The MIT ATIS System: February 1992 Progress Report[1]

Victor Zue, James Glass, David Goddeau, David Goodine, Lynette Hirschman, Michael Phillips, Joseph Polifroni, and Stephanie Seneff

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This paper describes the status of the MIT ATIS system as of February 1992, focusing especially on the changes made to the SUMMIT recognizer. These include context-dependent phonetic modelling, the use of a bigram language model in conjunction with a probabilistic LR parser, and refinements made to the lexicon. Together with the use of a larger training set, these modifications combined to reduce the speech recognition word and sentence error rates by a factor of 2.5 and 1.6, respectively, on the October '91 test set. The weighted error for the entire spoken language system on the same test set is 49.3%. Similar results were also obtained on the February '92 benchmark evaluation.

INTRODUCTION

This paper presents an update on the MIT ATIS system, which has been under development since 1990. We will describe several changes made to our system since the last official common evaluation in February, 1991 [8], with particular emphasis on the speech recognition component. We will also present our evaluation results for the October '91 "dry-run" test set and the February '92 test set. We have also modified our natural language component to include a robust parsing strategy. This change is described in detail in a companion paper [9].

SPEECH RECOGNITION

In this section we will describe the changes we have made over the past year to the speech recognition component (SUMMIT) of our ATIS system. These include improvements to both the phonetic and language models, and refinements on the lexicon. We have also implemented the acoustic models on a set of DSP boards to allow near real-time evaluation and demonstration.

The baseline SUMMIT system uses a mixture of up to 16 diagonal Gaussian models for each lexical unit. In recent months, we have been able to simplify the input representation of the models significantly with no loss in performance. The current representation consists of 39 segmental measurements for each hypothesized segment. This vector is rotated via principal component analysis prior to mixture Gaussian modelling. Segment duration is modelled separately, in the log domain, using a mixture of Gaussians. At the moment, spontaneous disfluencies are represented by one model, and are required to be one segment long.

Training and Testing Corpora

The multi-site ATIS data collection effort has resulted in a significant increase in the amount of speech data available to the community [6]. For speech recognition system development, we started with all the MADCOW data released by NIST, and augmented them with ATIS data collected earlier at MIT. Some 9,711 utterances in this pool were designated as training material, and an additional 1,595 utterances were set aside as a development set for independent evaluation.

To facilitate a meaningful comparison, all the experiments described in this section are performed on the October '91 "dry-run" test set, containing some 362 utterances collected at BBN, CMU, MIT, and SRI. The experiments that we conducted are summarized in Table 1, and will be described in this section.

In order to monitor progress internally, we also ran the same test set through our system as reported a year ago [8]. Our February '91 system had a vocabulary of 577 words. That system constrained the N-best search with the use of a word-pair grammar with a perplexity of 92. The N-best outputs were subsequently resorted using our natural language component TINA. It was trained on some 2400 utterances collected at TI and MIT. The recognition performance of that system on the October '91 "dry-run" test set, with and without the word-pair language model, is shown in the first two rows of Table 1 (labelled as AW and WP, respectively).

Lexicon

With the availability of a larger amount of training data we enlarged our vocabulary to contain 841 words. This was done by examining word frequency counts in

[1]This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

the training data and adding all reasonable words that occurred more than once. Examples of words that were not added included misspellings or people's names.

Other improvements to the lexicon included refinement of the pronunciation baseforms and the phonological rules used to generate the pronunciation networks. In part, this involved improving pre-existing rules such as the flapping rule. We also introduced a number of specific allophones for certain phonemes in certain contexts, such as a retroflexed /f/ or a stop closure following a fricative, and a number of new diphone units, allowing a sequence of two phonemes to be treated as a diphthong, such as /**ɛl**/ or /**ar**/. The inventory of phonetic units in the expanded lexicon contained 115 distinct labels.

As shown in the third row of Table 1 (labelled as AW, Small Training), these changes combined to reduce the word error rate from 62.5% to 55.4% for the system a year ago using an all-word language model. The next row in the same figure (labelled as AW, Full Training) shows that the word error rate is further reduced to 51% by using the full training set described earlier[2]. This result is identical to the results of the February '91 system using a *word-pair* language model, although the latter achieved better sentence recognition accuracy. Unless otherwise specified, the remaining experiments described in this section all use the full training set.

Bigram Language Model

The current SUMMIT system uses significantly more language constraints than were used by its predecessor [8]. With the help of the available large training set, we constructed a smoothed bigram grammar. As has been done elsewhere, the bigram was smoothed by interpolating the bigram estimates with the prior probabilities of each word [2,4]:

$$p(\omega_b|\omega_a) = \lambda(\omega_a)\hat{p}(\omega_b|\omega_a) + (1 - \lambda(\omega_a))\hat{p}(\omega_b)$$

where

$$\hat{p}(\omega_b|\omega_a) \equiv \frac{N(\omega_a, \omega_b)}{N(\omega_a)}$$

$$\hat{p}(\omega_b) \equiv \frac{N(\omega_b)}{N(all\ \ words)}$$

The interpolation weights were set to vary with the number of times we had observed the conditioning context:

$$\lambda(\omega) \equiv \frac{N(\omega_a)}{N(\omega_a) + K}$$

where K is a single constant that was optimized so as to minimize the measured perplexity on the development data set. For the ATIS training data, we found that the perplexity had a broad minimum when K was around 20. On our development data set this smoothed bigram had a perplexity of 20.1. The perplexity measures did not include out of vocabulary words since our recognition system does not currently have the capability of detecting these words. Including out-of-vocabulary words in the perplexity measure increased the value slightly to 20.8.

Recognition results using the bigram language model are shown in row 5 of Table 1 (labelled as BG). The bigram language model is the single most effective change we made to our system, reducing the word-error rate by more than twofold from the best results obtained previously.

Probabilistic LR Parser

A probabilistic LR parser was used in addition to a bigram model to provide language constraints. The LR algorithm is a deterministic, table-driven, left-to-right parsing algorithm for a subset of context-free grammars [1]. The probabilistic LR (PLR) model extends this algorithm to assign a probability

$$P(w_0...w_n) = \prod_{i=0}^{n} P(w_i|w_0...w_{i-1})$$

to each word string, (rather than a binary value). In the PLR model the conditional word probabilities are approximated using the parser state.

If $P(Q_j|w_0...w_{i-1})$ is the probability that the parser is in state Q_j having just parsed the substring $w_0...w_{i-1}$ (without making any moves based on the value of w_i), then the conditional word probability can be re-written as:

$$P(w_i|w_0...w_{i-1}) = \sum_j P(w_i Q_j|w_0...w_{i-1}).$$

Making the assumption that the parser state captures much of the information in the substring $w_0...w_{i-1}$ relevant to the conditional probabilities, this can be approximated by:

$$P(w_i|w_0...w_{i-1}) \approx \sum_j P(w_i|Q_j)P(Q_j|w_0...w_{i-1}).$$

The set of Q_j for which $P(Q_j|w_0...w_{i-1})$ is non-zero is determined by the grammar. In particular, if the grammar is deterministic, then $P(Q_j|w_0...w_{i-1}) = 1$, for some $j = j_i$, and

$$P(w_0...w_n) = \prod_i P(w_i|w_0...w_{i-1}) \approx \prod_i P(w_i|Q_{j_i}).$$

The probabilities $P(w_i|Q_j)$ can be estimated from a corpus of training utterances using the ratio of the number of times w_i is the next word when the parser is in state Q_j to the number of times the parser is in state Q_j.

[2]Due to computational limitations, we did not use the entire designated training set for training. Instead, a subset of about 7,500 utterances were used.

System	Characteristics	Sub (%)	Del (%)	Ins (%)	Wd. Error (%)	Sent. Error(%)
Feb '91	AW	41.7	10.5	10.4	62.5	98.3
	WP	33.6	10.0	7.4	51.0	93.9
Feb '92	AW, Small Training	39.7	7.5	8.2	55.4	97.8
	AW, Full Training	37.0	7.3	6.7	51.0	98.1
Feb '92	BG	15.3	5.4	3.5	24.1	72.7
	BG+PLR	13.2	6.3	2.5	22.0	66.9
	BG+CD	12.4	5.5	2.7	20.6	67.7
	BG+CD+PLR	11.7	5.3	2.3	19.3	61.6
	BG+CD+PLR+NL	11.6	5.1	2.1	18.8	58.6

Table 1: Speech recognition results on the October '91 test set for the various experiments described in this paper. In addition to the word and sentence error rates, errors due to substitution, insertion and deletion are also provided. Performance of the systems from a year ago on the same data set is included for reference. The symbols are: AW=all-word language model, WP=word-pair language model, BG=bigram language model, CD=context-dependent modelling, PLR=probabilistic LR parser, NL=NL filtering using TINA.

In previous work using the PLR model for the VOYAGER task [3], the language model implemented was strict, that is, it assigned probability 0 to word strings not generated by the input grammar. In order to apply this model to speech recognition (i.e., optimizing word accuracy), the parse table was extended to "accept" all word strings. This was accomplished by adding explicit error states to the parse table, and computing recovery actions to allow normal parsing to resume in an appropriate state after an error[3]. Other extensions to the model described previously [3] include various mechanisms for smoothing the probabilities by changing the conditioning state.

The ATIS grammar contains 971 rules, the vast majority of which introduce lexical items, and the resulting parse table contains about 1600 states. The lexicon of the parser is the same as that used by the recognizer. The probabilities were trained on all 9,711 utterances in the training set. The perplexity measured on the October '91 test set was 17.6.

Row 6 of Table 1 (labelled as BG+PLR) shows that further reduction in error rate is possible by incorporating the PLR. PLR is incorporated by using the parse score in place of the bigram score to reorder the 50 N-best outputs produced by the recognizer. The sentence error rate is reduced more than the word error rate, presumably due to the fact that PLR can deal with some of the long distance constraints better than the bigram.

Context-Dependent Modelling

At the last DARPA meeting we first described our work towards accounting for contextual effects on the phonetic modelling component of SUMMIT [5]. We proposed using regression tree analysis to find the contextual factors that provided the greatest reduction in the distortion of our phonetic models. In an initial experiment, regression tree analysis was used to form a set of context-specific models for each phonetic unit. However, we found that we were able to obtain the best performance by using the regression trees to independently learn a context-normalization factor for each of the input dimensions of the model. The model for each phonetic unit is then trained using these context-normalized inputs for all of the training samples in that class.

We have extended this work by considering more contextual effects, including phonetic labels two phones away and whether or not the current segment is in a syllable before a pause or at a sentence boundary. The new effects were simply added to the list of questions that could be asked at each node in the tree splitting algorithm.

When we applied this context-normalization to the ATIS domain, we found that the word error rate dropped from 24.1% to 20.6%, as shown in rows 5 and 7 in Table 1) (labelled as BG and BG+CD, respectively). This represents a 15% reduction in error rate. In the Resource Management domain, we found a decrease in word error rate from 10.3% to 7%, or 32% [5]. We believe that we are achieving a smaller reduction in error rate in the ATIS domain because a greater number of errors can be attributed to problems other than phonetic modelling (e.g., out-of-vocabulary words, mismatch of language model, spontaneous speech effects, etc.). In fact, if we look at the performance of the phonetic models in terms of their ability to match the "forced-recognition" phonetic string (the string obtained during recognition allowing only the correct word string), we see a much larger reduction in error rate in the ATIS domain (37.5%) than in the Resource Management domain (18.8%). This may not be surprising, since we are now considering more contextual effects. In addition, it is likely that there are stronger

[3]This is roughly equivalent to parsing the word string as a sequence of fragments rather than as a complete sentence.

Input	Correct	Incorrect	No Answer	Error
Text	87.7(%)	8.5(%)	3.9(%)	20.9(%)
Speech	64.8(%)	14.1(%)	21.1(%)	49.3(%)

Table 2: Overall system performance, for both text and speech input, on the October '91 test set.

contextual effects in a spontaneous speech corpus such as ATIS than in a more carefully spoken "read" corpus such as Resource Management.

The combined effect of our improved phonetic and language modelling is shown in row 8 of Table 1 (labelled as BG+CD+PLR). In this case, the PLR score is used in conjunction with the acoustic score to resort the N-best outputs. As expected, there is again a more significant improvement on the sentence error rate.

Finally, we incorporated our natural language system TINA as a filter on the N-best outputs produced by the recognizer (with $N = 40$), and the results are shown in the last row of Table 1 (labelled as BG+CD+PLR+NL). Not surprisingly, the natural language component is able to reduced the sentence error rate much more than the word error rate.

OTHER IMPROVEMENTS

The most significant improvement in the back-end, the augmentation of the system with a robust parsing capability is described separately. However, in addition, we have continued to expand the capabilities of the back-end at all levels (syntactic coverage, concepts understood, discourse modelling, dialogue aspects, etc.) We continue to improve the level of sophistication of the booking dialogue, towards the goal of a natural and effective mixed-initiative dialogue to achieve a successful booking.

The performance of our current spoken language system on the October '91 test set is summarized in Table 2. The significant improvement in our NL result can be attributed to the robust parsing strategy that we have adopted. Discussion of these results can be found in a companion paper [9].

FEBRUARY '92 BENCHMARK

The February '92 benchmark results were obtained by running the official test set released by NIST through our system once. This test set contains 971 utterances collected AT&T, BBN, CMU, MIT, and SRI. The speech recognition results are shown in Table 3. Comparing Table 3 with the last row of Table 1, we see that the performance of our system on the two test sets is quite similar.

Sub (%)	Del (%)	Ins (%)	Wd. Error (%)	Sent. Error (%)
11.5	4.4	2.3	18.1	59.6

Table 3: Speech recognition results for the February '92 test set.

Input	Correct	Incorrect	No Answer	Error
Text	80(%)	13(%)	7(%)	32.5(%)
Speech	61(%)	14(%)	25(%)	52.8(%)

Table 4: Overall system performance, for both text and speech input, on the February '92 test set.

The performance of our current spoken language system on the February '92 test set is summarized in Table 4. Although the system's performance for speech input is similar to that on the October '91 test set, the NL results are not as good. This is a direct reflection of our research priority since October 1991. That is, we have focused our group's attention almost entirely on improving the speech recognition component, to the neglect of expanding our NL system capabilities to adequately conform to the principles of interpretation. Again, discussion of these results can be found elsewhere in these proceedings [9].

SUMMARY AND FUTURE WORK

This paper describes the improvements that we have made to the recognition component of our ATIS system. By incorporating more language constraints (using a bigram and a probabilistic LR parser) and performing context dependent phonetic modelling, a significant reduction in recognition error rates is realized. This has led to a corresponding decrease in weighted error of the overall spoken language system. Much of the phonetic recognition parts of our system has been ported to a set of off-the-shelf DSP boards. The complete system, using an IBM RS6000 for lexical access and a Sun SPARCstation-II for the rest of the processing, now runs in 2-3 times real-time.

In the coming months, we plan to conduct research in several directions that will hopefully lead to further improvement in system performance. These areas include the introduction of gender-specific acoustic models, modelling out-of-vocabulary words, modelling, spontaneous speech effects such as pauses, increasing the size of the lexicon and training set size, and better language models.

Our results show that better language modelling is crucial to improved performance. Our future research

in this area falls in several categories. In addition to developing a bigram grammar, we have begun to explore the use of class bigram's as well as more general N-grams. The class bigrams we examined grouped similar words together in order to reduce the number of unseen word pairs. We investigated both grouping the conditioning context into classes:

$$p(\omega_b|\omega_a) \approx \hat{p}(\omega_b|C(\omega_a))$$

as well as the word itself:

$$p(\omega_b|\omega_a) \approx \hat{p}(\omega_b|C(\omega_b))\hat{p}(C(\omega_b)|\omega_a)$$

where $C(\omega_b)$ is the general class of words that ω_b belongs to. We explored a number of different classes and found that we could reduce the development set perplexity by a small amount, to 19.5.

We have also begun to explore the use of more general N-grams and class N-grams. The N-gram language model store *all* word sequences observed in the training data. In order to represent these grammars efficiently we store them in the form of a hierarchical tree, where each node deeper in the tree represents one word farther back in the past. Smoothing becomes extremely important for the N-gram. Thus far we have used the generalization of the bigram interpolation procedure so that N-gram smoothing is done recursively:

$$p_i = \lambda_i \hat{p}(\omega_n|\omega_{n-i}\ldots\omega_{n-1}) + (1-\lambda_i)p_{i-1}$$

$$\lambda_i = \lambda(\omega_{n-i}\ldots\omega_{n-1})$$

$$p_0 = \hat{p}(\omega_n)$$

so that

$$p(\omega_n|\omega_0\ldots\omega_{n-1}) = p_n$$

Our initial experiments suggest that, by incorporating a class 4-gram directly into the N-best search, we can reduce our sentence error rate from 62.5% to 56.3% (for $N = 1$) on the development set, although there is a corresponding increase in the amount of search required. We have also found that by simply adding the class 4-gram scores into our N-best resorting algorithm we can reduce our sentence error rate from 59.6% to 56.0%on the February '92 test set.

The PLR model implemented for the ATIS system and the N-gram models described above are both instances of a larger class of language models based on Stack Automata. A Probabilistic Stack Automata language model approximates conditional word probabilities as:

$$P(w_i|w_0...w_{i-1}) \approx P(w_i|\text{stack after parsing } w_0...w_{i-1})$$

This model can also be considered as an extension to a class N-gram, in which the class members can be phrases as well as words.

Future work in language modelling will focus on application of this general model. In particular, conditioning the probabilities on the entire parse stack rather than on the current state (essentially the top of stack) should further reduce perplexity and bring long-distance constraints to bear.

ACKNOWLEDGEMENTS

We would like to acknowledge the assistance of Hong C. Leung, a former member of our group, whose has contributed directly and indirectly to several aspects of our spoken language system development effort.

REFERENCES

[1] Aho, A. and Ullman, J., *The Theory of Parsing, Translation, and Compiling*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[2] S. Austin, P. Peterson, P. Placeway, R. Schwartz, J. Vandergrift, "Toward a Real-Time Spoken Language System Using Commercial Hardware," *Proc. DARPA Speech and Natural Language Workshop:* 72-77 June 1990.

[3] Goddeau, D. and Zue, V., "Integrating Probabilistic LR Parsing into Speech Understanding Systems," *Proc. ICASSP 92,* March, 1992.

[4] Katz, S., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP,* ASSP-35, 3, 400-401, March, 1987.

[5] Phillips, M., Glass, J., and Zue, V. "Modelling Context Dependency in Acoustic-Phonetic and Lexical Representations," *Proc. DARPA Speech and Natural Language Workshop*: 71-76, February 1991.

[6] MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," *These Proceedings.*

[7] Polifroni, J. Seneff, S., and Zue, V., "Collection of Spontaneous Speech for the sc Atis Domain and Comparative Analyses of Data Collected at MIT and TI," *Proc. DARPA Speech and Natural Language Workshop*: 360-365, February 1991.

[8] Seneff, S., Glass, J., Goddeau, D., Goodine, D., Hirschman, L., Leung, H., Phillips, M., Polifroni, J. and Zue, V., "Development and Preliminary Evaluation of the MIT ATIS System," *Proc. DARPA Speech and Natural Language Workshop*: 88-93, February 1991.

[9] Seneff, S., "A Relaxation Method for Understanding Spontaneous Speech Utterances," *These Proceedings.*