1

Combining Linguistic and Statistical Knowledge Sources in Natural-Language Processing for ATIS*

Robert Moore, Douglas Appelt, John Dowding, J. Mark Gawron, and Douglas Moran

Artificial Intelligence Center
SRI International
Menlo Park, California 94025

ABSTRACT

During the past year, significant improvements have been made in the natural-language processing technology used in the SRI ATIS spoken-language understanding system. The principal developments have been (1) the incorporation of information from the natural-language grammar and lexicon into a statistical language model that is used in both recognition and understanding, (2) implementation of a robust interpretation component that constructs queries out of grammatical fragments when an utterance cannot be analyzed as a single phrase or utterance, and (3) a new context mechanism for air travel planning that constructs an explicit model of the user's intended itinerary.

1. INTRODUCTION

The principal natural-language component of SRI's ATIS system is Gemini [1, 2], a unification-based natural-language processing system. The Gemini system combines a set of general syntactic and semantic rules for English with a domain-specific lexicon and sortal/selectional restrictions, to produce sorted logical forms for natural-language utterances. The output is derived by a bottom-up parser/interpreter with limited top-down prediction that integrates syntax and semantics on a phrase-by-phrase basis and incorporates strategies for detecting and correcting verbal repairs. The resulting logical forms are fed through a set of simplification rules that produce simplified logical forms intended to map more easily onto the ATIS database. If Gemini fails to produce a simplified logical form that can be translated to a database query, the system falls back to the Template Matcher [3], a key-word-and-phrase-based pattern-matching module. In addition to performing utterance interpretation, Gemini is used to rescore the output of the DECIPHER^{TM} ATIS speech recognition system [4] to improve recognition accuracy.

2. USING NATURAL-LANGUAGE CONSTRAINTS IN RECOGNITION

2.1. Our Previous Approach

A long-standing goal of research in spoken-language understanding has been to find an effective way to use a natural-language understanding system as a knowledge source in speech recognition. Achieving significant improvements in fair tests of recognition performance has proved difficult, at least in part because of lack of robustness when the natural-language system is unable to completely analyze an utterance. A year ago [5] we reported on a new approach based on the observation that, even when our grammar fails to provide a complete analysis of an utterance, it is usually possible to find a small number of semantically meaningful phrases that span the utterance. We used the Gemini parser/interpreter to find the minimal number of semantically meaningful phrases needed to span a recognition hypothesis and computed a natural-language score for the hypothesis based on this number. The Gemini score was computed as a somewhat ad hoc combination of the number of phrases needed to cover the hypothesis, a bonus if the hypothesis could be analyzed as a single sentence (as opposed to any other single grammatical phrase), and penalties for using grammar rules dispreferred by the parse preference component. This score was then scaled by an empirically optimized parameter and added to the recognition score in an N-best rescoring pass.

In the December 1993 ATIS SPREC benchmark evaluation, incorporation of this knowledge source reduced both word and utterance recognition error for the A+D+X test set by about 5%. (The word error rate went from 5.4% to 5.2%, and the utterance error rate went from 27.5% to 26.0%.) These improvements, while modest, were measured by NIST to be statistically significant at the 95% confidence level according to the matched-pair sentence segment (word error) test and the McNemar (sentence error) test.

2.2. Our Current Approach

Analysis of the December 1993 ATIS SPREC evaluation and our observation of other recent work in language modeling led us to conjecture that a closer integration of linguistic and statistical factors might lead to better results, and we have developed a new form of language model that represents a move in that direction. In our current language model, we use Gemini to analyze a recognition hypothesis as a sequence of semantically meaningful fragments (much as in our previous model), but then we use n-gram statistics to estimate the probability of the hypothesis under that analysis. The natural-language score is based on the logarithm of this probability estimate.

The resulting language model is a kind of multi-level n-gram model. The top level is a trigram model of the probability of a hypothesis as a sequence of types of fragments. That is, we model the probability of an utterance being a sequence of, for

*This research was supported by the Advanced Research Projects Agency under Contract N00014–93–C–0142 with the Office of Naval Research and Contract N66001–94–C–6046 with the Naval Command, Control, and Ocean Surveillance Center.

instance, a sentence followed by a modifier phrase followed by a nominal phrase. The fragment types we use are sentence, nominal phrase, modifier phrase, filler (e.g., "please"), and "skipped". This last category consists of the sequences of words that are left over after we have determined the best coverage of the utterance in terms of well-formed phrases of semantic classes that can be mapped into ATIS database entities. We chose a trigram model for this level in order to model well the important case of an utterance consisting of the "begin-utterance" token, a single fragment, and the "end-utterance" token.

The next level of the model is a word and word-class four-gram model of each fragment. We initially estimated separate models for each type of fragment, but we found that this suffered from splitting the training data into smaller pools. The method of modeling fragments we settled on is to have a single four-gram model, but to treat each fragment as a sequence starting with a token such as "begin_sentence" or "begin_nominal_phrase". The result is that the probability of the first few words of each fragment is conditioned on what type of fragment it is, but once we have several words of context, we base our probability estimate on that rather than on the type of fragment. The four-gram model is smoothed by linear combination with lower-order models, the weights being estimated by deleted interpolation.

For the word-class models, the classes are generated semi-automatically from the Gemini lexicon. For each noun and noun-like syntactic category, we define classes consisting of all items in the Gemini lexicon having the same syntactic and semantic features and semantic class, ignoring certain features that play little or no role in the distributional probabilities of the lexical items. Multi-token combinations that are entered in the Gemini lexicon as fixed items (e.g., "d c ten") are treated as single words. For the version of the lexicon used to create the model for the December 1994 evaluation, there are 131 word classes having more than one member, ranging in size from two members to several dozen. If an item has more than one entry in the lexicon, or if an entry is assigned more than one semantic class, the item can be a member of more than one word class. Any lexical item having a unique combination of features and semantic class (and any non-nominal lexical item) is in effect treated as a singleton class. The probability distributions for most classes across their members are estimated from training data. Because most of the training data is ATIS2 data, while the test data is ATIS3 data, we assumed uniform distributions for the classes we thought most likely to be affected by the change from the ATIS2 database to the ATIS3 database and by the dates of data collection. The classes for which we assumed uniform distributions were names of months, cities, states, city and state combinations, and airports.

Another feature of the model is the way it integrates information from the Gemini repair component. Sections of hypotheses that are analyzed by Gemini as being deleted by verbal repairs are indicated by the markers "begin_repair" and "end_repair". The word sequence model incorporates estimates of the probability of beginning or not beginning a repair at every point in the sequence where a repair is not in progress, and of ending or not ending a repair at every point where a repair is in progress. In addition, the estimate of the probability of the sequence following the end of the repair is conditioned on the context at the point the deleted material begins. That is, given the sequence "i want to fly begin_repair to san end_repair to los angeles", the probability estimate for "to los angeles" is conditioned on the context "i want to fly" rather than the context "i want to fly to san."[1]

All the probability estimates described above were combined into a simple joint probability estimate for a hypothesis under an analysis as a sequence of fragments. Since the Gemini-based model incorporates a four-gram word and word-class model, we initially conjectured that it could simply replace the statistical language model used in the baseline DECIPHER ATIS system. We discovered in development testing, however, that the Gemini-based model performed less well alone than the baseline language model, but both models together seemed to perform better than the baseline model alone. We also discovered that adding word-insertion and fragment-insertion penalties helped performance on development test material, so we incorporated those in our final model.

2.3. SPREC Test Results

Final December 1994 ATIS SPREC benchmark evaluation results for SRI (word error and utterance error for the class A+D+X and class A+D subsets) are given in Table 1. As the table shows, improvements in recognition by rescoring DECIPHER output with Gemini ranged from 13.3% to 15.3%. These differences were measured to be statistically significant at the 95% confidence level on all four metrics used by NIST.

	A+D+X Word Error	A+D+X Utt. Error
Baseline DECIPHER	2.5%	15.5%
DECIPHER+Gemini	2.1%	13.1%
Improvement	14.6%	15.1%

	A+D Word Error	A+D Utt. Error
Baseline DECIPHER	2.3%	15.2%
DECIPHER+Gemini	1.9%	12.8%
Improvement	13.3%	15.3%

Table 1: December 1994 final ATIS SPREC test results.

The improvements due to rescoring DECIPHER output with Gemini were measured to be statistically significant at the 95% confidence level on all four metrics used by NIST.

We believe that it is important that this improvement in recognition performance by adding a natural-language-based knowledge source was achieved relative to a state-of-the-art

[1] A bug in the version of the system used in the evaluation prevented this from working exactly as intended. The principal effect of the bug was that sequences following the end of repairs were not conditioned on any prior context. This does not seem to have had any significant effect in the overall performance of the system, however.

baseline recognizer. Repeatedly in the past, improvements have been obtained with natural-language-based knowledge sources in recognition, only to have the improvements disappear as baseline recognition performance has improved. The results reported here represent the only significant improvement we are aware of obtained by using a natural-language-based knowledge source in conjunction with a current state-of-the-art recognizer.

3. ROBUST INTERPRETATION

For high performance in query understanding, it is necessary to extract an interpretation from an utterance, even if it is not possible to find a complete linguistic analysis of the utterance. In previous years, SRI's ATIS system relied solely on the Template Matcher [3] for this robust interpretation capability in cases where Gemini failed to find a complete analysis. This had the disadvantage that, since the Template Matcher is a completely separate system from Gemini, the performance of the Template Matcher could not benefit from improvements in Gemini's coverage of phrases that could be used by the Template Matcher. To remedy this problem we have now implemented, within Gemini, a robust interpretation module that uses the linguistic analyses produced by Gemini for grammatical utterance fragments, heuristically combining them into an overall interpretation.

When Gemini is unable to find a complete analysis of an entire utterance, we first find Gemini's best analysis of the utterance as a sequence of fragments. The fragments we look for are filler words, whole sentences, noun phrases of semantic classes that can be mapped into database entities, and modifier phrases that can be interpreted as expressing restrictions on database entities. All other words in the utterance are skipped. Given this set of desired phrase types, we carry out a Viterbi search through the chart of possible phrases produced by the Gemini parser/interpreter to find the analysis that skips the fewest words and, for a given number of skipped words, has the smallest number of fragments. In general, there may be more than one analysis that is optimal by these criteria, and in that case we apply the statistical model described in Section 2.2, choosing the analysis with the highest probability according to that model. For instance, the utterance

> leave boston eastern flight one forty three at twelve forty

can be analyzed by Gemini as two fragments two different ways:

> [leave] [boston eastern flight one forty three at twelve forty]
> [leave boston] [eastern flight one forty three at twelve forty]

The statistical model assigns a higher probability to the second analysis, which is not only the one that intuitively seems correct, but is also much easier to generate a correct database query from. (With the first bracketing, it is more difficult to tell that Boston is the intended origin rather than destination.) We have not yet had a chance to directly measure the performance of the statistical model in choosing among possible bracketings, but anecdotally it seems to be quite accurate.

Once a single analysis is chosen, the semantic interpretations of the fragments are combined into a single interpretation for the whole utterance. In many cases, this is very simple. For instance, if one phrase denotes a flight, and another phrase is a flight modifier, we can assume that the flight modifier is modifying the flight referred to by the first phrase. In other cases, the relationships between the semantic classes found in the phrases need to be inferred, and new entities possibly introduced, to construct an interpretation. In these cases, we use a "database graph" that connects all the types of entities in the database according to their most salient relations, to find a minimal set of relations to connect the interpretations of the phrases we are trying to combine. For instance, for the two fragments in the utterance

> [what types of aircraft] [can i get a first class ticket from philadelphia to dallas]

a flight variable must be introduced into the query even though no flight is mentioned in the utterance, because flights are needed to relate aircraft to first class tickets.

4. CONTEXT MODELING

Our current ATIS system incorporates a context-handling mechanism that is improved in a number of respects over that in previous versions of the system. The context mechanism attempts to determine whether the current scenario involves travel planning, or the user is merely seeking information without a particular travel itinerary in mind. If the system determines that the user is engaged in travel planning, it attempts to recognize his travel itinerary, and it associates constraints with each leg of the itinerary. When a new query is processed, the system attempts to identify which leg of the itinerary is being discussed, or if the itinerary needs to be revised or extended, and then retrieves the set of constraints associated with that branch of the itinerary for possible incorporation into the current query. This explicit representation of an itinerary is useful in recognizing resumptions of previous dialog segments, and in identifying implicitly characterized origin/destination pairs, such as in "the return flights".

The new context mechanism also provides the means to refer to sets of entities either implicitly, in terms of the set of constraints characterizing the set, or explicitly, by directly referring to the actual set of entities produced as an answer to a previous query. When it is clear that the user has a particular set of entities in mind (e.g., he uses a demonstrative such as "those flights") the system identifies the context in which the intended set of entities was generated, and then directly retrieves the entities mentioned in that context.

5. NL AND SLS TEST RESULTS

Final December 1994 ATIS benchmark NL and SLS test results for SRI (unweighted utterance understanding error for class A+D, class A only, and class D only) are given in Table 2. SLS results are given for two systems; "SLS 1-best"

refers to the results of feeding the 1-best output of the baseline DECIPHER recognizer to the natural-language understanding system, and "SLS N-best" refers to the results of feeding the Gemini-rescored DECIPHER output to the understanding system. For class A+D, the NL result is a 41% improvement over the previous year, the SLS 1-best result is a 34% improvement over the previous year, and the SLS N-best result is a 41% improvement over the previous year.

	A+D	A	D
NL	10.7%	7.0%	16.4%
SLS 1-best	13.7%	10.6%	18.5%
SLS N-best	12.7%	9.7%	17.4%

Table 2: December 1994 final ATIS NL and SLS test results.

6. FUTURE DIRECTIONS

Of the work described here, we plan to extend that dealing with robust interpretation and the use of natural-language-based knowledge sources in recognition. In the former area, our system needs more work so as always to be able to generate a database query from the results of fragment combining. In principle, fragment combining should always result in a valid database query, but for a small percentage of utterances, the system still fails to generate a query and falls back on the Template Matcher. We also need to give the fragment-combining module other options than simply using the database graph to find the minimal set of relations, when reference is made to things like "return flights".

In the area of natural-language-based knowledge sources for recognition, we want to see if we can do a better job of choosing parameters for combining our current language model with the recognizer score. This is motivated by postmortem analysis of preliminary December 1994 SPREC results, which showed that the choice of parameters optimal for this particular test set would have improved performance from 14.6% better than the baseline recognizer to 19.0% better, for A+D+X word error. We also plan to undertake a more complete integration of statistical modeling with the Gemini system, incorporating probability estimates for lexical choices, syntactic and semantic rule choices, and semantic classes in predicate-argument combinations.

Finally, the results we have achieved using natural-language-based knowledge sources in recognition make it appear worthwhile to look once more at alternatives to N-best rescoring as a search architecture for speech and natural-language integration. Prior to the results reported here, it had generally not been shown to be useful to consider more than about five recognition hypotheses for natural-language rescoring, which suggested that the overhead of a more complex search architecture than N-best rescoring was unlikely to be repaid. In the system we used in the December 1994 benchmark tests, however, we corrected recognition errors by selecting recognition hypotheses ranked as low as 22, and with retrospectively optimized knowledge source weights, as low as 63. Since it appears that with this type of model it is useful to consider not just a handful of recognition hypotheses, but several dozen, we believe we should examine other search architectures for speech and natural-language integration that will be more efficient as a result of sharing analyses of word sequences common to several hypotheses. We have carried out some preliminary design work for such an approach based on SRI's progressive search architecture [6], which we intend to continue pursuing.

References

1. J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A Natural Language System for Spoken-Language Understanding," in *Proceedings ARPA Workshop on Human Language Technology*, Merrill Lynch Conference Center, Princeton, New Jersey, 21–24 March 1993, pp. 43–48 (Morgan Kaufmann Publishers, Inc., San Francisco, California).
2. J. Dowding, R. Moore, F. Andry, and D. Moran, "Interleaving Syntax and Semantics in an Efficient Bottom-Up Parser," in *Proceedings 32nd Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, New Mexico, 27 June – 1 July 1994, pp. 110–116.
3. E. Jackson, D. Appelt, J. Bear, R. Moore, and A. Podlozny, "A Template Matcher for Robust NL Interpretation," in *Proceedings Speech and Natural Language Workshop*, Pacific Grove, California, 19–22 February 1991, pp. 190–194 (Morgan Kaufmann Publishers, Inc., San Francisco, California).
4. M. Cohen, Z. Rivlin, and H. Bratt, "Speech Recognition in the ATIS Domain Using Multiple Knowledge Sources," to appear in *Proceedings ARPA Workshop on Spoken Language Technology*, Barton Creek Resort Conference Center, Austin, Texas, 22–25 January 1995.
5. R. Moore, M. Cohen, V. Abrash, D. Appelt, H. Bratt, J. Butzberger, L. Cherny, J. Dowding, H. Franco, J. M. Gawron, and D. Moran, "SRI's Recent Progress on the ATIS Task," to appear in *Proceedings ARPA Spoken Language Systems Technology Workshop*, Merrill Lynch Conference Center, Princeton, New Jersey, 6–8 March 1994.
6. H. Murveit, J. Butzburger, V. Digalakis, and M. Weintraub, "Progressive-Search Algorithms for Large-Vocabulary Speech Recognition," in *Proceedings ARPA Workshop on Human Language Technology*, Merrill Lynch Conference Center, Princeton, New Jersey, 21–24 March 1993, pp. 87–90 (Morgan Kaufmann Publishers, Inc., San Francisco, California).