance requires more than just knowledge of language; it can require arbitrary pieces of knowledge in the domain of the subject matter of the utterance. For this reason, language understanding is sometimes said to be an "AI-complete" problem. That is, understanding language in general seems to require a solution to the artificial intelligence problem in general.

3. TOWARDS DOMAIN-INDEPENDENT COMPONENTS OF UNDERSTANDING

Even though deep language understanding can require arbitrary amounts of domain-specific knowledge, there are still steps that could be taken towards isolating components of understanding that are largely domain independent. Ideally, deep understanding systems could be built by combining such domain-independent components with domain-specific knowledge bases.

What could such an approach hope to accomplish? Consider the following English expressions:

> *John broke the glass*
>
> *the glass was broken by John*
>
> *the glass broken by John*
>
> *the glass that was broken by John*

Although the syntactic phrasing in these expressions varies significantly, the semantic relations among the verb *break* and the noun phrases *the glass* and *John* are the same in all cases. Moreover, recognizing that these relationships are the same does not appear to require any deep domain-specific knowledge about the concepts involved. Only general syntactic knowledge and superficial lexical knowledge seem to be required. A level of representation that captures such relatively shallow semantic regularities might well provide a basis for partitioning the understanding problem into domain-independent and domain-specific parts.

In 1993–4, participants in the DARPA Spoken Language Program put a fairly substantial effort into developing a proposed methodology, called "SemEval" [2], for evaluating performance in identifying semantic relations of the sort discussed here. (In SemEval, these were called "predicate-argument relations".) The SemEval proposal also included evaluation of two other types of semantic analysis, coreference relation identification and word-sense identification.

The goal of the predicate-argument evaluation was to produce a structure that represents the semantic contribution of each word in an utterance and how it relates to the semantic contribution of the other words in the utterance. For example, under one proposal, the utterance *Every blue block is tall* might be represented as something like:

```
(decl (tall 1:(every <(block 1)
                      (blue 1)>)))
```

The coreference identification component of the evaluation was to include a number of different types of contextual relationships including:

1. Strict coreference, where one expression denotes exactly the same entity as some other expression:

   > *Show the flights from Boston to Dallas and the times they arrive.*
   >
   > *they = the flights from Boston to Dallas*

2. Relational coreference, where one expression denotes something bearing a specific relation to an entity denoted by some other expression:

   > *Show flights from Boston to Dallas and discount fares.*
   >
   > *discount fares = discount fares for flights from Boston to Dallas*

3. General constraints from context:

   > *I need to go from Boston to Dallas. Show me all the morning flights.*
   >
   > *the morning flights = the morning flights from Boston to Dallas*

Finally, the goal of word-sense identification component was to mark the content words of an utterance with sense tags from a lexical database such as WordNet [3];

and for function words, identify the cases where different words express the same semantic relationship between content words, such as:

ticket's price
price of a ticket
price for a ticket
price on a ticket

This was an ambitious proposal that was never put into practice for a number of reasons, the foremost one being that DARPA turned away from any common evaluation of understanding performance towards an emphasis on demonstrations of practical utility in tasks of military relevance. Nevertheless, something along the lines the SemEval proposal still seems like a possible way to promote research on the domain-independent aspects of spoken-language understanding.

4. RECOMMENDATIONS

To conclude, here are some personal recommendations for corpus development and evaluation methodologies to promote domain-independent component technology for speech understanding. These recommendations are based on the author's personal experience participating in both the ATIS evaluations and the development of the SemEval proposal.

First, the most important thing is to have a publicly available semantically annotated corpus of sufficient size for experimentation and testing. Even without any formal program of evaluations such as the DARPA benchmarks, the mere existence of such a corpus can stimulate research activity. This has proved to be the case with the Penn Treebank corpus annotated with syntactic bracketings [4]. This corpus has become the benchmark for testing statistical parsing models by numerous researchers, even though no formal program of evaluations exists in this area.

It should be understood that the development of a semantically annotated corpus is a much more ambitious undertaking than transcribing a speech corpus for recognition evaluations or annotating a corpus with syntactic bracketings. The existence of widespread agreement on how to write and spell in the world's major languages makes speech recognition evaluation simple by comparison. Developing an annotated corpus for semantic evaluation is a bit like transcribing a corpus for recognition evaluation would be if we had to simultaneously invent written language in order to carry out the transcription. Annotating syntactic bracketing, while perhaps more difficult than transcribing speech, involves not much more than making explicit the robust intuitions about syntactic structure that are tapped when school children are taught to "diagram" sentences.

The fact that the semantic structure of language is so much murkier than its lexical or syntactic structure means that developing a semantically annotated corpus would be much more of a research project than previous corpus development efforts; and the results will necessarily be more controversial, because of lack of widespread agreement on principles of semantic analysis. However, having some sort of semantically annotated corpus, with all the faults it would be sure to have, is certainly far preferable to having none at all, as is currently the case.

The fact that developing a semantically annotated corpus is problematical in so many ways also argues for using a very light hand in establishing any evaluation based on such a corpus. Any evaluation should be viewed as completely voluntary by all parties concerned. Indeed, my personal view is that DARPA's penchant for establishing complex, multidimensional benchmark evaluations for speech and language technology that all contractors feel obligated to participate in has at best produced very mixed results with respect to accelerating progress. It is undoubtedly true that the contractor community as a whole makes rapid progress on the *particular* tasks that DARPA has chosen to evaluate on. However, by choosing complex, multidimensional tasks, DARPA tends to force all contractors to spread their efforts across all aspects of the task, often with the result that contractors end up concentrating their efforts on whatever is the weakest link in their evaluated system, whether or not they have any good ideas about that subtask or whether it is the most important thing to be working on for the long term. I believe we saw this in the ATIS evaluations, and I fear the pattern may be repeating in the broadcast news evaluations.

Moreover, in the case of evaluating attempts to recover domain- and task-independent semantic structure, it is important to realize that this is not the only valid approach to developing speech understanding technology. Another approach is to concentrate on building tools to make domain- and task-specific systems easier to create. Based on the results of the ATIS efforts and other task-specific speech understanding systems, this seems to be a much more practical approach for the near-to-medium term; and it would be a mistake to try to specify one grand, field-defining evaluation that excludes making improvements in tools for building task-specific systems.

Nevertheless, creating domain-independent technologies for speech understanding is certainly an important goal that should be encouraged. My recommendation then, is to take an "if you build it, they will come" approach. Create a semantically annotated corpus, make research funds available to exploit it, and let science happen.

5. REFERENCES

[1] T. Winograd, *Procedures as a Representation for Data in a Computer Program for Understanding Natural*

Language, Project MAC TR–84, Massachusetts Institute of Technology, Cambridge, MA (1971).

[2] R. C. Moore, "Semantic Evaluation for Spoken-Language Systems," in *Proc. ARPA Human Language Technology Workshop*, Plainsboro, NJ, pp. 126–131 (1994).

[3] G. A. Miller (ed.), "WordNet: An On-Line Lexical Database," International Journal of Lexicography (special issue), Vol. 3, No. 4, pp. 235–312 (1990).

[4] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330 (1993).