Language Understanding using Two-level Stochastic Models with POS and Semantic Units

F. Pla, A. Molina, E. Sanchis, E. Segarra, and F. García

Universitat Politècnica de València
Departament de Sistemes Informàtics i Computació
Camí de Vera s/n 46020 València (Spain)
{fpla,amolina,esanchis,esegarra,fgarcia}@dsic.upv.es

Abstract. Over the last few years, stochastic models have been widely used in the natural language understanding modeling. Almost all of these works are based on the definition of segments of words as basic semantic units for the stochastic semantic models.
In this work, we present a two–level stochastic model approach to the construction of the natural language understanding component of a dialog system in the domain of database queries. This approach will treat this problem in a way similar to the stochastic approach for the detection of syntactic structures (Shallow Parsing or Chunking) in natural language sentences; however, in this case, stochastic semantic language models are based on the detection of some semantic units from the user turns of the dialog. We give the results of the application of this approach to the construction of the understanding component of a dialog system, which answers queries about a railway timetable in Spanish.

1 Introduction

Language Understanding systems have many applications in several areas of Natural Language Processing. Typical applications are train or plane travel information retrieval, car navigation systems or information desks. In the last few years, many efforts have been made in the development of natural language dialog systems which allow us to extract information from databases. The interaction with the machine to obtain this kind of information requires some dialog turns. In these turns, the user and the system interchange information in order to achieve the objective: the answer of a query made by the user. Each turn (a sequence of natural language sentences) of the user must be understood by the system. Therefore, an acceptable behavior of the understanding component of the system is essential to the correct performance of the whole dialog system.

There are different ways to represent the meaning of natural language sentences in the domain of database queries. One of the most usual representations in tasks of this kind is based on frames. In a frame, we can represent a user turn of the dialog as a concept (or a list of concepts) and a list of constraints made over this concept.

Many works in the literature use rule-based techniques in order to obtain the translation of a user turn into the corresponding frame. They are mainly based on the detection of keywords, which characterize the semantic constituents of the sentences. Other approaches are based on statistical modeling. Over the last few years, stochastic models, which estimate the models automatically from data, have been widely used in natural language understanding modeling [6] [12] [7] [14].

Almost all of these works are based on the definition of segments of words as basic semantic units for the stochastic semantic models. In most of them, the definition of classes of words is necessary in order to obtain high coverage models from the given data, (The problem of the lack of training data is always present when automatic learning techniques are used). This approach to the natural language understanding problem presents a strong parallelism with the stochastic approach applied in recent years [2][8][9] to the problem of tagging texts, when the objective is not only to associate POS tags to words but to detect some syntactic structures such as NP, VP, PP, etc. In the first case, the segments represent semantic units and, in the second one, they represent syntactic units.

In this work, we present a two–level stochastic model approach to the construction of the natural language understanding component of a dialog system in the domain of database queries. This approach will treat this problem in a way similar to the stochastic approach for the detection of syntactic structures (Shallow Parsing or Chunking) in natural language sentences. However, in this case, stochastic semantic language models are based on the detection of some semantic units from the user turns of the dialog. We describe the application of this approach to the construction of the understanding component of a dialog system, which answers queries about railway timetable in Spanish.

2 System Overview

The dialog system has been developed in the BASURDE project [1], and it follows a classic modular architecture. The input of the understanding module is a sequence of words, and the output is the semantic representation of the sentence (one or several frames). This output constitutes the input to the dialog manager, which will generate the corresponding answers following the dialog strategy. The knowledge sources used by the understanding module are the syntactic and semantic models, and, optionally, the dialog act predicted by the dialog manager.

The semantics of input sentences is represented by frames [13], like in other dialog systems [5]. Each type of frame has a list of cases associated to it. These cases, which will be filled through the understanding process, are the constraints given by the user. We consider two types of frames: *task–dependent*, if the sentence is a complete or incomplete query, or a confirmation of some information and *task–independent*, if the sentence is an affirmation, negation, courtesy, etc.

Each sentence can be represented by one or more frames. In the railway timetable task, a list of 11 types of frames and 22 cases were defined. An example of an input sentence and the corresponding frames are:

```
INPUT SENTENCE: Yes. I would like to know the price and the type
                of the train that leaves at 23 hours and 5 minutes.
OUTPUT FRAMES:
                (AFFIRMATION)
                (PRICE)
                DEPARTURE-TIME: 23:05
                (TYPE)
                DEPARTURE-TIME: 23:05
```

The understanding process consists of two phases, (see Figure 1). In the first phase, the input sentence is sequentially transduced into a sentence of an intermediate semantic language [11], and, in the second phase, the frame or frames associated to this sentence are generated. The models used for the sequential transduction are stochastic models, which are automatically learnt from training samples. Some rules are applied in the second phase in order to obtain the corresponding frames from the semantic sentence.

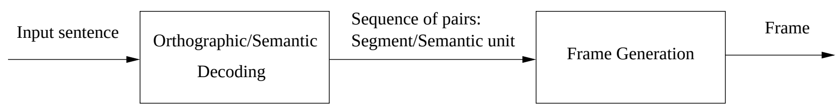


Fig. 1. General Description of the Understanding Process

The intermediate semantic language is defined over a vocabulary of semantic units. Each semantic unit stands for a specific meaning. Due to the fact that the defined semantic language is sequential with the input language, we can perform a segmentation of the input sentence into a number of intervals which is equal to the number of semantic units in the corresponding semantic sentence. An example of segmentation of an input sentence is:

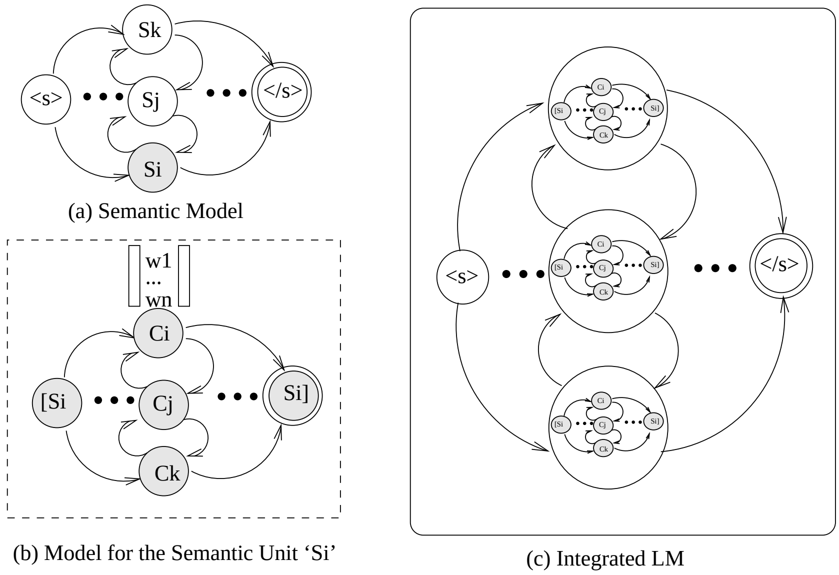
Input sentence:
me podría decir los horarios de trenes para Barcelona
(can you tell me the railway timetable to Barcelona)

Spanish		English	
u_1= me podría decir	v_1=consulta	u_1= can you tell me	v_1= query
u_2= los horarios de trenes	v_2=<hora_salida>	u_2= the railway timetable	v_2=<departure_time>
u_3=para	v_3=marcador_destino	u_3= to	v_3=destination_marker
u_4=Barcelona	v_4=ciudad_destino	u_4= Barcelona	v_4=destination_city

3 General Description of the Two-level Stochastic Models

In order to implement the first phase, we propose an approach based on a two-level stochastic model. This model combines different knowledge sources at two different levels. The top level models the intermediate semantic language, that is, the set of sequences of semantic units. The lower level represents the internal structure of each semantic unit in terms of the linguistic units considered (words, POS, lemmas, etc.). The formalism [9] that we use for the models in the two levels is finite–state automata. To be exact, we use models of bigrams which are smoothed using the back–off technique [4] in order to achieve full coverage of the language considered. The bigram probabilities are obtained by means of the SLM TOOLKIT [3] from the sequences of different units in the training set. They are then represented as finite-state automata.

All these models are estimated from a corpus of sentences which are segmented in semantic units. From this training set, we learn the following models:



(a) Semantic Model

(b) Model for the Semantic Unit 'Si'

(c) Integrated LM

Fig. 2. Integrated Language Model

– The Semantic Language Model (Figure 2 (a)), that is, a bigram model which represents the concatenation of semantic units (semantic units are associated to states in the figure). This model is learnt from the sequences of semantic units associated to each sentence of the training set. The sequence of semantic units in the above example is:

  consulta <hora_salida> marcador_destino ciudad_destino
  (query <depart_time> destination_marker destination_city)

– The Models for the Semantic Units. The structure of each semantic unit can be established directly in terms of words (word bigram models). This approach produces models that are very big and very dependent on the vocabulary of the application. For this reason, we propose an alternative method based on POS tags. To do this, we use a Spanish POS tagger [9] which supplies the corresponding POS tag for every word. In this situation, we obtain a new training set annotated with morphological information. For each semantic unit we learn a HMM in which the states represent POS tags and the words are emitted from them according to certain lexical probability (Figure 2 (b)). This HMM is estimated from the segments of POS associated to this semantic unit.

Once the different models have been learnt, a regular substitution of the lower models into the upper one is made. In this way, we get a single integrated model (Figure 2 (c)) which shows the possible concatenations of semantic units and their internal structure. This integrated model includes the transition probabilities as well as the lexical probabilities.

In order to obtain a more accurate modelization of the semantic units, we used a technique to enrich the HMM [10]. This technique consists of incorporating new categories to the POS tag set. These new categories are strongly related to some selected words, which can be established empirically from the training set or following other criteria. We obtain lexicalized models with this process. Although this lexicalization produces more complex models, semantic units models are improved. For instance, if we lexicalize the prepositions ‘to’ and ‘from’ we can distinguish between two strongly different meanings in the railway timetable task.

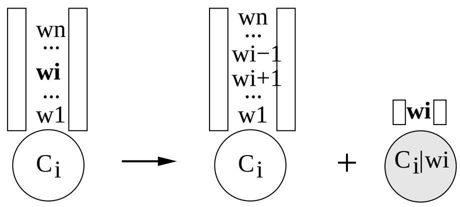


Fig. 3. An example of a Lexicalized State

In Figure 3, we show the effect of this lexicalization over a generic state C_i, belonging to a certain syntactic unit, when it is particularized for a certain word w_i. In this way, we obtain a new state (filled state in the figure) in which only the word w_i can be emitted with lexical probability equal to 1.

Our system can be considered as a two-level transducer. The upper level describes contextual information about the structure of the sentences, and the lower level modelizes the structure of the semantic units considered.

The Understanding process consists of finding out the sequence of states of maximum probability on the integrated model for an input sentence. Therefore, this sequence must be compatible with the contextual and lexical constraints. This process can be carried out by Dynamic Programming using the Viterbi algorithm, which we adapted to our models. From the Dynamic Programming trellis, we can obtain the best segmentation of the input sentence into semantic units.

4 Experimental Results

The language understanding model obtained was applied to an understanding task which was integrated into a spoken dialog system answering queries about a railway timetable in Spanish [1].

The corpus consisted in the orthographic transcription of a set of 215 dialogs, obtained through a Wizard of Oz technique. Using only the user utterances, we defined a training set of 175 dialogs with 1,141 user utterances, and a test set of 40 dialogs with 268 user utterances. The number of words in these two sets was 11,987 and the medium length of the utterances was 10.5 words.

We used several measures to evaluate the accuracy of the models:

- The percentage of correct sequences of semantic units (*%cssu*).
- The percentage of correct frames (*%cf*).
- The precision (*%P*), that is, the rate between the number of correct proposed semantic units and the number of proposed semantic units.
- The recall (*%R*), that is, the rate between the number of correct proposed semantic units and the number of semantic units in the reference.
- The score $F_{\beta=1} = \dfrac{2 \times P \times R}{P + R}$, which combines the last two rates (*%P* and *%R*).

We evaluated the segmentation accuracy and the correct interpretation of the user utterances using these measures.

Table 1. Experimental results

Models	% cssu	% cf	% P	% R	$F_{\beta=1}$
BIG-BIG	32.3	41.0	55.9	51.0	53.3
BIG-BIG-word	58.7	67.3	78.9	79.2	79.0
BIG-BIG-lemma	59.9	72.5	79.6	81.0	80.3

In Table 1, we show the experimental results with three approaches: a two-level approach using POS bigram models in the two levels (BIG-BIG), a two-level approach using POS bigram models which were lexicalized taking into account the most frequent words (BIG-BIG-words), and a two-level approach using POS bigram models which were lexicalized taking into account the most frequent lemmas (BIG-BIG-lemmas). In the third case, the lexicalized units were lemmas instead of words. These lemmas were obtained through a morphological analyzer.

On one hand, it can be observed from Table 1 that there is a big difference between the *%cssu* and the *%cf* measures. This difference is due to the fact that, although the obtained semantic sentence is not exactly the same as the reference semantic sentence, their corresponding frame is the same.

On the other hand, we observed that the best performance was achieved using lexicalized models. That is because the lexicalization process gave a more accurate relation between words and semantic units. In the BIG-BIG-word model, the best results were achieved taking into account the words whose frequency in training data was larger than 9. In the BIG-BIG-lemma model, we considered lemmas instead of words in order to specialize the model. This kind of specialization slightly improves the performance of the system.

In [14], some results on the same corpus are presented using word bigram models. Although this approach gave a slightly better performance than our approach, models based on categories are more independent from the task than models based on words.

5 Conclusions and Future Work

We have presented an approach to Natural Language Understanding based on Stochastic Models, which are automatically learnt from data. In particular, we have used some techniques from the areas of POS tagging and Shallow Parsing.

The evaluation has been done on a task of language understanding in a Spanish dialog system which answers queries about a railway timetable. Considering that the available data was small, the results were relatively reliable.

In any case, from the experimental results, we observed that the best performance was achieved using lexicalized models. That is because the lexicalization process gives a more accurate relation between words and semantic units. We hope that a more appropriate definition of the list of words/lemmas to be lexicalized will provide better performance.

On the other hand, the incorporation of contextual knowledge from the dialog manager, that is, the prediction of the next dialog act, could improve the models.

Acknowledgments

This work has been supported by the Spanish Research Projects CICYT TIC2000–0664–C02–01 and TIC98-0423–C06–02

References

1. A. Bonafonte, P. Aibar, N. Castell, E. Lleida, J.B. Mari no, E. Sanchis, and M.I. Torres. Desarrollo de un sistema de diálogo oral en dominios restringidos. In *Proceedings of I Jornadas en Tecnologa del Habla*, 2000.

2. T. Brants. Cascaded Markov Models. In *Proceedings of the EACL99*, Bergen, Norway, 1999.
3. P. Clarksond and R. Ronsenfeld. Statistical Language Modeling using the CMU-Cambridge Toolkit. In *Proceedings of Eurospeech*, Rhodes, Greece, 1997.
4. S. M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35, 1987.
5. L. Lamel, S. Rosset, J.Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts. The LIMSI ARISE System. *Speech Communication*, 31(4):339–353, 2000.
6. E. Levin and R. Pieraccini. Concept–Based Spontaneous Speech Understanding System. In *Proceedings of EUROSPEECH'95*, pages 555–558, 1995.
7. W. Minker. Stocastically–Based Semantic. Analysis for ARISE – Automatic Railway Information Systems for Europe. 2(2):127–147, 1999.
8. F. Pla, A. Molina, and N. Prieto. Tagging and Chunking with Bigrams. In *Proceedings of the COLING–2000*, Saarbrücken, Germany, August 2000.
9. F. Pla, A. Molina, and N. Prieto. An Integrated Statistical Model for Tagging and Chunking Unrestricted Text. In *Proceedings of the Text, Speech and Dialogue 2000*, Brno, Czech Republic, September 2000.
10. F. Pla, A. Molina, and N. Prieto. Improving Chunking by means of Lexical-Contextual Information in Statistical Language Models. In *Proceedings of 4th CoNLL-2000 and LLL-2000*, Lisbon, Portugal, September 2000.
11. E. Sanchis, E. Segarra, M. Galiano, F. García, and L. Hurtado. Modelización de la Compresión mediante Técnicas de Aprendizaje Automático. In *Proceedings of I Jornadas en Tecnologa del Habla*, 2000.
12. R. Schwartz, S. Miller, D. Stallard, and J. Makhoul. Language Understanding using hidden understanding models. In *ICSLP*, pages 997–1000, 1996.
13. E. Segarra, V. Arranz, N. Castell, I. Galiano, F. García, A. Molina, and E. Sanchis. Representación Semántica de la Tarea. In *Internal Report UPV DSIC–II/5/00*, 2000.
14. E. Segarra, E. Sanchis, M. Galiano, F. García, and L. Hurtado. Extracting Semantic Information through Automatic Learning Techniques. In *IX Spanish Symposium on Pattern Recognition and Image Analysis-SNRFAI'01*, 2001.