BBN ATIS System Progress Report - June 1990

M. Bates, R. Bobrow, S. Boisen, R. Ingria, D. Stallard

BBN Systems and Technologies Corp.
10 Moulton Street
Cambridge, MA 02138

Abstract

This paper reports recent progress on the development of the Delphi natural language component of the BBN spoken language system for the ATIS domain, focussing on the comparative evaluation performed by NIST in June, 1990.

1. Delphi and Parlance/Learner

Delphi is BBN's research NL system (formerly called CFG), which is based on a unification grammar and which incorporates semantics into the unification framework. Delphi is the NL component of the BBN HARC System. Delphi can be changed very quickly, but has no easy way to build knowledge bases rapidly.

The Parlance™ system is a commercial NL interface to relational databases, and is based on an ATN parser and grammar. Parlance has the advantage of an extensive knowledge acquisition system called the Learner™, so our previously reported approach (Ingria and Ramshaw 1989) has been to use the Learner to create a lexicon with morphological and syntactic information, a domain model with semantic information, and mapping rules from the domain model to the database. These knowledge bases were then imported for use by Delphi.

As a side-effect of using the Learner, a Parlance configuration for the ATIS domain was created, so results using that system are reported here for comparison.

2. Analysis of Delphi's Performance on the Blind Test

The original score of the BBN Delphi system on the 93 sentence ATIS blind test was 53 sentences correct, 2 sentences not correct and 38 sentences not answered.

Of the two sentences judged not correct, one is the result of a mistake in the canonical answer set provided by NIST. The other resulted from a lack of agreement on which table a fare is computed from. (There are a few cases in the ATIS database where information can be retrieved by several different paths, sometimes resulting in different data.) These reulsts indicate that the production of incorrect answers is not a significant problem for Delphi, so the remainder of this section focusses on the sentences not answered.

Several causes account for the 38 sentences not answered, with the primary one being words or word senses that were not seen in the training material, or which simply did get entered into the lexicon. Another important cause was the failure on the part of the system to infer a relationship between known word senses that a human user would have recognized as implicit. A quantitative breakdown is given in Table 1, and more detailed discussion follows.

Reason	#	%
Word senses missing:	14	37%
Lack of inference:	10	26%
Grammar:	5	13%
SNOR bugs:	2	5%
Miscellaneous:	7	18%
TOTAL	38	

Table 1. Analysis of BBN Delphi Performance on "No Answer" Queries.

2.1 Word Senses Not Previously Encountered

An example of a word not seen in the training is the verb "to service", seen in the following test set sentence:

I need information on airlines servicing Boston flying from Dallas.

The word "service" was present only in noun form in the training data, as in "class of service". Here it clearly has the same meaning as the verb "serve".

Another example of a word meaning not seen in the training but appearing in the test was "fly" in the following:

What type of aircraft is flying United Airlines flight 953?

Normally, "fly" appears intransitively, and is something a flight does by itself. In the sense of "fly" seen above, it would appear that "flying" is something an aircraft does TO a flight. Both this example and the one above are trivial to correct: the relevant word sense simply needs to be added to the lexicon.

The major concept in the domain that we did not cover was the notion of a "ticket". Rather than associate the word

"ticket" with a concept on its own, as something different from the meanings of "flight" and "fare", we chose to make it synonymous with the word "fare". This meant that the following sentence did not appear meaningful to Delphi:

Are there any excursion fares for round-trip tickets from Dallas to Boston?

Phrases such as "excursion fares for round-trip fares" would have similarly been meaningless. The solution for this case simply requires that a new concept for "ticket" be added to the domain model, that its relations with already existing concepts be established, and that the word "ticket" be associated with this concept. Eight sentences in the blind test used the word "ticket" in this way.

2.2 Lack of Inference

The second most important category is one where combinations of existing word senses occur which, while meaningful to a human being, do not appear meaningful to the system because it is unable to infer a missing element. An example from the ATIS test set is:

I need flight times from Boston to Dallas leaving on Saturday morning.

The problem here is that "flight times" are times, and times are not "from Boston", nor do they "leave on Saturday morning". The utterance would make sense, of course, given the following paraphrase which any human speaker could easily supply:

I need flight times for flights from Boston to Dallas leaving on Saturday morning.

There are number of examples in the ATIS test set of this kind, including several in which information about tickets that "fly" from place A to place B is requested.

Building an inference facility into the system has been a goal of our project even before the present test exercise, and we feel that the structure of our knowledge representation mechanisms (see particularly the description in Bobrow, Ingria & Stallard, 1990) will enable us to undertake this effort in the near future.

2.3 Other Reasons

Other understanding failures arose due to grammar issues (notably, the lack of preposed PP complements), a minor problem in our handling of SNOR input (e.g. ``12nd" for ``12th") which affected the performance of 3 queries, and errors in other phases of the system.

3. Parlance Performance

A few comments are in order about the performance of the Parlance system on the test data. Parlance achieved 58 correct answers, 7 not correct, and 28 not answered. Of the 7 which were classified as incorrect, 4 of them came from interpretations which a human being could easily have produced in response to the same queries (e.g., whether "the distance from San Francisco airport to downtown" refers only to downtown San Francisco or whether it also includes downtown Oakland); 1 came from a minor problem in handling SNOR input; 1 was the result of a wrong canonical answer; only 1 represented a bug in the system.

Of the 28 queries Parlance did not answer, 13 involved new words or word senses, 4 required inference beyond the capabilities of the system, 5 were not parsable becuase they involved structures that we believe occur commonly only in spoken language (Parlance's grammar has been constructed for typed input, and we did not make changes in the grammar specifically to accommodate spoken language), 2 involved a minor problem with SNOR input, and 4 were the result of miscellaneous other problems.

Most of the problem queries could be handled by Parlance with very little additional effort; within a day after the blind test, Parlance understood correctly 83 of the 93 test queries (89%).

4. Conclusions

There is evidence that intra-speaker varability in linguistic structure is farily low, but that inter-speakers variability is very high. In other words, a given speaker, at least in a single session, tends to use the same forms over and over again (e.g., "tickets flying"), and each new speaker (at least so far) tends to use locutions different from previous speakers.

This leads us to conclude that much more training data is needed in order to adequately prepare for evaluations, particularly when the test material is drawn from subjects not represented in the training data. It would further increase the validity of the test if more than one domain were used.

Our approach of having using the Learner for knowledge acquisition and the Delphi system as the primary NL component of HARC is successful and should be continued.

References

Ingria, Robert J.P. and Lance Ramshaw, Porting to New Domains Using the Learner™, in *Proceedings of the DAPRA Speech and Natural Language Workshop*, October, 1989.

Bobrow, Robert, Robert J.P. Ingria, and David Stallard, *Syntactic and Semantic Knowledge in Unification Grammar*, this volume.