BBN HARC and DELPHI Results on the ATIS Benchmarks - February 1991

S. Austin, D. Ayuso, M. Bates, R. Bobrow, R. Ingria, J. Makhoul, P. Placeway, and R. Schwartz, D. Stallard

BBN Systems and Technologies
10 Moulton Street
Cambridge, MA 02138

ABSTRACT

This paper presents the test results of running BBN's HARC spoken language system and DELPHI natural language understanding system on the ATIS benchmarks.

We give a brief system overview, and review the major changes that have been made in Delphi since the last DARPA SLS workshop. We will briefly discuss the development and training process, and then present our test results and an analysis of their meaning.

SYSTEM OVERVIEW

Delphi is BBN's research NL system, which is based on a unification grammar and which incorporates semantics into the unification framework. Delphi is the NL component of the BBN HARC (Hear and Respond to Continuous Speech) system; integrated with the BYBLOS speech recognition system using an N-best architecture [1,2].

Figure 1 shows the relationships among the components of HARC, and their inputs and outputs.

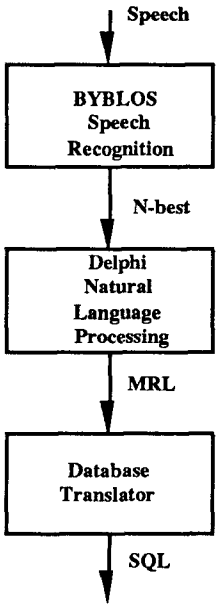

Figure 1: The BBN HARC System

RECENT CHANGES IN DELPHI

The BBN Delphi natural language understanding system which was reported in June, 1990 [3] has been changed and improved in a number of ways:

1. The addition of statistical agenda capabilities to the parser. This achieved a considerable reduction in parse times while at the same time producing a desirable parse as the first interpretation in most cases. It is reported on in detail elsewhere in this volume [4].

2. A streamlined semantic processor. This component now uses "mapping units" to handle a number of phenomena that would otherwise result in a combinatorial explosion of rules. This allows the rules to be expressed more simply, with less possibility of forgetting to include a particular syntactic pattern. It also makes possible a more general treatment of the kinds of metonomy which occur most frequently in the ATIS domain. Mapping units are described elsewhere in this volume [5].

3. An extended and improved dialogue component. In addition to covering domain-independent discourse phenomena, this component now also utilizes a domain-dependent frame-like representation of the discourse state, which makes it possible for Delphi to recognize implicit references to prior context as well as explicit reference. Implicit reference is frequent in the ATIS domain (e.g., "Show the flights from Boston to Dallas. What meals are served?").

4. An N-best integration of speech recognition output with Delphi's NL processing. Our initial results in using this architecture for integration have been very positive.

5. A military application task. We began to apply HARC to a demonstration task involved with military logistical planning. This system, called DART (Dynamic Analytical Replanning Tool), our initial integration of speech understanding with it, and an outline of our plans to expand that integration are described elsewhere in this volume [6].

Source	Total # of Sentences	Class A Sentences	Class A Answers	Class D1 Pairs	Class D1 Answers
June '90 to Feb '91					
SRI	??	177	220	20	20
TI	??	106		88	44
MIT	1647	not classified		not classified	
CMU	632	not classified		not classified	
Up to June '90					
TI	776	551	551		
TI (June '90 test data)	93	90	90		
Total	>3431	924	861	108	64

Figure 2: Common Training Data

NL TRAINING

Training data for this phase of the SLS program was primarily the 551 queries of training data that were available before the evaluation in June, 1990. A summary of the training data is given in figure 2.

Figure 2 also shows that although over 3400 queries were collected from all sources, fewer than 900 Class A queries with reference answers are available for training purposes, and only 64 Class D1 dialogue pairs with reference answers are available.

The data from MIT and CMU, although initially promising because of its volume, proved not to be very useful, because the queries were not classified (as Class A, Class D1, etc,), and reference SQL and answers were not provided. This meant that it was not possible during the development period to run these queries through our system and automatically determine whether the answers that were produced were correct or not.

PERFORMANCE

Figure 3 gives the results of BBN's performance on various benchmark NL and SLS tests, as of the February 19, 1991, the date of the workshop.

These results are comparable to Delphi's performance last June as reported by NIST [7]. Had the currrent scoring metric been in place then, Delphi would have scored 57.8% on Class A.

	NL, Class A	NL, Class A	NL, Class D1	NL, Class D1	NL, Class D1	SLS, Class A
Notes	1	2	1, 3	1, 3	2,4	1
System	Delphi	Delphi	Delphi	Delphi	Delphi	HARC
Test set size	145 S's	145 S's	38 pairs	38 pairs	38 pairs	145 S's
Date submitted	2/7/91	2/16/91	2/7/91	2/13/91	2/16/91	2/7/91
Right	58.6%	62.8%	26.3%	68.4%	68.4%	57.9%
Wrong	7.5%	6.9%	7.9%	7.8%	2.6%	15.2%
NA (not answered)	33.7%	30.3%	65.8%	23.6%	28.9%	26.8%
Weighted Error	49.0%	44.1%	81.6%	39.5%	34.0%	57.2%
Score	**51.0%**	**55.9%**	**18.4%**	**60.5%**	**66.0%**	**42.8%**

Figure 3: BBN's ATIS Benchmark Results, February 1991

Notes:

1. These results were scored by NIST before the workshop; these numbers reflect the rescoring NIST did after the workshop.
2. These results were submitted to NIST before the workshop, but were not scored. The only change made to the system between the first NL run and the second was to fix a minor bug in the SNOR translator which formats the input data for the parser.
3. The first class D1 test uncovered a problem in our system's backend translator, which was fixed for the second run. See the discussion section below for more information.
4. The difference between this and the previous run involved how to score pairs which gave NA for Q1. See discussion below.

DISCUSSION

There are several global points to make before discussing each test separately.

As was the case last June, some problems showed up in the test set itself. Several queries that were not actually Class A were included in the original test set; their removal resulted in the 145 item test. (More items may have been removed before the final official scoring.) Also, the reference answers for several queries had to be augmented during the scoring, to account for ambiguities that had gone unnoticed during the preparation of the test set.

We believe that such problems are unavoidable, but minor and easy to fix, so we do not recommend any major change in the evaluation methodology, but recommend that sufficient time for sites to check the reference answers be allocated in the schedule for the next evaluation.

A system which performs well on Class A but less well than expected on Class D might be using rather brittle techniques to deal with Class A which do not generalize effectively to discourse. It is also interesting to note that this is not at all a problem for us.

In fact, the best versions of our system did better on Class D than Class A, which is counterintuitive. One would expect that the probability of getting a D1 pair correct is less than the product of the probability of getting a Class A sentence correct, because not only must two sentences be processed, but the processing of the second is likely to be harder than a simple Class A sentence, since it must involve reference resolution or other discourse processing. This is best explained by the fact that the D1 test set was short, rather easy, and involved more repetition of similar query types than the Class A test.

Out of Vocabulary Words

One of the main reasons for the relatively high number of NA answers to Class A utterances was simply vocabulary: Nineteen of the test utterances (13% of the test set) contained vocabulary outside Delphi's lexicon. The lack of training data clearly had an impact here, since one of the great benefits of training data is increased vocabulary.

It is worthy of note that since there is no control vocabulary among the various systems, it is very difficult to meaningfully compare the performance of multiple systems. Using the official data presented, it is impossible to tell the difference between a system that simply lacks some vocabulary entries and one that has a larger lexicon but which cannot syntactically or semantically process many of the test utterances.

NL Class A

The only difference between the original score and the second one is that a small problem in the formatting of SNOR input for the parser was fixed. The understanding component (syntax, semantics, and discourse processing), which is what the Class A test is attempting to measure, was completely unchanged.

NL Class D1

The initial results of our D1 evaluation were shocking, but a quick investigation revealed several interesting facts:

1. Sixteen of the utterances that yielded a NA response were in fact understood perfectly correctly by the syntactic, semantic, and discourse components of Delphi, and produced correct MRL expressions (refer to figure 1). But there was a simple bug in the backend translator that turns MRL expressions into SQL expressions, and all 16 utterances tickled that bug.

2. Of those 16 utterances, 14 of them were extremely similar in words, syntactic form, and semantic import. That is, 37% of the test pairs has this single form. The fact that the test set was significantly skewed toward one particular type of second utterance enormously magnified the effect of what was actually a very small problem.

3. Fixing that one problem resulted in all 16 of those utterances going through to SQL, and producing the correct answer.

Because the problem was not in the language understanding component of Delphi, because the test set was so skewed, and because the purpose of the D1 test at this stage was to test the test methodology more than to test the dialogue systems, we fixed the problem and resubmitted the results to NIST. The resulting score (60.5%) is much more representative of the true capabilities of our dialogue component than the original score.

An additional problem with scoring D1 surfaced during this evaluation. In a case where Q1 of a Q1-Q2 pair is not answered by the system, what should be done with Q2? We allowed our system to run Q2 as a context-independent query if possible, but expected that the scoring software would treat it as NA, since it is never possible to get a correct answer to a context-dependent query if the context is not understood. But the scoring package counted such answers as wrong. The final run (66%) of our Class D test produces NA for these cases.

SLS Class A

It is remarkable, and quite unexpected, that the score for the speech test of Class A should be so close to the NL test on exactly the same set of utterances. This indicates that the N-best strategy for integrating speech and NL processing seems to be working. Because the speech recognition component is currently producing about 16.2% and a sentence error rate of about 54.1% [2], it is obvious that the natural language component is making up for some of the errors made by the speech recognition component

Some interesting results emerged from our analysis of how the speech and NL components worked together. The following statistics are from the original 148 utterance Class A test set (which was later reduced to 145 by NIST after removal of 3 queries which were not actually Class A).

In 58.8% of the cases, NL chose the 1-best utterance. Of these,
- 72.2% were correct speech hypotheses.
- 14.9% had the correct speech hypothesis in the N-best.
- 12.6% didn't have the correct speech hypothesis in the N-best.

In 20.3% of the cases, NL chose one of the N-best utterances.
26.7% of these were correct speech hypotheses.
6.7% had the correct speech hypothesis in the N-best.
66.7% didn't have the correct speech hypothesis in the N-best.

In 20.9% of the cases, NL chose none of the utterances.

Looking at the correctness of the answers produced by the HARC SLS system, we find the following (again, from 148 Class A utterances):

In 58.5% of the cases, NL chose the 1-best hypothesis.
77.0% of these were T
19.5% of these were F
3.5% of these were NA.

In 20.3% of the cases, NL chose one of the N-best utterances.
50% of these were T
30% of these were F
20% of these were NA.

In 20.9% of the cases, NL chose none of the N-best, so
100% of these were NA.

CONCLUSIONS

After the last evaluation, our primary conclusion was as follows [3], p 126.:

> "There is evidence that intra-speaker variability in linguistic structure is fairly low, but that inter-speaker 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 evaluation..."

Our experience in this evaluation only serves to underscore and reinforce our original conclusion. Large amounts (several thousand queries) of adequately prepared training data (classified, with reference SQL and reference answers) must be available in time for sites to use it for several months of development before a truly meaningful evaluation can be conducted.

We have also developed some additional suggestions for dialogue evaluation, which are detailed in a separate paper [8].

ACKNOWLEDGEMENTS

The work reported here was supported by the Advanced Research Projects Agency and was monitored by the Office of Naval Research under Contract No. N00014-89-C-0008. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research projects Agency or the United States Government.

REFERENCES

1.Chow, Y.L.., et al., "BYBLOS: The BBN Continuous Speech Recognition System", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas TX, April 1987, pp89-92, Paper No. 3.7.

2. Kubala, F., et al., "BYBLOS Speech Recognition Benchmark Results", (in this proceedings).

3. Bates, M., Boisen, S., Ingria, R. and Stallard, D. "BBN ATIS System Progress Report - June 1990", *Proceedings of the Speech and Natural Language Workshop* (June, 1990), Morgan Kaufmann Publishers, Inc., 1990.

4. Bobrow, R. "Statistical Agenda Parsing", (in this proceedings).

5. Bobrow, R., Ingria, R., Stallard, D, "The 'Mapping Unit' Approach to Subcategorization", (in this proceedings).

6. Bates, M., Ellard, D., Peterson, P., and Shaked, V. "Using Spoken Language to Facilitate Military Transportation Planning", (in this proceedings).

7. Pallett, D.S, et al, "DARPA ATIS Test Results June 1990", in *Proceedings Speech and Natural Language Workshop, June 1990*, Morgan Kaufmann Publishers, Inc., June, 1990.

8. Bates, M., and Ayuso, D., "A Proposal for Incremental Dialogue Evaluation", (in this proceedings).