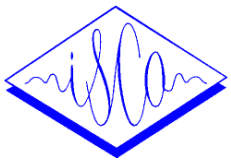

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A Spoken Language System For Information Retrieval

S.K. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L. Lamel, W. Minker

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{bennacef,hbm,gauvain,lamel,minker}@limsi.fr

ABSTRACT

Spoken language systems aim to provide a natural interface between humans and computers by using simple and natural dialogues to enable the user to access stored information. The LIMSI spoken language work is being pursued in several task domains. In this paper we present a system for vocal access to database for a French version of the Air Travel Information Services (ATIS) task. The ATIS task is a designated common task for data collection and evaluation within the ARPA Speech and Natural Language program. A complete spoken language system including a speech recognizer, a natural language component, a database query generator and a natural language response generator is described. The speaker independent continuous speech recognizer makes use of task-independent acoustic models trained on the BREF corpus and a task-specific language model. A case-frame approach is used for the natural language component. This component determines the meaning of the query and builds an appropriate semantic frame representation. The semantic frame is used to generate a database request to the ddatabase management system and the returned information is used to generate a response. First evaluation results for the ATIS task are given for the recognition and understanding components, as well as for the combined system.

INTRODUCTION

Our work is aimed at developing a natural interface between humans and computers by using simple and natural dialogues to enable the user to access stored information. Since our goal is to develop system components that are as task-independent as possible, this work is oriented toward several application domains. This paper presents our spoken language system for information retrieval in one of the chosen applications, a French ATIS system. The ATIS (Air Travel Information Services) task is a designated common task for data collection and evaluation within the ARPA Speech and Natural Language program. An ATIS system allows the user to acquire information derived from the Official Airline Guide about fares and flight schedules available between a restricted set of cities within the United States and Canada. Other information such as the meals served on the flight or the type of aircraft, is also available.

Another application is for access to rail travel information such as timetables, tickets and reservations. A spoken language system for such a service kiosk is under development at LIMSI as part of the ESPRIT project MASK using the same architecture as the one described in this paper.

The work on the French version of the ATIS task was initialized thanks to a collaboration with the MIT-LCS Spoken Language Systems Group. The natural language (NL) component of the MIT ATIS system[1] was ported to French, which enabled us to collect data with a Wizard of Oz (WOZ) setup[2]. We have since developed a natural language component based on a case-frame analysis.

An overview of the spoken language information retrieval system is given in Figure 1. The main components are the speech recognizer, the natural language component which includes a semantic analyzer and a dialog manager, and a response generator that also handles database access. While any spoken language system will necessarily have some dependence of the chose task, our goal is to develop underlying technology that is as task- and language-independent as possible. Thus, in this figure the task and language dependencies are explicitly denoted for the data. The spoken request is recognized by a speaker-independent, continuous speech recognizer, whose ouput is then passed to the natural language component. While the acoustic models used by the recognizer are task-independent, but language dependent, the language models are both task and language-dependent. The semantic analyzer carries out a case-frame analysis to determine the meaning of the query, and builds an appropriate semantic frame representation. The dialog history is used to complete the semantic frame when needed. The cases specifying the case-frame grammar, and the trigger keywords are also to a large part task- (or at least domain) dependent. The response generator uses the semantic frame to generate a database request to the ddatabase management system (DBMS), and presents the result of the database query and an accompanying natural language response to the user. The information access clearly relies on task-dependent data, and the response generation makes use of task and language-dependent databases.

In the remainder of this paper, we describe the major system components, in particular the speech recognizer and the NL component, and give experimental results for the ATIS task.

SPEECH RECOGNITION

The speech recognizer is a large vocabulary, speaker-independent, continuous speech recognizer[3, 4]. It makes use of continuous density HMM (CDHMM) with Gaussian mixture for acoustic modeling and a bigram-backoff language model. The acoustic analysis results in a 48-component feature vector computed every 10 ms. This

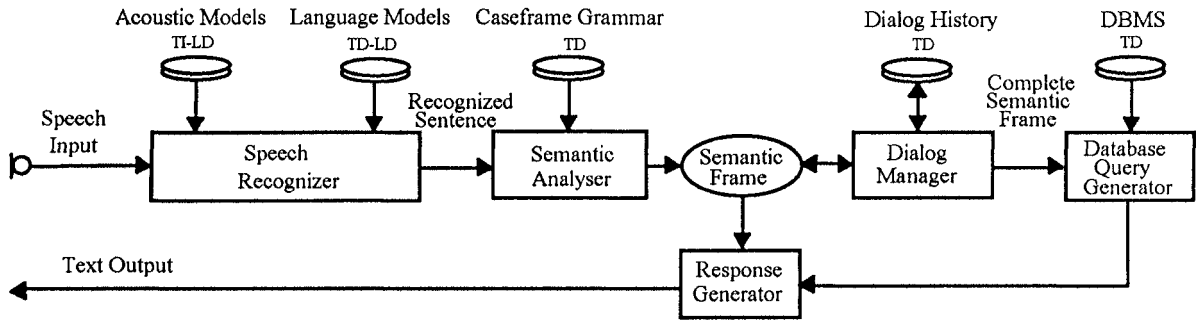

Figure 1: Overview of the spoken language information retrieval system. The system components are task-independent (TI) and language-independent (LI). The databases may be task-dependent (TD) and/or language-dependent (LD).

feature vector contains of 16 Bark-frequency scale cepstrum coefficients computed on the 8kHz bandwidth and their first and second order derivatives. The acoustic models are sets of speaker-independent, context-dependent (but position independent) phone models. The contexts are automatically selected based on their frequencies in the training data. The phone models include triphone, right- and left-context phone models, and context-independent phone models. Each phone model is a left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components). The covariance matrices of all the Gaussians are diagonal.

Since one of the goals of our speech recognition research is to develop acoustic models that are independent of the speaker and the vocabulary, none of the ATIS speech data has been used to train the acoustic models. The acoustic models, which have been trained on 15,200 sentences from 80 speakers taken from the BREF corpus of read newspaper text[5], are the same as are used for our research in large-vocabulary, speaker-independent dictation[4].

The recognizer uses a time-synchronous graph-search strategy[6] which includes the intra- and inter-word context-dependent phone models, phonological rules, and a bigram language model[3]. The HMM-based word recognizer graph is built by putting together word models according to the grammar in one large HMM. Each word model is obtained by concatenation of the phone models for each word, according to its phone transcription as found in the lexicon. The ATIS lexicon contains 730 words, and is represented phonemically with a set of 35 symbols including silence. The bigram language model was estimated on a total 1953 queries, including 505 typed, 793 translated from English, 655 spoken queries. Since the amount of LM training data is small, some grammatical classes (such as cities, days, months, etc) were used to provide more robust estimates of the bigram probabilities.

UNDERSTANDING COMPONENT

A case-frame approach is used to extract the meaning of a sentence. The original linguistic concept of a case-frame as described by Fillmore [7] is based on a set of universally applicable cases which show the relationship between a verb and its nouns. We use the terminology defined by Bruce[8], ie. a case is a relationship which holds between a predicate (usually but not necessarily the verb) and one of its arguments. A case marker is a surface structure indicator (preposition, case affix) of the case. A case frame of a predicate is a set of cases which is related to that predicate. A case system is a complete set of cases for the language. The case-frame approach appears very suitable for speech understanding systems where the need for semantic guidance in parsing is especially relevant. Case-frame grammars have already been successfully applied in several systems[10, 11].

The idea behind our understanding procedure does not require verifying the correct syntactic structure of the sentence, but rather is to extract its meaning using syntax as a constraint only. Therefore, in our system the predicate of the case frame is realized as a concept and not as a verb and the arguments are the constraints of that concept. In the request *"Je voudrais savoir le type d'avion qui va de Denver à Boston le 14 juillet à 14 heures et qui arrive vers 22 heures. (I would like to know the type of aircraft that goes from Denver to Boston on July 14th at 14 hours and arrives around 22 hours.)"* the predicate is the concept type and the constraints (cases) are departure city, arrival city, date and time.

Identification of the concept categories[9] is quite an important (and difficult) job which is obviously task dependent but hopefully language independent. In order to extract the possible case-frame categories of the ATIS task and their cases, a set of queries from the training corpus were manually analyzed to augment our a priori task knowledge. Five categories have been identified and are given in the form of concepts in Table 1. We have chosen to merge in a unique case frame concepts related to requests for time and flight information, because the information returned in response to these types of queries is the same.

A set of 54 cases are used to represent the different types of information in all the case frames. These cases may be classed according to the types of information: flight-designation, city, date, time and fare. Case markers are used to provide syntactic constraints which are necessary to extract the meaning of the request. In "**de** *Boston* **à** *Denver*", the preposition *de* designates *Boston* to be the

Semantic Category	*Example*
flight-time	*I would like to go from Denver to Boston*
fare	*Show me the fares from Denver to Boston*
stop	*What are the stopovers for flight 296*
type	*What is the type of aircraft for flight 1234*
reserve	*I would like to book a place in economy class*

Table 1: ATIS concepts.

```
CASEFRAME flight-time
{KEYWORDS: vol, voyager, aller, partir...
from: (quitte, de) @city
to: (a, pour, vers) @city
stop: (escale-a) @city
relative-departure-time: (partir+) avant, apres
departure-time: (partir+) @hour-minute
...}
CASEFRAME @city
{ city: denver, boston, dallas, atlanta...}
CASEFRAME @hour-minute
{...}
```

Figure 2: Example of a case frame.

departure town, and *à* designates *Denver* to be the arrival town. In the phrase "*à 14* **heures**", *heures* is an example of a postmarker, designating *14* to be a time. Pre- and post-case markers which are not necessarily located adjacent to the case may provide information useful to determine the context of the case. In "*qui* **arrive** *vers 22 heures*", the value *22* corresponds to the case arrival time, because it is preceded, although not directly, by the marker *arrive*.

A declarative language containing a list of possible case frames and associated cases is used to describe the case frames. The case frame structure is given in Figure 2. The case-frame list is organized in a the conceptual level (flight-time, fare, ...) and a level corresponding to common structures (@city, @hour-minute, ...). The KEYWORDS specify the words to select the given case frame during parsing. In the case frame flight-time, for instance, from and relative-departure-time are cases. The words in parantheses are premarkers, the symbol "+" indicates that the appropriate premarker may be located far away. The structure @city is the sub-case frame and the case city contains a list of towns. Hour-minute is also a sub-case frame, but since understanding of numbers is very relevant to the ATIS task (appearing also in dates and flight numbers), a restricted local grammar is used to extract the corresponding values.

The case-frame parser attempts to parse an input sentence using the whole set of case frames. It generates a semantic frame representing the meaning of the sentence. Sometimes the sentence may contain multiple queries which will result in the generation of multiple semantic frames. Sentence parsing is done by first selecting the corresponding case frame using keywords and then building a semantic frame representation of the meaning of the sentence by instanciating its slots. The parser is recursively applied on the

Q1) *Je veux aller demain matin de Denver à Boston avec escale à Atlanta (I would like to go tomorrow morning from Denver to Boston with a stop at Atlanta)*

```
<flight-time>
from:  denver
to:  boston
stop:  atlanta
relative-day:  tomorrow
morning-afternoon:  morning
```

Figure 3: Example query and corresponding semantic frame.

sub-case frames until there are no suitable words left to fill in the slots.

Figure 3 shows the resulting frame for an example utterance. For query **Q1**, the parser selects the case frame flight-time triggered by the keyword *aller* and constructs the complete semantic frame by instanciating the slots *from, to* and *stop* with the corresponding words *Denver, Boston* and *Atlanta* respectively. The analysis is driven by the order of the cases appear in the case frame flight-time. The following query will result in the same case frame as that of query **Q1**: *Montrez-moi les vols de demain matin allant à Boston en provenance de Denver avec escale à Atlanta (Show me the flights for tomorrow morning to Boston from Denver with a stopover in Atlanta).*

RESPONSE GENERATION

The response generator forms the database query, accesses the database, and presents the result to the user. An SQL request is built from the semantic frame using specific rules, where each rule forms a part of the SQL request. For example, the instanciated semantic frame flight-time of query **Q1** produces the SQL request "SELECT from-airport, to-airport, departure-time, arrival-time FROM flight". If the slots from and to contain values, then "WHERE from-airport=@from AND to-airport=@to" is concatenated to the SQL request, taking appropriate values for from and to from the semantic frame. The rules are described in a declarative file to allow easy modification. Once generated, the SQL request is used to access the database and retrieve information. The retrieved information is reformatted for presentation to the user along with an accompanying natural language response.

FRENCH ATIS CORPUS

We have chosen to work on a subset of the ATIS domain, which does not include ground transportation or meals, since these are not very meaningful to the subject population. All data were collected using the French version of the MIT-ATIS system. The initial parsing rules for French were obtained by translating ATIS queries from English. This shell system was used to obtain 505 typed queries, which were used to extend the coverage of the system.

The collection of the spoken data used a WOZ setup, where a wizard typed a paraphrased version of the spoken query to the system. The subjects were asked to solve a set of task-specific scenarios selected among 11 scenarios

Corpus	*WAcc*	*NL*	*SLS*
Type A only	86.7%	91.3%	79.0%
All Sentences	83.1%	91.7%	-

Table 2: Results on 138-class A sentences and entire 456 development sentences. The recognition word accuracy, natural language sentence understanding, and spoken language sentence understanding are given.

translated from English. Each session lasted about 50 minutes, during which the subject solved on average 6 scenarios and produced on average 50 queries. The recordings were made in an acoustically isolated room, simultaneously with a close-talking, noise cancelling Shure SM10 and a table-top Crown PCC160 microphone. The wizard monitored the recordings and could communicate with the subject via a microphone.

20 subjects were recorded, providing a set of 1111 spoken queries. There are on average 12 words per sentence including hesitations, false starts and reparations. The data from 12 speakers (655 sentences) were used to extend the vocabulary, to train the language model and to develop the case-frame grammar. The remaining 456 sentences from 8 speakers were reserved for development test material. From these, a set of 138 class 'A' utterances (able to be answered without the use of discourse history) were selected.

EXPERIMENTAL RESULTS

The experimental results are summarized in Table 2. The speech recognizer was evaluated *without the use of any task-specific* acoustic training data. The average word accuracy is 86.7% for the class-A sentences, with 1.2% out-of-vocabulary words. The word accuracy is 83.1% for all 456 development sentences, with 2.1% out-of-vocabulary words.

The understanding component was evaluated using the *exact transcriptions* of spoken queries including all spontaneous speech effects, such as hesitations or repetitions. Since evaluating the understanding component is quite delicate, our definition of a correct semantic frame is if the system produces an appropriate response after database access. An automatic method was developed for evaluation, which makes use of reference semantic frames for the 456 development utterances. The reference semantic frames were obtained by verifying and manually correcting the output of the case-frame parser. Correction for the class-A queries was done by running the parser on an paraphrased version of the query. For non-class-A queries, the reference semantic frames were manually corrected. The reference semantic frames are used to automatically evaluate the parser by comparison with the resulting semantic frames.

A correct semantic interpretation of 91.3% was obtained for the class-A queries. The semantic frame was judged to be correct for 91.7% of all 456 queries.

The spoken language understanding was evaluated by passing the output of the recognizer to the comprehension component. On the 138 class-A sentences, a correct semantic frame was instantiated for 79.0% of the queries.

SUMMARY AND PERSPECTIVES

In this paper we have presented a spoken language system which has be evaluated in the ATIS domain. The system is formed by feeding the output of our speech recognizer to a comprehension component. The recognizer uses acoustic models trained on the BREF corpus, and only the vocabulary items and language model make use of the ATIS training data. The average word accuracy on the class-A sentences is 87%, which resulted in correct semantic frames for 79% of these queries. The correct semantic interpretation was obtained by the NL component for 91% of all spoken queries using the exact word transcription.

One of our immediate goals is to use the current system to record new data. A near real-time version of the recognizer will be used for data collection instead of WOZ recordings. This data will then be used to extend the lexicon and to improve the language model. The task-specific data will also be used to train acoustic models in order to compare to the speech recognition accuracy using task-independent acoustic models. This data will provide examples of spontaneous speech effects that are not present in the BREF training data.

The understanding component was developed using the *typed* queries and the *exact* transcriptions of the spoken queries. The understanding component will be modified to take into account speech recognition errors.

REFERENCES

[1] S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, 18(1), 1992.

[2] H. Bonneau-Maynard, J.L. Gauvain, D. Goodine, L.F. Lamel, J. Polifroni, S. Seneff, "A French Version of the MIT-ATIS System: Portability Issues," *EUROSPEECH-93*.

[3] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *EUROSPEECH-93*.

[4] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "Continuous Speech Dictation in French," *ICSLP-94*.

[5] L.F. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *EUROSPEECH-91*.

[6] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. ASSP*, **32**(2), 1984.

[7] Ch.J. Fillmore, "The case for case," in *Universals in Linguistic Theory*, Emmon Bach & Robert T. Harms (eds.), Holt, Rinehart and Winston, Inc., 1968.

[8] B. Bruce, "Case Systems for Natural Language," *Artificial Intelligence*, 6, 1975.

[9] S.E. Levinson, K.L. Shipley, "A Conversational-Mode Airline Information and Reservation System Using Speech Input and Output," *Bell Sys. Tech. Journal*, Vol. 59, No. 1, pp. 119-137, 1980.

[10] P. Hayes, A. Hauptman, J. Carbonnell, M. Tomita, "Parsing Spoken Language, a Semantic Caseframe Approach," *COLING-86*.

[11] A. Matrouf, J.L. Gauvain, F. Néel, J. Mariani "An Oral Task-Oriented Dialog for Air-Traffic Controller Training," SPIE **1293**, *Applications of Artificial Intelligence, VIII*, April 1990.