

THE BBN/HARC SPOKEN LANGUAGE UNDERSTANDING SYSTEM

*Madeleine Bates, Robert Bobrow, Pascale Fung, Robert Ingria,
Francis Kubala, John Makhoul, Long Nguyen, Richard Schwartz, David Stallard*

BBN Systems and Technologies
Cambridge, MA 02138, USA

ABSTRACT

We describe the design and performance of a complete spoken language understanding system currently under development at BBN. The system, dubbed HARC (Hear And Respond to Continuous speech), successfully integrates state-of-the-art speech recognition and natural language understanding subsystems. The system has been tested extensively on a restricted airline travel information (ATIS) domain with a vocabulary of about 2000 words. HARC is implemented in portable, high-level software that runs in real time on today's workstations to support interactive online human-machine dialogs. No special purpose hardware is required other than an A/D converter to digitize the speech. The system works well for any native speaker of American English and does not require any enrollment data from the users. We present results of formal DARPA tests in Feb. '92 and Nov. '92.

1. OVERVIEW

The BBN HARC spoken language system weds two technologies, speech recognition and natural language understanding, into a deployable human-machine interface. The problem of understanding goal-directed spontaneous speech is harder than recognizing and understanding read text, due to greater variety in the speech and language produced. We have made minor modifications to our speech recognition and understanding methods to deal with these variabilities. The speech recognition uses a novel multipass search strategy that allows great flexibility and efficiency in the application of powerful knowledge sources. The natural language system is based on formal linguistic principles with extensions to deal with speech errors and to make it robust to natural variations in language. The result is a very usable system for domains of moderate complexity.

While the techniques used here are general, the most complete test of the whole system thus far was made using the ATIS corpus, which is briefly described in Section 2. Section 3 describes the techniques used and the results obtained for speech recognition, and Section 4 is devoted to natural language. The methods for combining speech recognition and language understanding, along with results for the combined system are given in Section 5. Finally, in Section 6, we describe a real-time implementation of the system that runs entirely in software on a single workstation.

More details on the specific techniques, the ATIS corpus, and the results can be found in the papers presented at the 1992 and 1993 DARPA Workshop on Speech and Natural Language [1, 2, 3, 4].

2. THE ATIS DOMAIN AND CORPUS

The Air Travel Information Service (ATIS) is a system for getting information about flights. The information contained in the database is similar to that found in the Official Airline Guide (OAG) but is for a small number of cities. The ATIS corpus consists of spoken queries by a large number of users who were trying to solve travel related problems. The ATIS2 training corpus consists of 12,214 spontaneous utterances from 349 subjects who were using simulated or real speech understanding systems in order to obtain realistic speech and language. The data originated from 5 collection sites using a variety of strategies for eliciting and capturing spontaneous queries from the subjects [4].

Each sentence in the corpus was classified as class A (self contained meaning), class D (referring to some previous sentence), or class X (impossible to answer for a variety of reasons). The speech recognition systems were tested on all three classes, although the results for classes A and D were given more importance. The natural language system and combined speech understanding systems were scored only on classes A and D, although they were presented with all of the test sentences in their original order.

The Feb '92 and Nov '92 evaluation test sets had 971 and 967 sentences respectively from 37 and 35 speakers with an equal number of sentences from all 5 sites. For both test sets, about 43% of the sentences were class A, 27% were class D, and 30% were class X. The recognition mode was speaker-independent – the test speakers were not in the training set and every sentence was treated independently.

3. BYBLOS – SPEECH RECOGNITION

BYBLOS is a state-of-the-art, phonetically-based, continuous speech recognition system that has been under development at BBN for over seven years. This system introduced an effective strategy for using context-dependent phonetic hidden Markov models (HMM) and demonstrated their feasibility for large vocabulary, continuous speech applications [5]. Over the years, the core algorithms have been refined primarily on artificial applications using read speech for training and testing.

3.1. New Extensions for Spontaneous Speech

Spontaneous queries spoken in a problem-solving dialog exhibit a wide variety of disfluencies. There were three very frequent effects that we attempted to solve – excessively long segments of waveform with no speech, poorly transcribed training utterances, and a variety of nonspeech sounds produced by the user.

We eliminated long periods of background with a heuristic energy-based speech detector. But typically, there are many untranscribed short segments of background silence remaining in the waveforms after truncating the long ones, which measurably

degrade the performance gain usually derived from using cross-word-boundary triphone HMMs. We mark the missing silence locations automatically by running the recognizer on the training data constrained to the correct word sequence, but allowing optional silence between words. Then we retrained the model using the output of the recognizer as *corrected* transcriptions.

Spontaneous data from naive speakers has a large number and variety of nonspeech events, such as pause fillers (um's and uh's), throat clearings, coughs, laughter, and heavy breath noise. We attempted to model a dozen broad classes of nonspeech sounds that were both prominent and numerous. However, when we allowed the decoder to find nonspeech models between words, there were more false detections than correct ones. Because our silence model had little difficulty dealing with breath noises, lip smacks, and other noises, our best results were achieved by making the nonspeech models very unlikely in the grammar.

3.2. Forward-Backward N-best Search Strategy

The BYBLOS speech recognition system uses a novel multi-pass search strategy designed to use progressively more detailed models on a correspondingly reduced search space. It produces an ordered list of the N top-scoring hypotheses which is then reordered by several detailed knowledge sources. This N-best strategy [6, 7] permits the use of otherwise computationally prohibitive models by greatly reducing the search space to a few (N=20-100) word sequences. It has enabled us to use cross-word-boundary triphone models and trigram language models with ease. The N-best list is also a robust interface between speech and natural language that provides a way to recover from speech errors.

We use a 4-pass approach to produce the N-best lists for natural language processing.

1. A forward pass with a bigram grammar and discrete HMM models saves the top word-ending scores and times [8].
2. A fast time-synchronous backward pass produces an initial N-best list using the Word-Dependent N-best algorithm[9].
3. Each of the N hypotheses is rescored with cross-word-boundary triphones and semi-continuous density HMMs.
4. The N-best list is rescored with a trigram grammar.

Each utterance is quantized and decoded three times, once with each gender-dependent model and once with a gender-independent model. (In the Feb '92 test we did not use the gender-independent model.) For each utterance, the N-best list with the highest top-1 hypothesis score is chosen. The top choice in the final list constitutes the speech recognition results reported below. Then the entire list is passed to the language understanding component for interpretation.

3.3. Training Conditions

Below, we provide separate statistics for the Feb(Nov) test as n1(n2). We used speech data from the ATIS2 subcorpus exclusively to train the parameters of the acoustic model. However, we filtered the training data for quality in several ways. We removed from the training any utterances that were marked as truncated, containing a word fragment, or containing rare nonspeech events. Our forward-backward training program also automatically rejects any input that fails to align properly, thereby discarding many sentences with incorrect transcriptions. These steps removed 1,200(1,289) utterances from consideration. After holding out a development test set of 890(971) sentences, we were left with a total of 7670(10925) utterances for training the HMMs.

The recognition lexicon contained 1881(1830) words derived from the training corpus and all the words and natural extensions from the ATIS application database. We also added about

400 concatenated word tokens for commonly occurring sequences such as WASHINGTON.D.C. and D.C.TEN. Only 0.4%(0.6%) of the words in the test set were not in the lexicon.

For statistical language model training we used all available 14,500(17,313) sentence texts from ATIS0, ATIS1, and ATIS2 (excluding the development test sentences from the language model training during the development phase). We estimated the parameters of our statistical bigram and trigram grammars using a new backing-off procedure[10]. The n-grams were computed on 1054(1090) semantic word classes in order to share the very sparse training (most words remained singletons in their class).

3.4. Speech Recognition Results

Table 1 shows the official results for BYBLOS on this evaluation, broken down by utterance class. We also show the average perplexity of the bigram and trigram language models as measured on the evaluation test sets (ignoring out-of-vocabulary words).

Sentence Class	Bigram Perplex	Trigram Perplex	Feb(Nov) % Word Errors
A+D	17	12	6.2(4.3)
A+D+X	20	15	9.4(7.6)
A	15	10	5.8(4.0)
D	20	14	7.0(4.8)
X	35	28	17.2(14.5)

Table 1: Official SPREC results on Feb(Nov) '92 test sets.

The word error rate in each category was lower than any other speech system reporting on this data. The recognition performance was well correlated with the measured perplexities. The trigram language model consistently, but rather modestly, reduced perplexity across all three classes. (However, we observed that word error was reduced by 40% on classes A+D with the trigram model.)

The performance on the class X utterances (those which are unevaluable with respect to the database) is markedly worse than either class A or D utterances. Since these utterances are not evaluable by the natural language component, it does not seem profitable to try to improve the speech performance on these utterances for a spoken language system.

4. DELPHI - NATURAL LANGUAGE UNDERSTANDING

The natural language (NL) component of HARC is the DELPHI system. DELPHI uses a definite clause grammar formalism, augmented by the use of constraint nodes [11] and a labelled argument formalism [3]. Our initial parser used a standard context-free parsing algorithm, extended to handle a unification-based grammar. It was then modified to integrate semantic processing with parsing, so that only semantically coherent structures would be placed in the syntactic chart. The speed and robustness were enhanced by switching to an agenda-based chart-parser, with scheduling depending on the measured statistical likelihood of grammatical rules [12]. This greatly reduced the search space for the best parse.

The most recent version of DELPHI includes changes to the syntactic and semantic components that maintain the tight syntactic/semantic coupling characteristic of earlier versions, while allowing the system to provide semantic interpretations of input which has no valid global syntactic analysis. This included the development of a "fallback component" [2], in which statistical estimates play an important role. This component allows DELPHI to deal effectively with linguistically ill-formed inputs that

are common in spontaneous speech, as well as with the word errors produced by the speech recognizer.

4.1. Parsing as Transduction – Grammatical Relations

The DELPHI parser is not a device for constructing syntactic trees, but an information transducer. Semantic interpretation is a process operating on a set of messages characterizing local “grammatical relations” among phrases, rather than as a recursive tree walk over a globally complete and coherent parse tree. The grammar has been reoriented around local grammatical relations such as deep-structure subject and object, as well as other adjunct-like relations. The goal of the parser is to make these local grammatical relations (which are primarily encoded in ordering and constituency of phrases) readily available to the semantic interpreter.

From the point of view of a syntactic-semantic transducer, the key point of any grammatical relation is that it licenses a small number of semantic relations between the “meanings” of the related constituents. Sometimes the grammatical relation constrains the semantic relation in ways that cannot be predicted from the semantics of the constituents alone (e.g. Given “John”, “Mary”, and “kissed”, only the grammatical relations or prior world knowledge determine who gave and who received). Other times the grammatical relation simply licenses the only plausible semantic relation (e.g. “John”, “hamburger”, and “ate”). Finally, in sentences like “John ate the fries but rejected the hamburger”, semantics would allow the hamburger to be eaten, but syntax tells us that it was not.

Grammatical relations are expressed in the grammar by giving each element of the right hand side of a grammar rule a grammatical relation as a label. A typical rule, in schematic form, is:

```
(NP ...) → :HEAD (NP ...) :PP-COMP (PP :PREP ...)
```

which says that a noun phrase followed by a prepositional phrase provides evidence for the relation PP-COMP between the PP and HEAD of the NP.

One of the right-hand elements must be labeled the “head” of the rule, and is the initial source of information about the semantic and syntactic “binding state” which controls whether other elements of the right-hand side can “bind” to the head via their labeled relation.

This view made it possible to both decrease the number of grammar rules (from 1143 to 453) and increase syntactic coverage. Most attachments can be modelled by simple binary adjunction, and since the details of the syntactic tree structure are not central to a transducer, each adjunct can be seen as being “logically attached” to the “head” of the constituent. This scheme allows the adjunction rules of the grammar to be combined together in novel ways, governed by the lexical semantics of individual words. The grammar writer does not need to foresee all possible combinations.

4.2. “Binding rules” – the Semantics of Grammatical Relations

The interface between parsing and semantics is a dynamic process structured as two coroutines in a cascade. The input to the semantic interpreter is a sequence of messages, each requesting the semantic “binding” of some constituent to a head. A set of “binding rules” for each grammatical relation licenses the binding of a constituent to a head via that relation by specifying the semantic implications of binding. These rules specify features

of the semantic structure of the head and bound constituent that must be true for binding to take place, and may also specify syntactic requirements. Rules may also allow certain semantic roles (such as time specification) to have multiple fillers, while other roles may allow just one filler.

As adjuncts are added to a structure, the binding list is conditionally extended as long as semantic coherence is maintained. When a constituent is syntactically complete (i.e., no more adjuncts are to be added), DELPHI evaluates rules that check for semantic completeness and produce an “interpretation” of the constituent.

4.3. Robustness Based on Statistics and Semantics

Unfortunately, simply having a transduction system with semantics based on grammatical relations does not deal directly with the key issue of robustness – the ability to make sense of an input even if it cannot be assigned a well-formed global syntactic analysis. In DELPHI we view standard global parsing as merely one way to obtain evidence for the existence of the grammatical relations in an input string. DELPHI’s strategy is based on two other sources of information. DELPHI applies semantic constraints incrementally during the parsing process, so that only semantically coherent grammatical relations are considered. Additionally, DELPHI has statistical information on the likelihood of various word senses, grammatical rules, and grammatical-semantic transductions. Thus DELPHI can rule out many locally possible grammatical relations on the basis of semantic incoherence, and can rank alternative local structures on the basis of empirically measured probabilities. The net result is that even in the absence of a global parse, DELPHI can quickly and reliably produce the most probable local grammatical relations and semantic content of various fragments.

DELPHI first attempts to obtain a complete syntactic analysis of its input, using its agenda-based best-first parsing algorithm. If it is unable to do this, it uses the parser in a fragment-production mode, which produces the most probable structure for an initial segment of the input, then restarts the parser in a top down mode on the first element of the unparsed string whose lexical category provides a reasonable anchor for top-down prediction. This process is repeated until the entire input is spanned with fragments. Experiments have shown that the combination of statistical evaluation and semantic constraints produces chunks of the input that are very useful for interpretation by non-syntactically-driven strategies.

4.4. Advantages of This Approach

The separation of syntactic grammar rules from semantic binding and completion rules greatly facilitates fragment parsing. While it allows syntax and semantics to be strongly coupled in terms of processing (parsing and semantic interpretation) it allows them to be essentially decoupled in terms of notation. This makes the grammar and the semantics considerably easier to modify and maintain.

5. COMBINED SPOKEN LANGUAGE SYSTEM

The basic interface between BYBLOS and DELPHI in HARC is the N-best list. In the most basic strategy, we allowed the NL component to search arbitrarily far down the N-best list until it either found a hypothesis that produced a database retrieval or reached the end of the N-best list. However, we have noticed in the past that, while it was beneficial for NL to look beyond the first hypothesis in an N-best list, the answers obtained by NL from speech output tended to degrade the further down in the N-best list they were obtained.

We optimized both the depth of the search that NL performed on the N-best output of speech and how we used the fall-back strategies for NL text processing [2]. We found that, given the current performance of all the components, the optimal number of hypotheses to consider was N=10. Furthermore, we found that rather than applying the fall-back mechanism to each of these hypotheses in turn, it was better to make one pass through the N-best hypotheses using the full parsing strategy, and then, if no sentences were accepted, make another pass using the fall-back strategy.

In Tables 2 and 3 we show the official performance on the Feb and Nov '92 evaluation data. The percent correct and the weighted error rate is given for the DELPHI system operating on the transcribed text (NL) and for the combined HARC system (SLS). The weighted error measure weights incorrect answers twice as much as no answer.

Corpus	NL Cor	NL WE	SLS Cor	SLS WE
A+D	76.7	33.9	71.8	43.7
A	80.1	26.4	74.9	35.8
D	71.9	44.6	67.4	54.7

Table 2: %Correct and Weighted error on the Feb '92 test set.

Corpus	NL Cor	NL WE	SLS Cor	SLS WE
A+D	85.0	22.0	81.0	30.6
A	88.8	15.7	84.6	23.7
D	78.6	32.8	74.9	42.5

Table 3: %Correct and Weighted error on the Nov '92 test set.

The weighted error on context-dependent sentences (D) is about twice that on sentences that stand alone (A). First, it is often difficult to resolve references correctly and to know how much of the previous constraints are to be kept. Second, in order to understand a context-dependent sentence correctly, we must correctly understand at least two sentences.

The weighted error from speech input is from 8%-10% higher than from text, which is lower than might be expected. Even though the BYBLOS system misrecognized at least one word in 25% of the utterances, the DELPHI system was able to recover from most of these errors through the use of the N-best list and fallback processing.

The SLS weighted error was 30.6%, which represents a substantial improvement in performance over the weighted error during the previous (February '92) evaluation, which was 43.7%. Based on end-to-end tests with real users, the system is usable, given that subjects were able to accomplish their assigned tasks.

6. REAL-TIME IMPLEMENTATION

A real-time demonstration of the entire spoken language system described above has been implemented. The speech recognition was performed using BBN HARKTM, a commercially available product for continuous speech recognition of medium-sized vocabularies (about 1,000 words). HARK stands for High Accuracy Recognition Kit. HARKTM (not to be confused with HARC) has essentially the same recognition accuracy as BYBLOS but can run in real-time entirely in software on a workstation with a built-in A/D converter (e.g., SGI Indigo, SUN Sparc, or HP715) without any additional hardware.

The speech recognition displays an initial answer as soon as the user stops speaking, and a refined (rescored) answer within 1-2 seconds. The natural language system chooses one of the

N-best answers, interprets it, and computes and displays the answers, along with a paraphrase of the query so the user can verify what question the system answered. The total response cycle is typically 3-4 seconds, making the system feel extremely responsive. The error rates for knowledgeable interactive users appears to be much lower than those reported above for naive noninteractive users.

7. SUMMARY

We have described the HARC spoken language understanding system. HARC consists of a modular integration of the BYBLOS speech recognition system with the DELPHI natural language understanding system. The two components are integrated using the N-best paradigm, which is a modular and efficient way to combine multiple knowledge sources at all levels within the system. For the Class A+D subset of the November '92 DARPA test the official BYBLOS speech recognition results were 4.3% word error, the text understanding weighted error was 22.0%, and the speech understanding weighted error was 30.6%.

Finally, the entire system has been implemented to run in real time on a standard workstation without the need for any additional hardware.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract Nos. N00014-91-C-0115, and N00014-92-C-0035.

REFERENCES

- [1] F. Kubala, C. Barry, M. Bates, R. Bobrow, P. Fung, R. Ingria, J. Makhoul, L. Nguyen, R. Schwartz, D. Stallard, "BBN BYBLOS and HARC February 1992 ATIS Benchmark Results", *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Feb. 1992.
- [2] Bobrow R., D. Stallard, "Fragment Processing in the DELPHI System", *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Pub., Feb. 1992.
- [3] Bobrow, R., R. Ingria, and D. Stallard, "Syntactic/Semantic Coupling in the DELPHI System", *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Pub., Feb. 1992.
- [4] MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus", *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Pub., Feb. 1992.
- [5] Chow, Y., M. Dunham, O. Kimball, M. Krasner, G.F. Kubala, J. Makhoul, P. Price, S. Roucos, and R. Schwartz (1987) "BYBLOS: The BBN Continuous Speech Recognition System," *IEEE ICASSP-87*, pp. 89-92.
- [6] Chow, Y.-L. and R.M. Schwartz, "The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses", *ICASSP90*, Albuquerque, NM S2.12, pp. 81-84.
- [7] Schwartz, R., S. Austin, Kubala, F., and J. Makhoul, "New Uses for the N-Best Sentence Hypotheses Within the BYBLOS Speech Recognition System", *ICASSP92*, San Francisco, CA, pp. I.1-1.4.
- [8] Austin, S., Schwartz, R., and P. Placeway, "The Forward-Backward Search Algorithm", *ICASSP91*, Toronto, Canada, pp. 697-700.
- [9] Schwartz, R. and S. Austin, "A Comparison Of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses", *ICASSP91*, Toronto, Canada, pp. 701-704.
- [10] Placeway, P., Schwartz, R., Fung, P., and L. Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora", To be presented at *ICASSP93*, Minneapolis, MN.
- [11] Stallard, D., "Unification-Based Semantic Interpretation in the BBN Spoken Language System", *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Pub., Oct. 1989, pp. 39-46.
- [12] Bobrow, R., "Statistical Agenda Parsing", *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Feb. 1991, pp. 222-224.