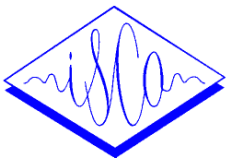

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

# STOCHASTIC REPRESENTATION OF SEMANTIC STRUCTURE FOR SPEECH UNDERSTANDING

Roberto Pieraccini Esther Levin

Speech Research Department
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974, USA

## Abstract

We propose a model for a statistical representation of the conceptual structure of a restricted subset of spoken natural language. The model is used for segmenting a sentence into phrases and labeling them with concept relations (or cases). The model is trained using a corpus of annotated transcribed sentences. An understanding system is being built around this model, allowing for unconstrained spoken input in a database retrieval task. The results on a test set of 148 sentences show that almost 97% of cases were correctly assigned.



## 1 Introduction

The application of the understanding system reported in this paper refers to the ATIS (Air Travel Information System) task [2] that was proposed by DARPA about two years ago. The ATIS task is built around a relational database, a subset of the Official Airline Guide (OAG), that includes the information about connections between 10 American cities. A number of spoken utterances within the domain were collected through a *wizard* [5] system and distributed to the sites participating in the project. Only the sentences classified as *context independent*, or class A sentences (i.e. where the meaning of the sentence does not depend upon previous sentences), are considered as legal inputs to the system described in this paper. Examples of class A sentences are:

*SHOW ME ALL THE NONSTOP FLIGHTS FROM DALLAS TO DENVER LEAVING ON APRIL TWENTY SECOND.*

*WILL DINNER BE SERVED ON AMERICAN FLIGHT 986 DEPARTING FROM SAN FRANCISCO AT NINE FORTY FIVE.*

*WHAT TRANSPORTATION IS AVAILABLE FROM THE DENVER AIRPORT TO BOULDER ON APRIL 22ND.*

Fig. 1 shows a block diagram of the proposed understanding system. The input can be either speech or text. The functions of the various blocks are described by the following example. Assume that the input (written or spoken) sentence is the first example above. The task of the *case decoder* is to provide a representation of the sentence in terms of a *conceptual segmentation*. Concepts (or *cases*) are the smallest units of meaning that are relevant to the task. The conceptual segmentation for the sample sentence is:

| | |
|---|---|
| **QUERY:** | *SHOW ME ALL* |
| **A_STOP:** | *THE NONSTOP* |
| **OBJECT:** | *FLIGHTS* |
| **ORIGIN:** | *FROM DALLAS* |
| **DESTINATION:** | *TO DENVER* |
| **DATE:** | *LEAVING ON APRIL TWENTY SECOND* |

The case called **QUERY** is associated with the part of the sentence that expresses the question, **OBJECT** is the object of the question, **ORIGIN** and **DESTINATION** refer to the origin airport and destination airport of the flight, **DATE** is the effective date of the flight, and **A_STOP** is an attribute of the query object, related to the number of stops of the flight. The sentence is segmented into cases through a stochastic modeling of the conceptual structure of the task. The second step, called *case to attribute mapping*, consists of converting this representation, where the cases are filled with English phrases extracted from the sentence, into a representation that is closer to the way data is encoded into the database.

| | |
|---|---|
| **QUERY:** | `LIST` |
| **OBJECT:** | `flight.airline, flight number, flight.departure_time` |
| **STOPS:** | `0` |
| **FROM_AIRPORT:** | `DFW` |
| **TO_AIRPORT:** | `DEN` |
| **DAY_NAME:** | `SUNDAY` |

The attribute **QUERY** is given the value `LIST`, the action required by the sentence. Since the object of the sentence is *flights*, the information that has to be given to the user is the airline (`flight.airline`), the flight number (`flight.flight_number`) and the departure time

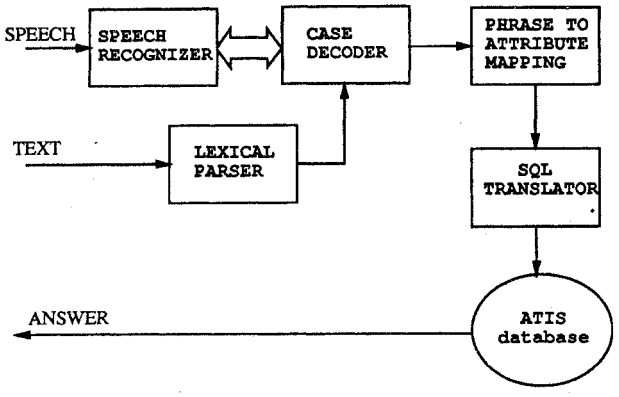


Figure 1: Block diagram of the proposed understanding system

(`flight.departure_time`). The word *NONSTOP* is translated into the value `0` for the attribute **STOPS**, and the words *DALLAS* and *DENVER* into the corresponding airport acronyms `DFW` and `DEN`. Finally, the phrase *LEAVING ON APRIL TWENTY SECOND* is translated into the corresponding day of the week (`SUNDAY`) needed for retrieving the appropriate flights. This module is currently implemented by a pattern-matching procedure that spots keywords in the phrases associated with the various cases, and yields the appropriate attribute value according to the possible values stored in the database. A final step translates the latter representation into the SQL query:

```
SELECT DISTINCT flight.airline,
flight.flight_code, flight.departure_time FROM
flight WHERE ( flight.from_airport='DFW')
AND (flight.to_airport='DEN') AND
flight.stops=0 AND (flight.flight_days IN (SE-
LECT flight.day_mask FROM flight_day WHERE
flight_day.day_name='SUNDAY'))
```

This paper focuses on the implementation of the first block, namely the *speech-to-case* decoder. The formulation of the case segmentation problem is described in section 2, section 3 gives details about the actual implementation, and section 4 reports experimental results.

## 2 The CHRONUS model

CHRONUS stands for Conceptual Hidden Representation of Natural Unconstrained Speech. The basic idea of this model consists in defining the speech decoding task in the following terms. An utterance, consisting of a sequence of acoustic observations,

$$\mathbf{A} = a_1, a_2 \ldots a_N, \tag{1}$$

corresponds to a sequence of words

$$\mathbf{W} = w_1, w_2 \ldots w_M. \tag{2}$$

Each word can be associated to a *case* label; hence the utterance also corresponds to the sequence of case labels:

$$\mathbf{C} = c_1, c_2 \ldots c_M, \tag{3}$$

We are interested in finding the sequence of words $\tilde{\mathbf{W}}$ and the sequence of cases $\tilde{\mathbf{C}}$ that maximizes the conditional probability

$$P(\tilde{\mathbf{W}}, \tilde{\mathbf{C}} \mid \mathbf{A}) = \max_{\mathbf{W} \times \mathbf{C}} \mathbf{P}(\mathbf{W}, \mathbf{C} \mid \mathbf{A}), \tag{4}$$

according to the maximum a posteriori decoding criterion. This conditional probability can be written using, the Bayes inversion formula, as:

$$P(\mathbf{W}, \mathbf{C} \mid \mathbf{A}) = \frac{\mathbf{P(A \mid W, C)P(W \mid C)P(C)}}{\mathbf{P(A)}}. \tag{5}$$

The denominator term in Eq. 5, being a constant, can be disregarded in the maximization. The expression to maximise consists then of three terms: the acoustic model of words $P(\mathbf{A} \mid \mathbf{W}, \mathbf{C})$ that can be reasonably assumed independent from the concepts, hence substituted for $P(\mathbf{A} \mid \mathbf{W})$, the concept-conditional language model $P(\mathbf{W} \mid \mathbf{C})$, and the conceptual model $P(\mathbf{C})$. The acoustic model of words can be implemented in the usual way by means of hidden Markov models of phonetic units. The language and conceptual model can be represented as:

$$P(\mathbf{W} \mid \mathbf{C})P(\mathbf{C}) = \tag{6}$$
$$\prod_{i=2}^{M} P(w_i \mid w_{i-1} \ldots w_1, \mathbf{C}) P(w_1 \mid \mathbf{C})$$
$$\prod_{i=2}^{M} P(c_i \mid c_{i-1} \ldots c_1) P(c_1).$$

Assuming that:

$$P(w_i \mid w_{i-1} \ldots w_1, \mathbf{C}) = \tag{7}$$
$$P(w_i \mid w_{i-1} \ldots w_{i-n}, c_i),$$

and

$$P(c_i \mid c_{i-1} \ldots c_1) = \tag{8}$$
$$P(c_i \mid c_{i-1} \ldots c_{i-m}).$$

we can represent the language/conceptual model as an HMM where the hidden states represent the concepts and the observations represent the words. Eq. 7 represents the observation probability as a state local $(n+1)$-gram language model, while Eq. 8 represents the transition probabilities of a $m$-th order Markov process. If $n = m = 1$, the

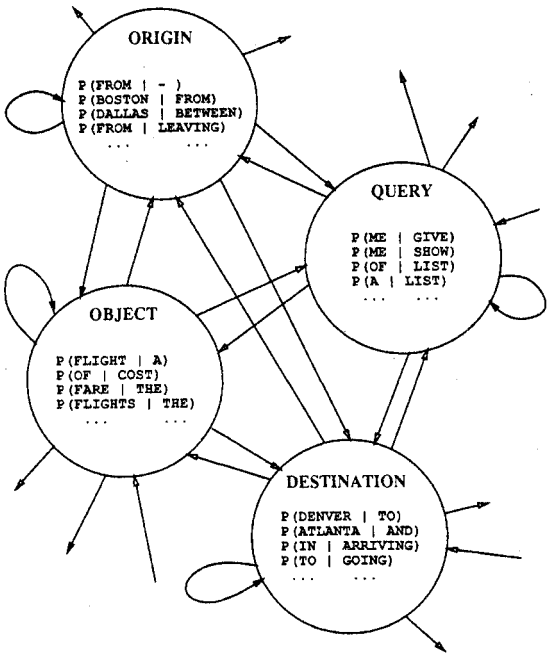


Figure 2: Language/conceptual model as an HMM

language/conceptual model can be represented as in Fig. 2. Viterbi decoding is used to perform the segmentation of a sentence into cases. More details on the CHRONUS model for the ATIS task can be found in [1]. The model, made up of 47 states, was trained using a set of 547 training sentences, represented in a textual format that were initially hand-segmented into cases. The test on the official June 1990 and February 1991 DARPA test sets gave high segmentation accuracy (95.0% and 94.1% cases were correctly segmented and labeled in the two test sets respectively).

## 2.1 Speech Recognition

At the time this paper is being written, the case decoder accepts only a textual input. Of course, the system was designed for understanding spoken language. We intend the integration with speech , according to Eq. 5, to follow the maximum a posteriori criterion. The idea is to represent each word in each concept as an acoustic hidden Markov model, based on phonetic subword units [4], so that an input utterance can be decoded in terms of words and concepts using the Viterbi algorithm.

# 3 The Super-Lexicon

With speech input, the format of a sentence can look rather different from a textual format (the format of the speech reference transcriptions provided in the DARPA ATIS project is called SNOR). This is mainly due to the fact that acronyms and numbers are generally expressed as a single word in a textual input (e.g. DC10, 747) but can be expressed with multiple words and in different ways in spoken language (e.g. D C TEN or D C ONE ZERO or D C ONE OW, SEVEN FOUR SEVEN or SEVEN HUNDRED AND FORTY SEVEN, etc.). An official vocabulary of 1065 words was defined for the ATIS task. This vocabulary includes alphabetic characters for spelling acronyms and words for pronouncing natural and ordinal numbers. Based on the official lexicon, we designed a *super-lexicon*, consisting of 764 items. Each item of the super-lexicon (a *super-word*) can be:

a. a word, like ABOUT, MONTH, RETURN, etc.

b. a word with optional morphological inflections, like AIRFARE(S), DAY(S), ADVANCE(D), etc.

c. a grammar, represented by a finite state automaton; for instance, the grammar for natural numbers (e.g. THIRTY SEVEN), the grammar for airport acronyms (e.g. D F W ), the grammar for compound words (e.g. SAN FRANCISCO).

The local bigram languages of the CHRONUS model are computed with reference to the super-words rather than the words of the lexicon. This reduces the number of parameters to be estimated and increases the robustness of the system, by giving the same probability to all the words in the same category. Words that are out of the vocabulary are associated to an UNKNOWN word class that is given a fixed small bigram probability for all concepts.

## 3.1 Smoothing of bigram probabilities

The major problem in the case segmentation of the ATIS sentences is that the training set for estimating the parameters of the CHRONUS model is, at the moment, very small. The 547 class A sentences do not provide a statistically significative sample for estimating the transition and the bigram probabilities. While we assume a floor value for the unobserved transitions between cases, we use a supervised method for smoothing the case-conditional word bigrams. The supervised smoothing relies on the knowledge that for a given concept there are several words that can be assumed to carry the same meaning. For instance, for the concept **ORIGIN**, the words

DEPART(S) LEAVE(S) ARRIVE(S)

can be considered as synonyms, and can be interchanged in sentences such as:

*THE FLIGHT THAT DEPART(S) FROM DALLAS*
*THE FLIGHT THAT LEAVE(S) FROM DALLAS*
*THE FLIGHT THAT ARRIVE(S) FROM DALLAS.*

A number of groups of synonyms were manually detected for each concept. The occurence frequencies inside a group were equally shared among the constituting words, giving the same bigram probability for synonymous words.

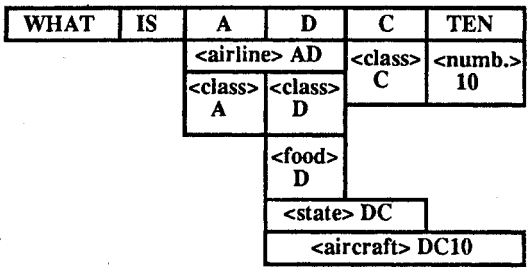


Figure 3: Example of lattice generated by the lexical parser

### 3.2 Lexical parser

The presence of acronyms, numbers, and compound words in the text makes the interpretation of a sentence in the SNOR format ambiguous at the lexical level. For instance the sentence:

WHAT IS A D C TEN

could be interpreted, at the lexical level, in any of the following ways:

*What is a DC10*
*What is A D C 10*
*What is A DC 10*

For a correct interpretation of the sentence, the case decoding must take into account all the possible lexical interpretations. Hence, given an input SNOR sentence, a lattice is generated which includes all the possible interpretations. An example of a lattice, for the previous sentence, is shown in Fig. 3. The parsed words are tagged with the name of the corresponding grammar in the super-lexicon. For instance, the word DC belongs to the *state* grammar, the word DC10 belongs to the *aircraft* grammar, the word AD to the *airline* grammar. The case Viterbi decoding must be modified in order to find the best interpretation, according to the concept/language model, on a lattice of word hypotheses rather than on a string of words. The modified Viterbi algorithm works on a three-dimensional grid [3] rather than on a bi-dimensional one and finds the best sequence of states along with the best sequence of contiguous lattice items.

## 4 Results

An accurate evaluation of the system, using the standard measure proposed for the ATIS project, will be possible only when all the modules of Fig. 1 are completed. At the moment, we can evaluate the accuracy of the case decoder only by comparing a manual segmentation of the sentences with the segmentation proposed by the system, and counting how often a case label is correctly assigned to a phrase. For the February 91 test set, consisting of 148 sentences, the first version of the case decoder [1] (starting from a textual input with acronyms and numbers transcribed by hand) gave 94.1% of correctly assigned cases, whereas 80.4 of the sentences did not show any segmentation error. With the new implementation (starting from textual SNOR format, and including the lexical parser), 96.8% of the cases were correctly labeled and 88.5% of the sentences did not show any error. The increase in the performance is due to the grouping of words into grammars (super-lexicon) and to the supervised smoothing of bigrams.

## 5 Conclusions

In this paper we propose a model for segmenting a sentence into semantic cases defined within a database retrieval task. The segmentation is performed according to the maximum a posteriori criterion, and the model can be integrated with existing speech recognizers. Although the training set size is not appropriate for the estimation of the model parameters, the system performed accurately on a set of 148 test sentences where more than 96% of the cases were correctly segmented and labeled.

## References

[1] Pieraccini, R., Levin, E., Lee, C. H., "Stochastic Representation of Conceptual Structure in the ATIS Task," *Proc. of 4th DARPA Workshop on Speech and Natural Language*, Asilomar (CA), February 1991.

[2] Price, P. J., "Evaluation of Spoken Language Systems: the ATIS Domain," *Proc. of 3rd DARPA Workshop on Speech and Natural Language*, pp. 91-95, Hidden Valley (PA), June 1990.

[3] Fissore, L., Laface, P., Micca, G., Pieraccini, R., "Lexical Access to Large Vocabularies for Speech Recognition," *IEEE Trans. on ASSP*, Vol. 37, No.8, pp. 1197-1213, August 1989.

[4] Lee, C. H., Rabiner, L. R., Pieraccini, R., Wilpon, J. G., "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language*, No. 4, pp. 127-165, 1990

[5] Hemphill, C. T., Godfrey, J. J., Doddington, G. R., "The ATIS Spoken Language System Pilot Corpus," *Proc. of 3rd DARPA Workshop on Speech and Natural Language*, pp. 96-101, Hidden Valley (PA), June 1990.