DEVELOPMENT AND EVALUATION OF THE ATOS SPONTANEOUS SPEECH CONVERSATIONAL SYSTEM

J. Álvarez, D. Tapias, C. Crespo, I. Cortazar, F. Martínez
Speech Technology Group
Telefónica Investigación y Desarrollo, S.A. Unipersonal
C/ Emilio Vargas, 6
28043 - Madrid (Spain)
jorge@craso.tid.es

ABSTRACT

In this paper we report our recent development work in Spanish spontaneous speech conversational systems. We describe the Automatic Telephone Operator Service (ATOS) and present the improvements introduced into it to deal with spontaneous speech, which are: (a) a task independent dialogue manager, that can be adapted to a new semantic domain by changing a configuration file. It also generates a prediction about the user's expected utterance to constrain the language model used by the speech recognizer. (b) a language modeling strategy, which allows to adapt the statistical language model to a new task with just few hundreds of sentences. This strategy reduces a 27% the word error rate.

We also report the results, conclusions and the speech database collected in the evaluation of the ATOS system, which has been tested by 30 real users.

1. INTRODUCTION

The Automatic Telephone Operator Service (ATOS) is a conversational system developed at Telefónica Investigación y Desarrollo (TID). It integrates large vocabulary continuous speech recognition, natural language processing, natural language generation and text to speech conversion. The ATOS system has been designed to provide users with four kind of services:

(a) personal agenda, that allows users to add or delete entries and also to retrieve private telephone numbers,

(b) telephone directory, that contains the telephone numbers of all the employees of TID and selects what telephone number is going to be used depending on the time of the day and the day of the week,

(c) mail box, which allows to play or delete a message, record the welcome message, etc..., and

(d) PABX services like recall, call transfer, do not disturb, etc... Section 2 describes the structure of the entire system.

The work we present deal with the spontaneous speech problem: It is not clear how users are going to interact with machines in real applications. For this reason the ATOS has been evaluated by real users. The evaluation has pursued four goals: (a) the performance evaluation of the speech recognizer, the natural language processor and the entire conversational system, (b) the collection of a speech database of real human-machine dialogue in a domain-dependent application, that is helping us to find out what the particularities of human-machine communication are, (c) the acquisition of a text database to improve the language model (LM), and (d) a subjective evaluation of the system in terms of usefulness, satisfaction and user suggestions.

Another problem related with conversational systems is that their implementation takes a long time: If the semantic domain changes, the acoustic and language models as well as the natural language processor and natural language generator will have to be adapted to the new domain. The acoustic models can be easily used to model unseen triphones by means of senonic decision trees [1]. Nevertheless, it is difficult to adapt a general language model to a specific task unless a large enough text database is available for that domain.

We have developed a method to adapt a generic language model to a specific domain using just few hundreds of sentences belonging to the new domain. This method is described in section 3. Section 4 presents the description of the Dialogue Manager (DM), which is easy to adapt to new domains by changing a configuration file and generates a prediction about the user's expected utterance. The prediction is then passed to the speech recognizer to constrain the LM. In section 5 we present in detail the experiments, results and conclusions.

2. SYSTEM OVERVIEW

The general scheme of the system is represented in Figure 1 and acts as follows: once an order has been uttered, a speaker independent continuous speech recognizer gets the most likely string of words. Then, a semantic parser, which is a flexible frame-based parser developed at Carnegie Mellon University (CMU) [2], is used to match substrings of the recognized sentences with semantic slots. The semantic slots represent the basic semantic entities known by the system. If all the slots needed to perform an action have been filled in, then the order is executed. Otherwise, the dialogue manager will ask the user, through the message generator and the text to speech system, for additional information.

The DM [3] controls the overall system. It can deal with different tasks by just changing a configuration file which specifies a semantic tree for each task. The DM takes all the filled semantic slots as input, and tries to fill the slots of the frame that are related with the current dialogue state. When an order is completed, the DM can engage one of the following actions: if it is a command, then the DM orders to execute it, but if it is an information request, the DM retrieves the information and passes it to the message generator. Finally, the DM introduces constraints into the LM, which depend on the state of the dialogue.

The continuous speech recognizer is a modified version of the sphinx-II decoder developed at CMU [4]. It uses a multipass search approach. It is based on semicontinuous hidden Markov models and uses a statistical language model. The acoustic models represent intra-word and cross-word context-dependent phone units.

3. LANGUAGE MODEL

It is well known the importance of the language model (LM) in the performance of continuous speech recognizers, hence, a well estimated LM is always convenient. This is achieved by using a large set of training sentences belonging to the task semantic domain and smoothing techniques to model poorly estimated n-grams. The problem is that it is difficult to get large enough textual corpus for every semantic domain, mostly if we want to model a LM for person-machine dialogues.

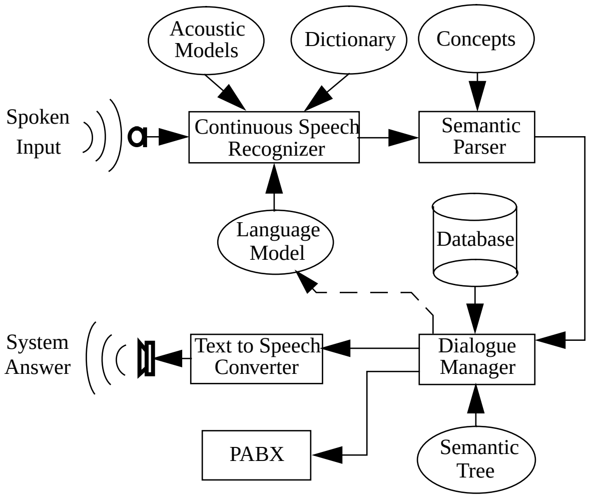

Figure 1: Block diagram of the ATOS conversational system

We propose a method which tries to overcome this problem. The key idea is to build a generic LM using as many sentences as possible, and then adapting it to the target domain by using a small set of target domain sentences [7]. The overall procedure can be summarized in four steps:

1.- A word-bigram back-off LM is built from a set of sentences of Spanish newspapers.

2.- A class-bigram back-off LM is built using the small set of sentences belonging to the target semantic domain. This set is composed of just three hundred sentences automatically tagged by a tagger developed at TID [5].

3.- A class-based LM is built using the following simplified expression for the bigram case:

$$P(w_2/w_1) = P(w_2/C_2) \times P(C_2/C_1)$$

where:

$P(w_2 / C_2)$: probability of word w_2 tagged as C_2 (the most likely class)

$P(C_2 / C_1)$: bigram C_1-C_2, probability according to the class pair backoff language model

4.- Finally, an interpolated class-based LM is built. This is carried out by interpolating the language models obtained in steps 1 and 3.

The back-off models have been generated with the Statistical Language Model Toolkit developed at CMU [6].

Our experiments show a 27% reduction in the word error rate when we use the step-4 LM instead of the one obtained in step-1.

4. DIALOGUE MANAGER

The goal of a conversational system is to provide users with the service they are asking for. In order to do this, the system has to dialogue to the user to find out what the user would like it to do. There are two constraints in the design of the DM: (a) the input sentence may have recognition errors, so that the *semantic parser* has to be flexible enough to get as much information as possible from incorrect sentences. (b) The information has to be extracted and *understood* in whatever order it comes; besides the system has to allow users to take the initiative, at least at the beginning of the dialogue.

In addition to this, the DM has to produce predictions about the user’s kind of expected utterance, so that this information can be used by the speech recognizer to increase the word accuracy. It also has to be easy to adapt to new semantic domains to facilitate the creation of new conversational system applications.

4.1 Semantic Analyzer

The semantic analyser scans the input sentence and looks for words or sequences of words that represent relevant information from the application point of view. Every piece of relevant information is called concept; concepts are the minimal meaningful entities used in the task. In a design stage all combinations of words that validate a concept have to be defined in a configuration file by means of a finite state network. During the analysis stage, when a sequence of words in the sentence validates a network, the semantic analyser labels those words with a concept.

4.2 Dialogue Manager

The Dialogue Manager is the engine of the system; it controls the work of each module in the conversational system and controls the human-machine dialogue by means of two acts: concept comprehension and answer generation.

During the concept comprehension stage, the DM receives all concepts that have been generated by the semantic analyser. At this moment the concepts are still separated pieces of knowledge. Then the DM performs *concept unification* by grouping concepts into semantic frames.

The DC has to *generate a proper answer* too. For this task, the DC uses two kinds of strategies: a global strategy oriented towards detection of missing concepts in a semantic frame; and a local strategy for finding the best way of interacting with the user at every moment.

The required information for the comprehension and the dialogue strategy is organized by means of “semantic trees”. Figure 2 shows a branch of the semantic tree which represents all the information related to extracting messages from a particular voice mail box. The upper most node represents the entire ATOS conversational system. The left child node, “Read_message”, represents one of the functions of the system, which allows to listen to a particular message (mail_id) from some user’s mail box (mail_box_id). Moreover, you can distinguish between new and recorded messages. The child nodes indicate two different ways of accessing these messages: by arrival order or time.

There are four kind of nodes in the semantic tree: (a) *necessary nodes*, whose parameters have to be obtained, (b) *optional nodes*, which provide extra-information, so that they may not be completed, (c) *complementary nodes*, where only one of them have to be completed, and (d) *request for confirmation nodes*, that must be confirmed by the user. The behavior of the DM depends on the kind of node, so for example only when a request for confirmation node is filled in with some information, the DM activates the WAITING_FOR_CONFIRMATION state.

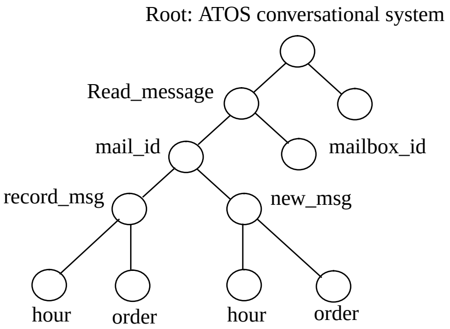

Figure 2: Semantic tree example

Additionally, the default behavior of the DM can be modified by means of the *implication* and *restriction rules*. If a node is modified by an implication rule, then when this node is filled in, another node is also filled in. If a node is modified by a restriction rule, once the node is filled in, the parameter associated with the node is checked to be validated. The DM considers two types of restrictions: comparative (greater than, equal to,...) and list ordering (after, before,...).

This way of structuring the dialogue allows to easily add new functions to the system or to adapt the DM to new semantic domains by just defining similar branches and integrating them in the semantic tree.

After a sentence is processed by the DM, the speech recognizer receives information about the expected user’s kind of answer. This information is then used during the recognition process to increase the probability of the words associated with the DM prediction.

5. EXPERIMENTS AND RESULTS

The ATOS conversational system has been evaluated by 30 users. This evaluation is playing an important role in the detection of problems related to all the components of the system. The users were asked to use all the features of the system and to evaluate each one of them in terms of usefulness and satisfaction. They also were asked to give a global evaluation of the system and were invited to suggest improvements. Additionally, they were told to speak spontaneously. The user’s utterances, the recognized sentences and the system messages were recorded and are being analyzed. In this paper we present the results and conclusions, which have been derived after analyzing all the dialogues.

The vocabulary size of the system is currently 2000 words and the Out-of-Vocabulary (OOV) word rate is 9.3%. Concerning the false starts rate, its percentage is 4.2%.

The quality of the system has been measure in terms of five parameters:

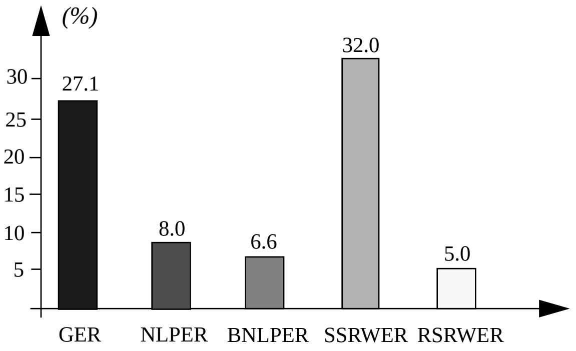

Figure 3: Error rates of the ATOS conversational system

(a) Global Error Rate (GER): This parameter shows the percentage of times a user's sentence is not understood by the system.

(b) Natural Language Processing Error Rate (NLPER): It is the percentage of *recognized sentences* that are not properly processed by the natural language processing (NLP) module. The NLP module is composed of the semantic parser, the dialogue manager and the natural language generator.

(c) Basic Natural Language Processing Error Rate (BNLPER): It is the error rate of the NLP module when the incoming sentences have no recognition errors.

(d) Spontaneous Speech Recognition Word Error Rate (SSRWER): It is the word error rate of the Speech Recognizer when it is recognizing spontaneous speech.

(e) Read Speech Recognition Word Error Rate (RSRWER): It is the word error rate of the Speech recognizer when it is recognizing read speech of the same semantic domain.

Figure 3 shows the error rates described above. As can be observed, just 27.1% of the utterances were totally or partially misunderstood by the system. The word error rate of the speech recognizer is 32% for spontaneous speech while it is 5% for read speech, what shows the degradation of the performance when there are hesitations, false starts, OOVs and other effects like competitive background speech or background stationary and non-stationary noise.

Concerning the Natural Language Processing module, the error rate for sentences with no recognition errors (BNLPER) is 6.6% while the error rate for the recognized sentences (NLPER) increases up to 8.0%.

5. CONCLUSIONS

In this article a conversational system has been introduced. The main source of errors in conversational systems comes from the speech recognizer module ought to the tremendous difficulty of dealing with spontaneous speech. The idea underlying this system relays on the analysis of recognition errors. If we take a look to this errors we can see that usually a human could recover the main information from the recognized sentence even when the sentence has some recognition errors. So perhaps, a NLP module can also recover from those errors in order to execute some simple functions.

We strongly think that one of the main characteristics of this kind of systems should be fast adaptation to different tasks: Continuous speech recognizers can be adapted to different domains almost automatically. However, the NLP module has to be adapted manually, so that one of our goals was to implement a NLP module able to reuse the source code. In this way, the adaptation process of the NLP to a new task is done by means of configuration files.

The obtained results of the entire system (speech recognizer and natural language processor) overcome the performance obtained just by the speech recognizer. This fact points out that the use of a natural language processor is a good method to augment the system robustness.

REFERENCES

[1] Mei-Yuh Hwang, “Subphonetic Acoustic Modeling for Speaker Independent Continuous Speech Recognition”, Doctoral Thesis, Carnegie Mellon University, Pittsburgh (USA), Dec. 1993.

[2] S. Issar and W. Ward, “CMU’s Robust Spoken Language Understanding System”, In Proc. Eurospeech’93, pp. 2147-2150. Sept. 1993.

[3] J. Caminero, J. Álvarez, C. Crespo, D. Tapias, “Data-Driven Discourse Modeling for Semantic Interpretation”, In Proc. ICASSP’96, pp. 401-404, Atlanta (USA), May 1996.

[4] M. K. Ravishankar, “Efficient Algorithms for Speech Recognition”, Doctoral Thesis, Carnegie Mellon University, Pittsburgh (USA), May 1996.

[5] M. A. Rodriguez, J. G. Escalada, A. Macarrón and L. Monzón, “AMIGO: Un Conversor Texto-Voz para el Español”, Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN’92, Vol. 13, pp. 389-400. Sept. 1992.

[6] R. Rosenfeld, “Adaptive Statistical Language Modeling: A Maximum Entropy Approach”, Doctoral Thesis, Carnegie Mellon University, Pittsburgh (USA), April 1994.

[7] C. Crespo, D. Tapias, G. Escalada, J. Álvarez, “Language Model Adaptation for Conversational Speech Recognition Using Automatically Tagged Pseudo-Morphological Classes”, In Proc. ICASSP’97. Munich (Germany).