INCORPORATING SEMANTIC KNOWLEDGE TO THE LANGUAGE MODEL IN A SPEECH UNDERSTANDING SYSTEM

Sergio Grau, Encarna Segarra, Emilio Sanchis, Fernando García, Lluís F. Hurtado

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, E-46022 València, Spain
{sgrau,esegarra,esanchis,fgarcia,lhurtado}@dsic.upv.es

ABSTRACT

The majority of the understanding systems follows an architecture based on two modules, a speech recognition module and an understanding module. Usually, only syntactic restrictions are incorporated to the speech recognition module through the language model and the semantic restrictions are incorporated in the understanding module. In this work, we present an approach to language understanding where the semantic knowledge involved in the understanding process is incorporated through an adequate definition of the language model of the automatic speech recognition module. Then, both the recognition and understanding processes incorporate semantic knowledge. An evaluation of the behavior of the proposed understanding system in the framework of a dialog system is also presented. The results show that the use of semantic information in the language model of the speech recognizer provides for the best performance.

1. INTRODUCTION

In many systems of human-machine interaction, the understanding process is one of the most important parts. This is the case of spoken dialog systems in which the information that must be extracted from the user utterances is not the exact sequence of words, but the meaning of the utterance as well as the specific values that appear in it.

Although approaches to language understanding have traditionally used hand-built semantic rules to detect keywords that are used to fill slots in a frame, other approaches that are based on the use of stochastic models have been developed. The BNN-HUM [1], the AT&T-CHRONUS [2], and the LIMSI-ARISE [3] [4] are some examples of the use of Hidden Markov Models and N-gram models to stochastically model the understanding process from training data. There are also other statistical approaches based on classification, transduction, and grammatical inference techniques: [5], [6], [7], [8] and [9]. These stochastic approaches tackle the understanding process as a problem of transduction of the input sentence into a semantic representation. They can model the variability of the language from real data and take into account the possible sources of error.

An interesting point of study in the speech recognition/understanding system is how to apply the syntactic/semantic restrictions during the decoding process. The classical approach considers a language model of words (typically bigrams or trigrams) during the recognition process, and in a second phase, the sentence (or n-best sentences or a graph of words) obtained is analyzed by the understanding module in order to obtain the semantic representation of the input utterance. However, there is another possibility: to use semantic restrictions during the recognition process in order to guide the search not only by the syntactic restrictions but also by the semantic ones.

In this work, we present an approach to language understanding where the semantic knowledge involved in the understanding process is incorporated through an adequate definition of the language model of the automatic speech recognition module. Then, both the recognition and understanding processes incorporate semantic knowledge. We also present an integrated speech recognition/understanding system where the system makes both the recognition and the understanding processes in just a single step. This approach has been applied to the recognition/understanding module of the DIHANA dialog system, which answers queries by telephone about railway timetables and prices in Spanish [10]. Some experimental results are presented.

In section 2, we describe the DIHANA task and the semantic representation designed for it. In section 3, we present our approach to the understanding process. In section 4, we evaluate the performance of the system; and, in section 5, we present some conclusions.

2. THE DIHANA TASK AND THE SEMANTIC REPRESENTATION

One of the objectives of this project was the acquisition of a corpus of dialogs. The DIHANA task consists of a telephone-based information service for trains in Spanish. A set of 900 dialogs was acquired by using the Wizard of Oz technique. Two hundred and twenty-five naive speakers collaborated in the acquisition of dialogs corresponding to different scenarios. Each one produced 4 dialogs. Three scenarios were defined: timetables for a one-way trip or a two-way trip, prices, and services. The number of user turns was 6,280 and the vocabulary was 823.

As in many other dialog systems [3], the semantic representation chosen for the task is based on the concept of frame. Therefore, the understanding module generates one or more frames with their corresponding attributes as output. In this task, we identified eight concepts. Some of them are: DEPART-TIME, ARRIVAL-TIME, PRICE, etc. Each concept has a set of attributes associated to it (ORIGIN, DESTINATION, DEPART-TIME, ARRIVAL-TIME, TRAIN-TYPE, etc). This set represents the restrictions that the user can place on each concept in an utterance.

3. THE UNDERSTANDING SYSTEM

We propose an understanding system that works in two phases. The first phase consists of a transduction of the input sentence in terms of an intermediate semantic language. In the second phase, a set of rules transduces this intermediate representation in terms of frames. As the intermediate language is close to

the frame representation, this phase only requires a small set of rules to construct the frame. This second phase consists of the following: the deletion of irrelevant segments of the input sentence, the reordering of the relevant concepts and attributes that appeared in the user sentence following an order which has been defined a priori, the instantiation of certain task-dependent values, etc.

In order to represent the meaning of the sentences in terms of the intermediate semantic language, a set of 31 semantic units was defined. Some of them are: *query*, *affirmation*, <departure_hour> <price>, *origin_city*, *destination_city*, *departure_hour*, *departure_date*, *depart_marker*, *arrival_marker*. (We used < and > to distinguish concepts from attributes).

Each semantic unit represents the meaning of words (or sequences of words) in the sentences. For example, the semantic unit $query$ can be associated to "can you tell me", "please tell me", "what is", etc. This way, an input sentence (sequence of words) has a semantic sentence (sequence of semantic units) associated to it, and there is an inherent segmentation.

In this work, we propose two approaches to carry out the first phase of this understanding process. In the first one, the decoupled approach, an automatic speech recognition module produces a sequence of words, and then an understanding module translates it to a sequence of semantic units. In the second one, the integrated approach, the understanding is made through the incorporation of the syntactic and semantic knowledge into the automatic speech recognition module. This module generates, not only the recognized sentence, but also the corresponding sequence of semantic units.

From the sequence of semantic units, the second phase of the understanding process is applied, and the corresponding frame is obtained through a set of rules.

3.1. The decoupled approach

In this approach [6], two kinds of models must be learnt from a training set of semantically tagged and segmented sentences: a semantic model that represents the concatenations of semantic units, and a model for each semantic unit that represents the language of sequences of words associated to that semantic unit.

In order to learn these stochastic models, a set of sequences of semantic units associated to the input sentences, as well as the corresponding association of segments of words to the semantic units must be available. That is, let W be the vocabulary of the task, and let V be the alphabet of semantic units; the training set is a set of pairs (u,v) where:

$$u = u_1, u_2 \dots u_n, u_i = w_{i_1} w_{i_2} \dots w_{i_{|u_i|}},$$
$$w_{i_j} \in W,\ i = 1, \dots, n,\ j = 1, \dots, |u_i|$$
$$v = v_1 v_2 \dots v_n,\ v_i \in V,\ i = 1, \dots, n$$

Each sentence from W has an associated pair *(u,v)*, where *v* is the sequence of semantic units and *u* the sequence of segments in which the original sentence has been divided. An example with a sentence, the associated sequence of semantic units, and the corresponding segmentation is shown in Figure 1.

u_1: I would like u_2: the train timetables u_3: from Valencia u_4: to Barcelona	v_1: query v_2: <departure_hour> v_3: origin_city v_4: destination_city
Input pair (u,v) = ($u_1 u_2 u_3 u_4$,$v_1 v_2 v_3 v_4$)	
Output v = query <departure_hour> origin_city destination_city	

Figure 1. An example of a pair *(u,v)*.

When a training corpus is available, the learning of the sequential translator is carried out through the learning of two models: a model for the semantic language and a set of models (with one model for each semantic unit v_i).

For the understanding process, all the models must be combined in order to take advantage of all the syntactic and semantic restrictions. To do that, in the stochastic automaton for the semantic language, each state (which is associated to each semantic unit) is substituted by the corresponding stochastic automaton, (which represents the sequences of words associated to that semantic unit). The understanding process is performed using the Viterbi algorithm, which supplies the best path in the integrated model. This path not only gives the sequence of semantic units, but it also gives the segmentation associated to it.

3.2. The integrated approach

In this approach, the understanding model represents the semantic knowledge involved in the understanding process through an adequate definition of the language model of the automatic speech recognition module. Then, the recognition and understanding processes are performed at the same time by a single module. In this approach, not only syntactic restrictions, but also semantic restrictions are applied through the language model during the speech recognition process. This language model is learnt using the Morphic Generator Grammatical Inference (MGGI) methodology [11].

The MGGI methodology is a grammatical inference technique that allows us to obtain a certain variety of regular languages. The application of this methodology implies the definition of a renaming function; that is, each symbol of each input sample is renamed following a given function g. Then, a classical grammatical inference algorithm can be chosen to infer an automaton with the renamed training samples. Finally, the renamed symbols are converted back to the original ones in the obtained automaton. An important characteristic of this methodology is that different definitions of the function g will produce different models. Therefore, we can choose an adequate renaming function depending on the characteristics we want to represent in the model.

In this work, we defined the renaming function g in such a way that it specialized each word in a segment u_i by adding to it information about its semantic unit v_i. Figure 2 shows the application of this renaming function g to the example in Figure 1. We used # to concatenate each word with the name of its semantic unit.

u_1: I would like u_2: the train timetables u_3: from Valencia u_4: to Barcelona	v_1: query v_2: <departure_hour> v_3: origin_city v_4: destination_city
Input pair (u,v) = ($u_1 u_2 u_3 u_4$,$v_1 v_2 v_3 v_4$)	
Output g((u,v)) = I#query would#query like#query the#<departure_hour> train#<departure_hour> timetables#<departure_hour> from#origin_city Valencia#origin_city to#destination_city Barcelona#destination_city	

Figure 2. An example of the use of the renaming function.

The training corpus labeled as described above can be used as the language model in the speech recognition process. In this way, both syntactic restrictions, and semantic constraints are considered in the search space of the recognition process. As a result, the recognition process not only gives the sequence of words, but also the semantic label associated to each word.

Therefore, we can obtain the segmentation associated to the sentence by including in the same segment all the consecutive words labeled with the same semantic label, as shown in Figure 3.

Recognizer Output = I#query would#query like#query the#<departure_hour> train#<departure_hour> timetables#<departure_hour> from#origin_city Valencia#origin_city to#destination_city Barcelona#destination_city	
Output = (u,v)	
u_1: I would like	v_1: query
u_2: the train timetables	v_2: <departure_hour>
u_3: from Valencia	v_3: origin_city
u_4: to Barcelona	v_4: destination_city

Figure 3. Obtaining the pair *(u,v)* from the output of the speech recognition process.

4. EXPERIMENTAL RESULTS

We used the CMU Sphinx-II recognizer [12] to decode the user utterances. We trained semi-continuous acoustic models from 3,600 telephone-quality utterances acquired in the DIHANA project. The models were trained using a set of 25 phones plus silence for the Spanish.

In this section, we describe the results of the evaluation of our understanding systems. Three different understanding systems were tested:

- (Trigrams+UM) A decoupled understanding system with a speech recognizer that used a trigram of words as language model, and the understanding module (UM) based on the finite-state models presented in subsection 3.1.
- (MGGI_LM) An integrated understanding system that incorporates both the syntactic and semantic knowledge into the language model (see subsection 3.2).
- (MGGI_LM+UM) A decoupled understanding system with the MGGI_LM module used as speech recognition module and the understanding module (UM) presented in subsection 3.1.

The experiments were performed using the user turns of the 900 dialogs obtained through the Wizard of Oz technique. The characteristics of the transcribed corpus are shown in Table 1. The orthographic transcriptions of the user turns were semi-automatically segmented and labelled in terms of semantic units. The characteristics of this labelled and segmented corpus are shown in Table 2.

Table 1. Characteristics of the transcribed corpus

Number of turns	6,227
Number of words	47,196
Vocabulary size	823
Average number of words in turn	7.58
Longest turn	55

A cross-validation procedure was used to evaluate the performance of our understanding models. To this end, the experimental set was randomly split into five subsets of 1,246 turns. Our experiment consisted of five trials, each of which had a different combination of one subset (taken from the five subsets) as the test set. The remaining 4,981 turns were used as the training set.

Table 2. Characteristics of the labelled and segmented corpus

Number of semantic segments	18,570
Average number of words per segment	2.54
Highest number of words in a segment	25
Average number of segments per turn	2.98
Highest number of segments in a turn	16
Number of semantic segments without semantic relevance	1,631
Number of semantic units	31
Number of different sequences of semantic segments	1,592

We defined four measures to evaluate the accuracy of the models:

- the percentage of correct sequences of semantic units (%cssu).
- the percentage of correct semantic units (%csu).
- the percentage of correct frame names (%cfn); i.e., the percentage of resulting frame names that are exactly the same as the corresponding reference frame names.
- the percentage of correct frame slot names (frame name and its attribute names) (%cfsn).

The measure $\%csu$ allows us to evaluate the first phase of our understanding system. This measure is the concept accuracy and is calculated in the same way as the word accuracy used in speech recognition. The measures $\%cfn$ and $\%cfsn$ evaluate the overall understanding system. As shown in Section 2, the semantic representation of a sentence is made by one or more frames. The measure $\%cfn$ considers the output to be correct only when the obtained frame (frame name and its attribute names) is the same as the reference one. The measure $\%cfsn$ is the frame slot accuracy, that is, the number of correctly understood units (frame name and its attribute names) divided by the number of units in the reference.

A first experiment was performed using the correct transcriptions of the sentences in order to evaluate the understanding models when working under the best conditions of the input. The understanding model used in these experiment was UM, presented in subsection 3.1. We used a bigram for the semantic model and also for the semantic unit model for each semantic unit; the smoothing technique for bigrams was back-off with Good-Turing discounting. The results, which are reported as Text+UM, are shown in Table 3.

Experiments taking into account the recognition and understanding processes were also performed. Table 3 shows the results for the three understanding systems, Trigrams+UM, MGGI_LM+UM, and MGGI_LM, in terms of the understanding measures and in terms of the word accuracy (WA) of the recognition process.

Table 3. Understanding results.

LM	WA	%cssu	%csu	%cfn	%cfsn
Text+UM	-	79.6	90.4	90.2	93.9
Trigrams+UM	77.3	63.0	79.0	73.2	83.9
MGGI_LM+UM	77.5	64.2	79.5	74.1	84.5
MGGI_LM	77.5	62.3	78.0	71.2	82.5

The word accuracy results of the recognizer that used syntactic-semantic language models (MGGI_LM) slightly outperformed the corresponding results of the recognizer that used word trigrams. That means that the use of the semantic restrictions in the recognition process can be useful for speech recognition in dialogue systems tasks. On the other hand, the integrated model did not outperform the decoupled approaches in

terms of speech understanding results. However, using the recognized sequence of words from the MGGI_LM as input for the understanding module improved the overall results of the speech understanding system.

5. CONCLUSIONS AND FUTURE WORK

We have presented an approach to incorporate semantic knowledge to the language model of a speech recognition module within the framework of a dialogue system. We have studied the use of this semantic information in order to improve not only the word accuracy of the automatic speech recognition module but also the overall speech understanding results. For this reason, three different understanding systems have been tested. The best results were obtained using the sequence of words recognized by the MGGI_LM speech recognizer as input for the understanding module. Therefore, when syntactic and semantic information is used through the language model in the recognition process, the results outperform the understanding results when word trigrams are used.

6. ACKNOWLEDGEMENTS

Work partially supported by the Spanish CICYT under contract TIN2005-08660-C04-02

7. REFERENCES

[1] R. Schwartz, S. Miller, D. Stallard, y J. Makhoul, "Language understanding using hidden understanding models," in *Proc. ICSLP'96*, Philadelphia (USA), Oct. 1996, pp. 997–1000.

[2] E. Levin y P. Pieraccini, "Concept-based spontaneous speech understanding system.," in *Proc. of Eurospeech'95*, 1995, pp. 555–558.

[3] W. Minker, A. Waibel, y J. Mariani, *Stocastically-Based Semantic Analysis*, Kluwer Academic Publishers, Boston, 1999.

[4] H. Bonneau-Maynard y F. Lefèvre, "Investigating stochastic speech understanding," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'01)*, 2001.

[5] K. Macherey, F.J Och, y H. Hermann Ney, "Natural Language Understanding Using Statistical Machine Translation," in *Proc. of EUROSPEECH*, September 2001, pp. 2205–2208.

[6] E. Segarra, E. Sanchis, F. García, y L.F. Hurtado, "Extracting semantic information through automatic learning techniques," in *International Journal of Pattern Recognition and Artificial Intelligence*, Salt Lake City (USA), 2002, pp. 16(3):301–307.

[7] Yulan He y S. Young, "A data-driven spoken language understanding system," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'03)*, 2003, pp. 583–588.

[8] S. Gupta, N.K. Bangalore, "Segmenting spoken language utterances into clauses for semantic clasiffication," in *Proc. of ASRU*, 2003, pp. 525–530.

[9] Y. Esteve, C Raymond, y R. Bechet, F. De Mori, "Conceptual Decoding for Spoken Dialog systems," in *Proc. of Eurospeech*, 2003, vol. 1, pp. 617–620.

[10] J.M. Benedí, A. Varona, y E. Lleida, "DIHANA: Sistema de diálogo para el acceso a la información en habla espontánea en diferentes entornos," in *Actas de las III Jornadas en Tecnología del Habla*, Valencia (España), 2004, pp. 141–146.

[11] E. Segarra y L. Hurtado, "Construction of Language Models using Morfic Generator Grammatical Inference MGGI Methodology," in *Proc. of Eurospeech*, 1997, pp. 2695–2698.

[12] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, y Ronald Rosenfeld, "The SPHINX-II speech recognition system: an overview," *Computer Speech and Language*, vol. 7, no. 2, pp. 137–148, 1993.