A DATA-DRIVEN SPOKEN LANGUAGE UNDERSTANDING SYSTEM

Yulan He *Steve Young*

Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, England
{yh213,sjy}@eng.cam.ac.uk

ABSTRACT

This paper presents a purely data-driven spoken language understanding (SLU) system. It consists of three major components, a speech recognizer, a semantic parser, and a dialog act decoder. A novel feature of the system is that the understanding components are trained directly from data without using explicit semantic grammar rules or fully-annotated corpus data. Despite this, the system is nevertheless able to capture hierarchical structure in user utterances and handle long range dependencies. Experiments have been conducted on the ATIS corpus and 16.1% and 12.6% utterance understanding error rates were obtained for spoken input using the ATIS-3 1993 and 1994 test sets. These results show that our system is comparable to existing SLU systems which rely on either hand-crafted semantic grammar rules or statistical models trained on fully-annotated training corpora but it has greatly reduced build cost.

1. INTRODUCTION

Substantial research has been done in spoken dialogue systems. Among the various spoken dialogue projects, the most influential one is the U.S. DARPA program. From 1990 to 1995, DARPA sponsored a spoken language understanding programme to develop and objectively measure the performance of various Spoken Language Understanding (SLU) systems. Different research sites worked on the same domain, the Air Travel Information Service (ATIS) [1], data for which were collected jointly by them. The utterance understanding error rates for spoken language input in the December 1994 benchmarks range from 6.5% to 44.9% for context-independent utterances (category A).

Work in the early 90's focused on the semantic parser module. The techniques used were either based on context-free semantic rules to extract keywords or phrases to fill slots in semantic frames (template matching), such as MIT's TINA [2], CMU's PHOENIX [3], and SRI's Gemini [4], or based on stochastic models, such as AT&T's Markov model-based CHRONUS [5] and BBN's hierarchical Hidden Understanding Model (HUM) [6]. Both approaches have drawbacks. The former is highly domain-specific and requires heavy manual processing, whilst the latter needs a fully-annotated corpus in order to reliably estimate model parameters.

More recently, the DARPA Communicator project [7] aims to support rapid, cost-effective development of multi-modal speech-enabled dialog systems. Members of the Communicator sites include AT&T, BBN, CMU, University of Colorado (CU), IBM, MIT, MITRE, and SRI. In most of the systems developed by these sites, semantic parsing is still based on the early versions of parse modules, such as the Phoenix parser used by CMU and CU, the TINA Parser used by MIT, and the Gemini parser used by the BBN's Talk'n'Travel system [8]. Only IBM uses a slightly different approach in that it uses a decision-tree based statistical semantic classer and parser for its natural language understanding module [9].

The above systems rely on either semantic grammar rules or statistical models trained on fully-annotated training corpora. Here, we propose a SLU system whose three major components, the speech recognizer, the semantic parser, and the dialog act decoder are all trained directly from data. In particular, it has a hierarchical semantic parser which is able to capture embedded semantic structure in user utterances and which is trained using constrained Expectation-Maximization (EM) directly on unannotated data. The evaluation results on the ATIS corpus using spoken input show that our system is comparable to the original DARPA ATIS SLU systems but with greatly reduced build cost.

The rest of the paper is organized as follows. Section 2 briefly describes the general framework of a statistical SLU system and Section 3 summarizes the training and evaluation procedures used. Section 4 discusses in detail each of the three major components, the speech recognizer, the semantic parser, and the dialog act decoder. The experimental setup and evaluation results are then presented in section 5. Finally, section 6 concludes the paper.

2. SPOKEN LANGUAGE UNDERSTANDING

Spoken language understanding (SLU) can be broadly viewed as a pattern recognition problem. It aims to interpret the meanings of users' utterances and respond reasonably to what users have said. A typical architecture of an SLU system is given in Fig. 1, which consists of a speech recognizer, a semantic parser, and a dialog act decoder. The user's input acoustic signal A is first translated into a word string W by the speech recognizer. Such word strings are then mapped into a set of semantic concepts C by the semantic parser. The dialog act decoder infers the user's intention or goals G_u based on the semantic concepts extracted and the current dialogue context. Finally, the deduced information may be passed to the dialogue manager to decide appropriate actions to take in response to the user's query.

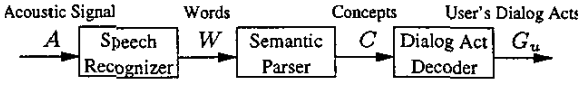

Fig. 1. Typical structure of a spoken language understanding system.

0-7803-7980-2/03/$20.00 © 2003 IEEE

Traditionally, the SLU problem is solved in three stages. First recognize the underlying word string W from each input acoustic signal A, i.e.

$$\hat{W} = \underset{W}{\text{argmax}}\, P(W|A) = \underset{W}{\text{argmax}}\, P(A|W)P(W) \quad (1)$$

then map the recognized word string $\hat{W}$ into a set of semantic concepts C

$$\hat{C} = \underset{C}{\text{argmax}}\, P(C|\hat{W}) \quad (2)$$

and finally determine the user's dialog acts or goals by solving

$$\hat{G}_u = \underset{G_u}{\text{argmax}}\, P(G_u|\hat{C}) \quad (3)$$

In the system described in this paper, each of these stages is modelled separately. we use a standard HTK-based [10] Hidden Markov Model (HMM) recognizer for recognition, the Hidden Vector State (HVS) model for semantic parsing [11], and Tree-Augmented Naive Bayes networks (TAN) [12] for dialog act decoding. Section 4 below describes each of these in more detail.

It should, however, be noted that sequential decoding is sub-optimal in the sense that the solution of each stage depends on the exact solution of the previous stage. In order to reduce the effect of this approximation, it is possible to retain a word lattice or N-best word hypotheses instead of the single best string $\hat{W}$ as the output of the speech recognizer. The semantic parse results may then be incorporated with the output from the speech recognizer to rescore the N-best list since it provides additional knowledge to the recognizer. This is considered further in Section 5. Similarly, it is possible to retain the N-best parse results from the semantic parser and leave the selection of the best hypothesis until the dialog act decoding stage. However, in practice, no gain was found for this and hence we do not pursue it further here.

3. SYSTEM TRAINING AND EVALUATION

Fig. 2 shows the organization of our SLS system for both training and evaluation. The ATIS training data contain the acoustic speech signal, word transcription and reference SQL query for each utterance. Each of the three major components, the speech recognizer, the semantic parser, and the dialog act decoder are trained separately. The acoustic speech signal is modelled by extracting 39 features every 10ms: 12 cepstra, energy, and their first and second derivatives. This data is then used to train the speaker-independent, continuous speech recognizer. The semantic parser is trained using the word transcriptions from the ATIS corpus combined with their abstract semantics extracted automatically from the reference SQL queries provided in the corpus. The parser is trained on this data using constrained EM as described further in Section 4.2. It is straightforward to identify the main topic or goal and the key semantic concepts of each utterance from the corresponding reference SQL query and this information is used to train the dialog act decoder.

During testing, the N-best lists from the speech recognizer are passed to the semantic parser to generate semantic concept sequences. Parse scores from the semantic parser are combined with the total acoustic and language model likelihoods from the speech recognizer and used to rescore the N-best list. Meaningful semantic concept/value pairs are then extracted from the resulting best hypothesis and the user's goals are inferred by the dialog act decoder from the semantic concept sequences generated. These extracted concept/value pairs and inferred goals are then fed into the SQL query generator to form an SQL query in order to fetch answers from the ATIS database.

Performance is measured at both the component and the system level. For the former, the recognizer is evaluated by word error rate, the parser by concept slot retrieval rate using an F-measure metric [13], and the dialog act decoder by detection rate. The overall system performance is measured using the standard NIST "query answer" rate.

4. SYSTEM COMPONENTS

This section discusses the three main components of our SLU system, the speech recognizer, the semantic parser, and the dialog act decoder.

4.1. Speech Recognizer

The speech recognizer was built using the HTK toolkit [10]. It comprises 14 mixture Gaussian HMM state-clustered cross-word triphones augmented by using heteroscedastic linear discriminant analysis (HLDA) [14]. Incremental speaker adaptation based on the maximum likelihood linear regression (MLLR) method [15] was performed during the test with updating being performed in batches of five utterances per speaker.

4.2. Semantic Parser

The semantic parser component was built using the *Hidden Vector State (HVS)* model [11]. The HVS model can be best explained using the example parse tree shown in Fig. 3 where the semantic information relating to each word is completely described by the sequence of semantic concept labels extending from the preterminal node to the root node. If these semantic concept labels are stored as a single vector, then the parse tree can be transformed into a sequence of vector states as shown in the lower portion of Fig. 3. For example, the word Denver is described by the semantic vector [CITY, FROMLOC, SS]. Viewing each vector state as a hidden variable, the whole parse tree can be converted into a first order vector state Markov model. This is the HVS model.

Each vector state is in fact equivalent to a snapshot of the stack in a push-down automaton and state transitions may be factored into a stack shift by n positions followed by a push of one or more new preterminal semantic concepts relating to the next input word. Such stack operations are constrained in order to reduce the state space to a manageable size. Natural constraints to introduce are limiting the maximum stack depth and only allowing one new preterminal semantic concept to be pushed onto the stack for each new input word. Such constraints effectively limit the class of supported languages to be right branching. The joint probability $P(N, \mathbf{C}, W)$ of a series of stack shift operations N, concept vector sequence $\mathbf{C}$, and word sequence W can be decomposed as follows

$$P(N, \mathbf{C}, W) = \prod_{t=1}^{T} P(n_t|W_1^{t-1}, \mathbf{C}_1^{t-1}) \cdot P(c_t[1]|W_1^{t-1}, \mathbf{C}_1^{t-1}, n_t) \cdot P(w_t|W_1^{t-1}, \mathbf{C}_1^{t}) \quad (4)$$

where:

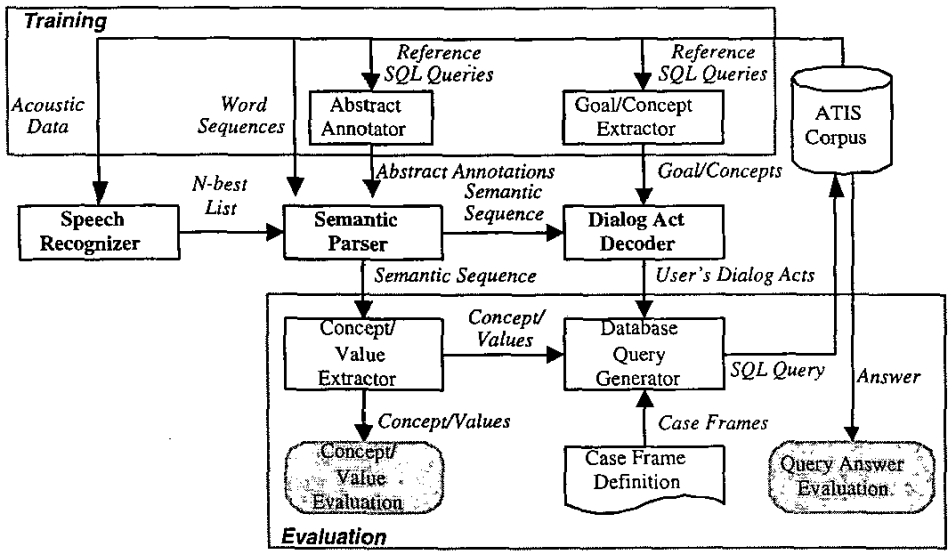

Fig. 2. Procedures on ATIS training and evaluation.

- $\mathbf{C}_1^t$ denotes a sequence of vector states $\mathbf{c}_1..\mathbf{c}_t$. $\mathbf{c}_t$ at word position t is a vector of D_t semantic concept labels (tags), i.e. $\mathbf{c}_t = [c_t[1], c_t[2], .., c_t[D_t]]$ where $c_t[1]$ is the preterminal concept and $c_t[D_t]$ is the root concept (SS in Fig. 3);
- $W_1^{t-1}\mathbf{C}_1^{t-1}$ denotes the previous word-parse up to position $t-1$;
- n_t is the vector stack shift operation and takes values in the range of $0, .., D_{t-1}$ where D_{t-1} is the stack size at word position $t-1$;
- $c_t[1] = c_{w_t}$ is the new preterminal semantic tag assigned to word w_t at word position t.

In the HVS model used by our SLU system, Equation 4 is approximated by

$$P(n_t|W_1^{t-1}, \mathbf{C}_1^{t-1}) \approx P(n_t|\mathbf{c}_{t-1}) \quad (5)$$

$$P(c_t[1]|W_1^{t-1}, \mathbf{C}_1^{t-1}, n_t) \approx P(c_t[1]|c_t[2..D_t]) \quad (6)$$

$$P(w_t|W_1^{t-1}, \mathbf{C}_1^{t}) \approx P(w_t|\mathbf{c}_t) \quad (7)$$

For training, we assume the availability of a set of domain-specific lexical classes and abstract semantic annotations for each utterance. In the case of ATIS, these can be extracted automatically from the relational database and SQL queries of the training utterances. The HVS model is then trained on the unannotated utterances using EM constrained by the lexical class information and the dominance relations built into the abstract annotations [11].

4.3. Dialog Act Decoder

The dialog act decoder was implemented using the Tree-Augmented Naive Bayes (TAN) algorithm [12], which is an extension of Naive Bayes Networks. The basic classifier learns from training data the conditional probability of each semantic concept C_i given the goal G_u, $P(C_i|G_u)$. Classification is done by picking the goal with the highest posterior probability of G_u given the particular instance of concepts $C_1 \cdots C_n$, $P(G_u|C_1 \cdots C_n)$. The strong independence assumption made is that all the concepts C_i are conditionally independent given the value of the goal G_u. TAN networks relax this independence assumption by adding dependencies between concepts. They are however still a restricted family of Bayesian networks in which the goal variable has no parents and each concept has as parent the goal variable and at most one other concept. An example of such a network is given in Fig. 4 where each concept may have one augmenting edge pointing to it. The procedure for learning these edges is based on the well-known Chow-Liu algorithm [16] except that instead of using the mutual information (MI) between two concepts, conditional mutual information (CMI) between concepts given the goal variable is used

$$CMI_{(C_x;C_y|G_u)} = \sum_{C_x,C_y,G_u} P(C_x, C_y, G_u) \cdot \log \frac{P(C_x, C_y|G_u)}{P(C_x|G_u)P(C_y|G_u)} \quad (8)$$

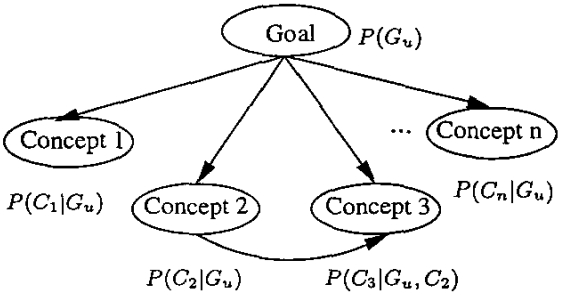

Fig. 4. Example of a Tree-Augmented Naive Bayes Network.

In our dialog act decoder here, one TAN was used for each goal, the semantic concepts which serve as input to its corresponding TAN were selected based on the MI between the goal and the

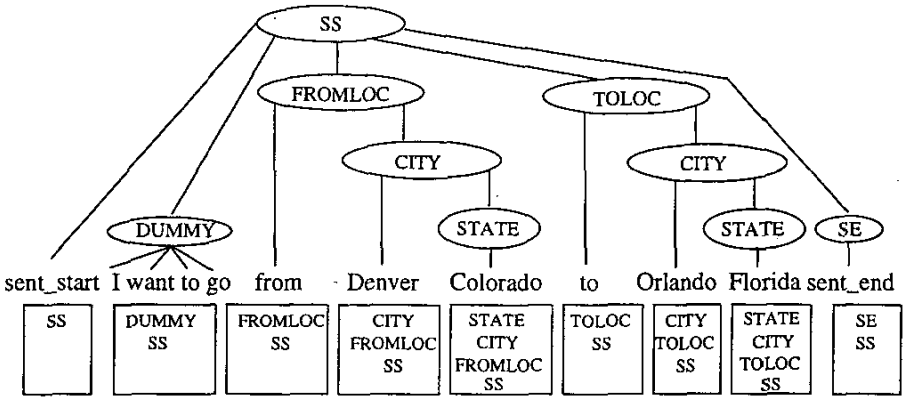

Fig. 3. Example of a Parse Tree and its Vector State Equivalent.

concept. Dependencies between concepts were then added based on the CMI between concepts given the goal.

5. EXPERIMENTS

Experiments have been conducted using the ATIS corpus and the ATIS-3 NOV93 and DEC94 data were selected as test sets. Utterances in the ATIS corpus are divided into three categories, context-independent (A), context-dependent (D), or unanswerable (X). The experimental results reported in this paper focus on category A utterances only unless otherwise specified.

5.1. Experimental Setup

Altogether 22316 spontaneous utterances recorded using Sennheiser microphone from ATIS-2 and ATIS-3 are used for acoustic model training. This includes the ATIS-2 FEB92 and NOV92 test sets in addition to the ATIS-2 and ATIS-3 training sets. The language model was trained on 23096 ATIS spontaneous utterances with vocabulary size 1644. It consists of a word trigram and a word trigram interpolated with a class-based trigram. The latter has 60 classes derived automatically using the Kneser-Ney clustering procedure [17]. The perplexity tested on the joint ATIS-3 NOV93 and DEC94 test sets is 16.5 and 15.5 for the word trigram alone and the interpolated model respectively.

The N-best word hypotheses generated from the speech recognizer were fed into the semantic parser to output semantic concept sequences. Given an acoustic speech signal A, translated into a word sequence W, and parsed into a semantic concept sequence C, the parse scores are combined with the total acoustic and language model likelihoods according to equation 9.

$$\begin{aligned} \hat{C}, \hat{W} &\approx \underset{C,W \in L_N}{\text{argmax}}\ P(A|W)P(W)P(C|W) \\ &\approx \underset{C,W \in L_N}{\text{argmax}}\ P(A|W)P(W)^{\gamma}P(C|W)^{\alpha} \end{aligned} \quad (9)$$

where $P(A|W)$ is the acoustic probability from the first pass, $P(W)$ is the language modelling likelihood, $P(C|W)$ is the semantic parse score, L_N denotes the N-best list, α is a semantic parse scale factor, and γ is a grammar scale factor which was set to 15.0 for the NOV93 test set and 17.0 for the DEC94 test set as determined experimentally.

For the dialog act decoder, 16 dialog acts or goals were defined in the ATIS domain with each goal corresponding to one TAN. The top 15 semantic concepts ranked by MI were used as input to each TAN.

The SQL query generator module was tested on the reference parse results of ATIS-3 NOV93 and DEC94 test sets. 5 out of 448 utterances from NOV93 test set and 3 out of 445 utterances from DEC94 test set did not return the correct answers, which gives the utterance understanding error rate 1.1% and 0.7% respectively. The analysis of the results shows that one context-dependent utterance has been misclassified as category A (context-independent) in each of these two test sets and the rest are too complicated for the SQL query generator to handle properly.

5.2. Experimental Results

Experiments were first conducted to evaluate individual components of the SLU system. Table 1 gives the results in word error rate (WER) for the speech recognizer by imposing different refinement techniques on the full test sets (A+D+X). The baseline was built using a word bigram language model (LM), then the HMM models were refined based on the HLDA technique. Subsequently, incremental adaptation test was performed and bigram word lattices were generated, which were then expanded to word trigram lattices by applying the word trigram LM. Finally, the class-based trigram LM was used to transform word bigram lattices to class-based trigram lattices.

Criteria	*NOV93*	*DEC94*
word bigram	7.3	6.0
+HLDA	6.8	5.4
+adaptation test	5.7	4.1
+word trigram	4.8	3.6
+class-based trigram	4.8	3.4

Table 1. Test results for the speech recognizer (%WER).

The semantic parser was tested using both text input (reference transcriptions) and spoken input (recognizer output). The F-measure scores together with recall and precision values are reported in Table 2.

For the dialog act decoder, the goal detection accuracy based

Measurement	NOV93		DEC94	
	Text	*Spoken*	*Text*	*Spoken*
Recall	89.2%	87.6%	91.3%	89.7%
Precision	91.4%	90.4%	92.6%	91.4%
F-measure	90.3%	89.0%	91.9%	90.5%

Table 2. Test results for the semantic parser.

on the parse results of both text input and spoken input is shown in Table 3.

Parser Input	*NOV93*	*DEC94*
Text input	91.7%	91.2%
Spoken input	91.5%	90.8%

Table 3. Test results for the dialog act decoder.

During the integrated system test, experiments were first conducted to determine the best possible performance in WER obtainable from the N-best lists output by the speech recognizer. This was done by picking the hypothesis with the lowest WER from each list for N ranging from 1 to 1000. As the system gave the same performance when N is beyond 25, only the results with values of N ranging from 1 to 25 are reported in Fig. 5. It can be observed that $N = 10$ gives the optimal WER and subsequent experiments were therefore conducted on 10-best lists only. Increasing the value of N degrades the system performance slightly. This is due to noise introduced by the lower ranks of N-best lists. The oracle WER of different N-best lists are also given to indicate the range of improvements possible by incorporating more knowledge sources.

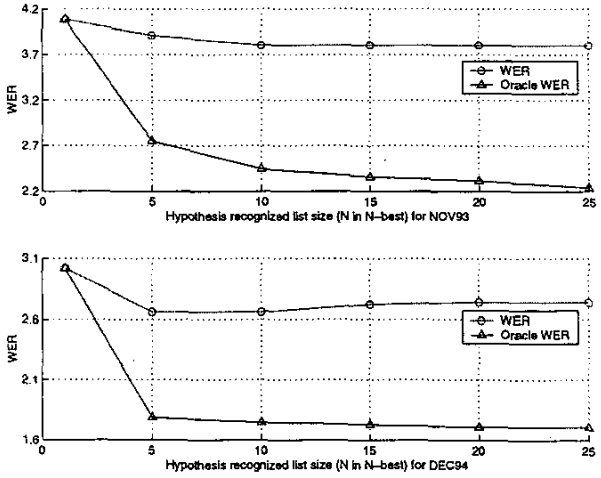

Fig. 5. Values of N (as in N-best list) vs WER.

Fig. 6 shows the WER obtained for rescored 10-best word hypotheses when the semantic parse scale factor α as defined in Equation 9 is varied. The optimal value for α is 10 as the lowest WER is obtained at this point for both NOV93 and DEC94 test sets. Increasing α value degrades the system performance since the semantic parse scores tend to dominate the rescored results.

The end-to-end evaluation results on both natural language understanding (NL) and spoken language understanding (SLS) evaluations are shown in Table 4. F-measure evaluates the extraction of concept/value pairs in terms of recall and precision, while answer error rate measures the minimum / maximum answers from the ATIS database using the NIST scoring package. The latter is the standard scoring metric used by DARPA ATIS SLU systems. For the NL test, the semantic parser used as input the reference transcriptions instead of the recognized output. The SLS(1) results were obtained by taking the best word hypothesis directly from the speech recognizer, while the SLS(10) results were obtained by taking the best word hypothesis from the rescored 10-best list after incorporating semantic parse scores.

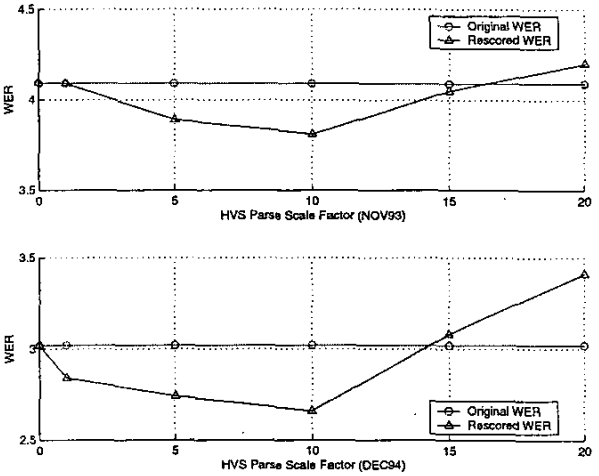

Fig. 6. Scale of semantic parse score vs WER.

	NOV93		DEC94	
	F-measure	*Answer Error*	*F-measure*	*Answer Error*
NL	90.3%	12.3%	91.9%	8.5%
SLS(1)	89.0%	18.3%	90.5%	13.9%
SLS(10)	89.3%	16.1%	90.6%	12.6%

Table 4. NOV93 and DEC94 NL and SLS test results.

6. CONCLUSION

This paper has discussed a purely data-driven spoken language understanding system. Its three major components, the speech recognizer, the semantic parser, and the dialog act decoder, are trained directly from corpus data. In particular, its two understanding components, the semantic parser and the dialog act decoder, are trained without the use of explicit semantic grammar rules or fully-annotated treebank style data.

The evaluation results on the ATIS corpus show that our SLU system is comparable to the original DARPA ATIS SLU systems which relied on either hand-crafted semantic grammar rules or fully-annotated training corpora to extract semantic information, but it can be built at much lower cost. We have also confirmed, as others have done [18, 19, 20, 21], that semantic knowledge extracted by a parser can be applied to rescore N-best word hypotheses from

the speech recognizer to improve both WER and overall end-to-end performance.

7. REFERENCES

[1] P. Price, "Evaluation of spoken language systems: the ATIS domain," in *Proc. of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, June 1990, pp. 91–95, Morgan Kaufman Publishers, Inc.

[2] S. Seneff, "Robust parsing for spoken language systems," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, San Francisco, 1992.

[3] W. Ward and S. Issar, "Recent improvements in the CMU spoken language understanding system," in *Proc. of the ARPA Human Language Technology Workshop*. 1996, pp. 213–216, Morgan Kaufman Publishers, Inc.

[4] J. Dowding, R. Moore, F. Andry, and D. Moran, "Interleaving syntax and semantics in an efficient bottom-up parser," in *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Maxico, June 1994, pp. 110–116.

[5] E. Levin and R. Pieraccini, "CHRONUS, the next generation," in *Proc. of the DARPA Speech and Natural Language Workshop*, Austin, TX, Jan. 1995, pp. 269–271, Morgan Kaufman Publishers, Inc.

[6] S. Miller, M. Bates, R. Bobrow, R. Ingria, J. Makhoul, and R. Schwartz, "Recent progress in hidden understanding models," in *Proc. of the DARPA Speech and Natural Language Workshop*, Austin, TX, Jan. 1995, pp. 276–280, Morgan Kaufman Publishers, Inc.

[7] MITRE, *DARPA Communicator homepage*, URL: http://fofoca.mitre.org/, 2003.

[8] D. Stallard, "Evaluation results for the Talk'n'Travel system," in *Human Language Technology Conference*, San Diego, California, Mar. 2001.

[9] T. Ward, "How long until a high school student can build a language understanding system," in *Proc. of Intl. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000.

[10] HTK, *Hidden Markov Model Toolkit (HTK) 3.2*, Cambridge University Engineering Department, URL: http://htk.eng.cam.ac.uk/, 2002.

[11] Yulan He and Steve Young, "Hidden vector state model for hierarchical semantic parsing," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, Hong Kong, Apr. 2003.

[12] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2, pp. 131–163, 1997.

[13] V. Goel and W. Byrne, "Task dependent loss functions in speech recognition: Application to named entity extraction," in *ESCA ETRW Workshop on Accessing Information from Spoken Audio*, Cambridge, UK, 1999, pp. 49–53.

[14] N. Kumar, "Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition," *PhD Thesis*, 1997, Johns Hopkins University, Baltimore MD.

[15] M.J. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, Oct. 1996.

[16] C.K. Chow and C.N. Liu, "Approximating discrete probability distributions with dependence tree," *IEEE Trans. on Information Theory*, vol. 14, pp. 462–467, 1968.

[17] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modelling," in *Proceedings of the European Conference on Speech Communication and Technology*, 1993, pp. 973–976.

[18] M. Rayner, D. Carter, V. Digalakis, and P. Price, "Combining knowdge sources to recoder N-best speech hypothesis lists," in *Proceedings of the ARPA Human Language Technology Meeting*, 1994.

[19] R. Moore, D. Appelt, J. Dowding, J.M. Gawron, and D. Moran, "Combining linguistic and statistical knowledge sources in natural-language processing for ATIS," in *ARPA Spoken Language Technology Workshop*, 1995.

[20] K. Hacioglu and W. Ward, "Combining language models: Oracle approach," in *Human Language Technology Conference*, San Diego, CA, Mar. 2001.

[21] A. Chotimongkol and A.I. Rudnicky, "N-best speech hypotheses reordering using linear regression," in *Proc. of Eurospeech*, Aalborg, Denmark, 2001, pp. 1829–1832.