Investigating Stochastic Speech Understanding

Hélène Bonneau-Maynard and Fabrice Lefèvre

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{maynard,lefevre}@limsi.fr

ABSTRACT

The need for human expertise in the development of a speech understanding system can be greatly reduced by the use of stochastic techniques. However corpus-based techniques require the annotation of large amounts of training data. Manual semantic annotation of such corpora is tedious, expensive, and subject to inconsistencies. This work investigates the influence of the training corpus size on the performance of understanding module. The use of automatically annotated data is also investigated as a means to increase the corpus size at a very low cost.

First, a stochastic speech understanding model developed using data collected with the LIMSI ARISE dialog system is presented. Its performance is shown to be comparable to that of the rule-based caseframe grammar currently used in the system. In a second step, two ways of reducing the development cost are pursued: (1) reducing of the amount of manually annotated data used to train the stochastic models and (2) using automatically annotated data in the training process.

1. INTRODUCTION

In our view the role of the speech understanding module in a dialog system is to extract the literal meaning of the user's query. Several well established ways of performing this extraction with relatively good performance exist. However, despite efforts made to ensure portability over domains and languages, defining a rule-based grammar for a new understanding component still requires extensive human expertise. During the last decade, several stochastic approaches for speech understanding have been proposed to reduce this need for human expertise [1, 2, 3].

In the LIMSI ARISE system, a rule-based caseframe grammar approach has been successfully applied to the speech understanding problem [7]. In this work we evaluate a stochastic understanding model developed using a corpus of dialogs collected with the LIMSI ARISE system. This work has given us the opportunity to address several issues which, to our knowledge, have not already been openly discussed. The first is: how sensitive is semantic component to tuning on a particular data set. A comparative evaluation is performed to address this question using both the rule-based and stochastic approaches.

A second important point is: what is the amount of data required to obtain good performance with a stochastic model? Effectively, if the need for human expertise is a weakness of rule-based approaches, the semantic annotation of a large amount of data, needed for stochastic approaches, is also a costly procedure. Therefore we have investigated the influence of the training corpus size on the performance of the understanding model in an attempt to determine the minimum amount of data required to ensure reasonable performance, thus reducing the development cost. A third question addressed is whether or not the development cost can be reduced by automatically annotating data to increase the training corpus size at a low cost.

The paper is organized as follows. The next section overviews the LIMSI ARISE task. In Section 3 a glass-box evaluation paradigm based on a Concept-Value representation of the domain is introduced. Then, the basis of the stochastic understanding approach is presented in Section 4. Finally, after the corpus description in Section 5, the experiments are reported in Section 6 .

2. THE LIMSI ARISE TASK

The LIMSI ARISE system [7] allows users to obtain travel information from the French national railway's static timetables by telephone. The system also provides information about services offered on the trains, reductions, fares, and fare-related restrictions. The system is composed of a speaker-independent real-time continuous speech recognizer, and components for natural language understanding, dialog management, database access and response generation.

The speech recognizer transforms the input signal into the most probable sequence of words and then forwards it to the natural language understanding component which carries out a literal understanding of the text string using a caseframe analysis and then reinterprets the query in the context of the ongoing dialog. The dialog manager ensures the communication between the user and the DBMS. If enough information is present, a pseudo-SQL request is generated to the DBMS. The result is given to the natural language response generation module and then to a synthesizer to provide vocal

User query	*dans la matinée*	*et*	*c'est*	*pas*	*Croisic*	*c'est*	*Roissy*
Recognized sentence	*dans la matinée*	*et*		*pas*	*Croisic*		*Roissy*
Concept sequence	`+/range-dep`	`+/null`		`-/m:mode`	`-/place`		`+/place`
Value normalization	matin				Croisic		Roissy

CVR	`+/range-dep`
	`-/place`
	`+/place`
		matin
		Croisic
		Roissy					

Table 1: Example of the semantic decoding for the sentence *"In the morning and it's not Croisic it's Roissy"*.

feedback.

The current understanding component carries out a rule-based caseframe analysis to determine the meaning of the query. Keywords are used to select the appropriate case structure. Case markers are used to provide syntactic constraints. In "*de Paris à Marseille*", for example, the preposition *de* designates *Paris* to be the departure town, and *à* designates *Marseille* to be the arrival town. Pre- and post-case markers which are not necessarily located adjacent to the case provide information useful to determine the context of the case. Sentence parsing is done by first selecting the corresponding caseframe using keywords and then building a semantic frame representation of the meaning of the sentence by instantiating its slots.

3. SEMANTIC ANNOTATION

A task-specific semantic representation has been defined for the ARISE domain. The feasibility of the evaluation process depends greatly on the semantic representation. We have chosen a frame concept/value representation (CVR). An example of a CVR is given in the last row of Table 1. The values are either numeric units, proper names or semantic classes merging lexical units which are synonyms for the task. The order of the concept/value pairs in the semantic representation follows their order in the user's utterance. A modal information (affirmative or negative) is assigned to each concept/value pair. The example given in Table 1 illustrates the use of the negative mode. The sequence *"c'est pas Croisic" (it's not Croisic)* is represented in the CVR with the `place` concept assigned with a negative mode (`-/place:` Croisic). The development of the CVR allows the definition of a concept dictionary, which specifies for each concept the set of possible values it can have. There are a total 64 concepts for the ARISE domain (128 with modality). A scoring tool has been developed to compare two semantic representations in terms of deletions, insertions, and substitutions [8]. The scoring is done on the whole triplet including mode, concept name and concept value.

For the stochastic understanding approach, the set of initial concepts is extended by 40 additional modal marker concepts (like `-/m:mode` for word "*pas*" in Table 1) and a `null` concept is associated to words not carrying any semantic information in the utterance (e.g. "*et*" in Table 1).

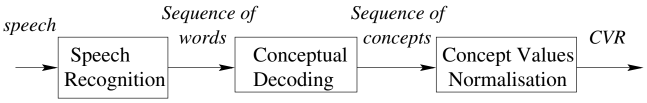

Figure 1: Block diagram of the speech understanding system

There are a total of 170 concepts used for the stochastic approach.

4. STOCHASTIC UNDERSTANDING

The aim of stochastic understanding is to find the sequence of semantic units (*concepts*) $C = c_1 c_2 \ldots c_N$ that will represent the meaning of the sentence, assuming that there is a sequential correspondence between the concept and word sequences [1].

Given $W = w_1 w_2 \ldots w_N$ the sequence of words in the sentence, the understanding process consists of finding the sequence of concepts which maximizes the *a posteriori* probability:

$$\hat{C} = \arg\max_C \Pr(C|W) \tag{1}$$

According to the Bayes formula, equation (1) can be rewritten as

$$\hat{C} = \arg\max_C \Pr(W|C) \Pr(C) \tag{2}$$

The term $\Pr(W|C)$ is estimated by means of n-gram probabilities of words given the concept associated to word i:

$$\Pr(W|C) = Pr(w_1) \prod_{i=2}^{N} Pr(w_i|w_{i-1}, \ldots, w_{i-n+1}, c_i)$$

and the term $\Pr(C)$ is estimated in terms of m-gram probabilities of concepts:

$$\Pr(C) = Pr(c_1) \prod_{i=2}^{N} Pr(c_i|c_{i-1}, \ldots, c_{i-m})$$

In the following experiments, the understanding model is limited to first order models *i.e.* $n = 1$ giving probabilities of words conditioned on a concept and $m = 1$ giving concept bigrams.

A block diagram of the stochastic speech understanding procedure is shown in Figure 1. An example of the representation used at different steps in the decoding process is given in Table 1. The speech recognizer transforms the acoustic signal into the most probable sequence of words (second row in Table 1). No *prior* transduction, such as lexical parsing, is performed. An exception concerns words that are always associated with the `null` concept in the training corpus. These *filler* words (such as *"euh", "ah'*, or "*je*") are removed before the conceptual decoding. The conceptual decoding is then carried out to segment the sentence into a sequence of concepts (third row in Table1).

The role of the concept value normalization module is to remove the `null` and marker concepts (e.g. `-/m:mode`) and to transduce each sequence of words assigned to a given concept into its normalized form, according to the CVR concept value list. In the example, the normalization module modifies the sequence *"dans la matinée"* assigned to the concept `range-dep` to the normalized form *"matin"* (fourth row in Table 1). The resulting CVR proposed by the whole understanding process for the example is given in the last row of Table 1.

5. CORPUS DESCRIPTION

The training set used in our experiments contains 14,582 sentences. These utterances have been extracted from the LIMSI ARISE corpus, which has over 10k dialogs of users interacting with the system. This corpus has been semi-manually annotated in terms of concepts [6]. The average number of words per utterance is 5. The total number of concepts in the training corpus is 44,812, giving an average number of 3 CVR concepts per utterance.

A development corpus of 400 utterances was used to check the evaluation procedure: mostly the relevance of the CVR representation and the scoring tool. The evaluation is performed on a test set of 496 utterances randomly selected from the remaining portion of the ARISE corpus. An iterative approach has been used to derive the reference CVR of the development and test sets [4].

In order not to bias in favor of one or the other of the systems, both understanding systems (rule-based and stochastic) were run on the manual transcriptions of the utterances. Then, the CVRs proposed by the stochastic approach were corrected by hand. The resulting CVRs were used to score the rule-based system result and hand corrections were made when appropriate.

Manual transcriptions are available for all utterances. The version of the ARISE speech recognizer used in our experiments has a recognition vocabulary of about 4k words including over 3k station names. The word error rate is 14.3% on the test corpus (13.4% for the development corpus).

Table 2 summarizes the characteristics of the training, development and test sets.

	Training	Dev.	Test
#Utt	14582	400	496
#Words	72380	2261	2880
#Concepts (in CVR)	44812	708	923
Word Error Rate	-	13.4%	14.3%

Table 2: Corpus description: number of utterances, words and CVR concepts of the training, development, and test sets. Word error rates of the recognized utterances are given for the development and test sets.

	Dev.		Test	
Approach	*Manual*	*Auto.*	*Manual*	*Auto.*
Rule-based	2.1	13.2	9.2	19.8
Stochastic	7.8	16.6	9.4	19.1

Table 3: Comparative understanding error rates (%) of the rule-based caseframe grammar and the stochastic model on manual (*Manual*) and automatic (*Auto.*) transcriptions for the development and test sets.

6. EXPERIMENTS

Three sets of experiments are reported. First, the stochastic approach is validated through a comparative evaluation with the rule-based approach. The development cost of a stochastic model is then addressed through the relation between training corpus size and model performance, and finally an attempt to use automatically annotated data is described.

Comparison with the rule-based caseframe approach

To assess the efficiency of the stochastic understanding method, its performance has been compared with the performance of the rule-based caseframe approach. The results are summarized in Table 3. The performance criterion is the understanding error rate measured on the reference CVR as described in Section 3.

The ability of the rule-based approach to fit a particular data set is shown on the development set where its UER is 2.1%, to be compared with 7.8% for the stochastic approach when the manual transcripts are used as input to the understanding component. The gap in performance between the approaches is significantly reduced when automatic transcriptions are used. The loss in performance due to recognition errors is 11% with the rule-based approach, compared to 9% for the stochastic one. It can also be observed that, unlike the rule-based approach, the stochastic approach obtains relatively close results on the development and test sets. Furthermore, the results on the test set show that both methods achieve comparable performance when confronted with unknown data: about 9% UER on the manual transcriptions and 19% UER on the automatic transcriptions.

Influence of the training corpus size

Manual annotation of a training corpus is a weak aspect of the stochastic understanding approach. The conceptual annotation of a 15,000 sentence corpus is a tedious and expensive process. As a first step in reducing the development cost, a series of experiments has been carried out to measure the impact of the training corpus size on the model performance.

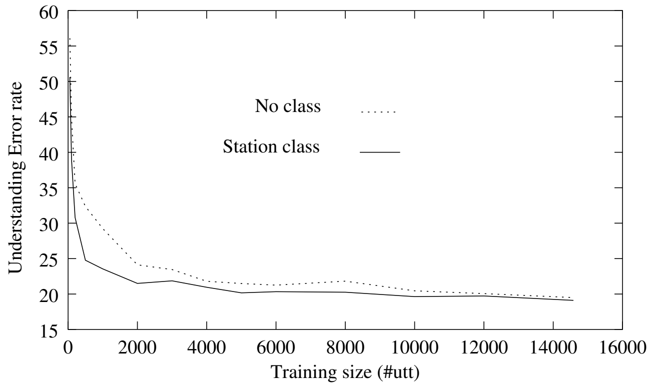

Figure 2: Variation of the understanding error rate on the automatic transcriptions of the test set as a function of the training set size (number of utterances). Two stochastic models have been tested: one using a lexical class for station names (solid line) and one without the lexical class (dotted line).

Figure 2 plots the variation of the UER as a function of the training set size (number of utterances). These results have been obtained using automatic transcriptions of the test set. Two stochastic models have been tested: one using a lexical class for the station names (solid line in Fig.2) and another without the lexical class (dotted line).

It can be seen in Figure 2 that the UER reaches an asymptote relatively quickly. In the case where a lexical class is used for the station names, the UER variations become small in the vicinity of 2,000 sentences. Without this lexical class, the same level of performance requires about twice the amount of data (4,000 utterances). The influence of the class on the model becomes negligible after about 11,000 utterances.

From the development point of view, a few thousand manually annotated utterances appear to be enough for the model to reach good performance (11.2% with 2,000 queries to be compared to 9.4% with the whole training set). On the other hand, it also shows that we have some leeway to increase the model complexity for it could make a better use of more data.

Training with automatically annotated data

A system-in-loop scheme has been successfully applied in the context of speech recognition [9] (for both acoustic and linguistic model adaptation). Here the same basic idea is investigated for the development of the stochastic understanding model. The approach consists of bootstrapping a model with a small amount of manually annotated data and using this model to automatically annotate newly collected data.

In order to simplify the procedure in these preliminary experiments, an initial model was built using 2,000 of the training utterances. This model was used to decode the remaining 12,482 utterances in the training corpus. Both subsets were then merged and a new understanding model built. The understanding error rate obtained with this new model showed no significant improvement over the initial model. This first result seems to show that supervision is still necessary during the development phase of the stochastic model.

7. CONCLUSIONS

Experiments with a stochastic speech understanding module for a train travel information task have been carried out and a comparative evaluation of this model with a rule-based caseframe approach has shown that comparable performances are obtained with both methods when confronted with unknown data. Since (semi)-manual annotation is an important cost in developing a stochastic model, we investigated the performance of the model as a function of the amount of annotated training data. These experiments indicate that the annotated training set can be reduced to a few thousand utterances without an important loss in performance. It leads us to question whether our simple model could take advantage of the additional data or whether a more complex model will need to be used.

A first attempt at using automatically annotated data did not improve the understanding performance, which seems to imply that we cannot yet avoid supervision during the development of the stochastic model. Future work will assess if, at least, the level of supervision can be reduced.

8. ACKNOWLEDGMENT

Special thanks go to Sophie Rosset for providing the results on the rule-based caseframe grammar approach and to Lori Lamel and Jean-Luc Gauvain for their valuable comments.

REFERENCES

[1] R. Pieraccini, E. Levin, “A learning Approach to Natural Language Understanding”, in NATO ASI Series, Springer-Verlag, Bubion, 1993.

[2] R. Schwartz et al, “Hidden Understanding Models for Statistical Sentence Understanding”, *Proc. ICASSP’97*, Munich.

[3] G. Riccardi, A. Gorin, “Stochastic Language Models for Speech Recognition and Understanding”, *Proc. ICSLP’98*, Sydney.

[4] R.C. Moore, “Semantic Evaluation for Spoken-Language Systems, *Proc. DARPA Speech and Natural Language Workshop*, pp. 122-127, 1994.

[5] S. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L. Lamel,W. Minker, “A spoken language system for information retrieval”, *Proc. ICSLP’94*, Yokohama.

[6] W. Minker, “Compréhension Automatique de la Parole Spontanée”, PhD Thesis, Université Paris XI, 1998.

[7] L. Lamel et al, “The LIMSI ARISE System”, Speech Communication, vol31, pp. 339-353, 2000.

[8] H. Bonneau-Maynard, L. Devillers “A Framework for evaluating contextual understanding,” *Proc. ICSLP’2000*, Beijing.

[9] F. Lefevre, J.-L. Gauvain, L. Lamel “Genericity and Adaptability Issues for Task-Independent Speech Recognition,” *ISCA ITRW on Adaptation Methods for Speech Recognition*, Nice, 2001.