Applying dialogue constraints to the understanding process in a Dialogue system

Emilio Sanchis, Fernando García, Isabel Galiano, Encarna Segarra

Departamento de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia (UPV),
Camino de Vera s/n, 46022 Valencia, Spain
{esanchis,fgarcia,mgaliano,esegarra}@dsic.upv.es

Abstract. In this paper, we present an approach to the estimation of a dialogue-dependent understanding component of a dialogue system. This work is developed in the framework of the BASURDE Spanish dialogue system, which answers queries about train timetables by telephone in Spanish. Modelization which is specific to the dialogue state is proposed to improve the behaviour of the understanding process. Some experimental results are presented.

1 Introduction

The construction of dialogue systems applied to limited domain information systems is an important objective in the area of Human Language Technologies. The advance in the design and analysis of the different knowledge sources involved in a spoken dialogue system, such as speech processing, language modeling, language understanding, or speech synthesis, has led to the development of dialogue system prototypes. Some characteristics of these systems are : telephone access, limited semantic domains and mixed initiative [6][4][1].

The work that we present in this paper is an approach to the construction of the understanding module of the BASURDE dialogue system [2]. The system's task consists in answering telephone queries about railway timetables in Spanish. In this system, the output of the speech recognizer is the input of the understanding module, which in turn supplies its output to the dialogue manager.

The representation of the meaning of the user utterances is made through the semantic-frames. The frame determines the type of communication of the user turn as well as the data supplied in the utterance.

There are two kinds of classical approaches to the problem of language understanding: the first is based on the use of rules to detect markers and keywords to obtain the frame type and its attributes; the second one is based on the use of models which are automatically learnt from samples. Some of these are based on stochastic models (HMM and Stochastic Regular Grammars) that can be automatically learnt by means of Grammatical Inference techniques [3][10][8].

One advantage of using models and learning techniques of this kind, is that it allows us to adapt to new tasks and situations, such as reestimation with new data, context changes, new tasks or different languages. However, one drawback is that it is necessary to have a large amount of training data, which is especially difficult to obtain in the case of dialogues.

In the BASURDE system the representation of the dialogue structure is done by means of a stochastic network of dialogue acts. One advantage of this structure is that it gives a prediction of the next dialogue acts expected from the user. In this work, this information is used in the understanding process. In particular, different understanding models are used depending on the last dialogue act of the system. Other approaches of dialogue-dependent understanding models have been recently investigated [5].

2 The BASURDE task

The BASURDE task [2] consists of telephone queries about the Spanish train timetables. The kind of queries (semantic restrictions) are: questions about timetables, prices, and services, for long distance trains. A corpus of 200 person-to-person dialogues corresponding to a real information system were recorded and analyzed. Then, four types of scenarios were defined (departure/arrival time for a one-way trip, a two-way trip, the prices and services, and one free scenario). A total of 215 dialogues were acquired using the Wizard of Oz technique. The total number of user turns was 1460 (14.902 words).

3 Dialogue act labels

One method of representing the structure of a dialogue is through dialogue acts. To do this, a set of labels must be defined. The number of labels must be large enough to show the different intentions of the turns and to cover all the situations. If the number is too high the models will be underestimated because there aren't enough training samples. On the other hand, if we define a small set of just a few labels, only general purposes of the turn can be modeled. In the BASURDE system a three-level label set was proposed [7]. The first level of each dialogue act describes the dialogue behavior. These labels are generic for any task. The second level is related to the semantic representation of a sentence and is specific to that task. The third level takes into account the data given in the utterance.

A stochastic network of dialogue acts can be learnt from training samples by using these labels. This network represents the dialogue structure and strategy and can be used by the Dialogue Manager to generate the dialogue acts of the system. It can also help in the recognition and the understanding process by means of the generation of hypotheses of the expected acts of the user.

We will center our interest on the first-level, which is used to direct the understanding process. The first level can take the following values:*Opening, Closing, Undefined, Not_understood, Waiting, New_query, Acceptance, Rejection, Question, Confirmation, Answer.*

An example of some turns annotated with the first-level labels is shown in Fig 1. Note that each turn can have more than one label associated to it.

S1: Bienvenido al sistema automático de información de trenes regionales y de largo recorrido, qué desea? (S:*Opening*)
(Welcome to the information system for train timetables. What information would you like?)
U1: Puede decirme a qué hora salen los trenes de Valencia a Barcelona? (U:*Question*)
(Can you tell me what time the trains from Valencia to Barcelona leave?)
S2: De Valencia a Barcelona,(S:*Confirmation*) qué dia quiere salir? (S:*Question*)
(From Valencia to Barcelona, what day do you want to leave?)
U2: El próximo jueves (U:*Answer*)
(Next Thursday)

Fig. 1. Example of a labelled dialogue

4 Semantic representation

In our system, the representation of the meaning of user turns is done using semantic frames; i.e. each frame represents a semantic concept and it can have some attributes and values associated to it. The set of frames defined for this task is the following: *Departure_time, Return_departure_time, Arrival_time, Return_arrival_time, Price, Return_price, Length_of_trip, Train_type, Services, Confirmation, Not_Understood, Affirmation, Rejection, Closing.*

As we have mentioned above, a set of attributes can be associated to each frame. For example, the sentence:

Cuál es el precio de un billete de Valencia a Barcelona?
(How much is a ticket from Valencia to Barcelona)

is represented as follows:

(PRICE)
Destination: Barcelona
Origen: Valencia

5 Categorization

As the number of instances of some attributes can be very low, problems of coverage and lack of training samples can occur when learning stochastic models from a corpus. Therefore, we had to reduce the number of words in our lexicon by using lemmas and categories. Due to the complex conjugation of Spanish verbs we had to substitute any conjugated form of a verb with its corresponding infinitive form. We also had to substitute any instance of a city name with the category City-Name; any instance of day of the week with the category Day-Name; and the same with numbers, months, etc. We defined seven categories in the lexicon. This way, we reduced the size of our lexicon from 637 to 370 different words.

6 The Understanding process

The understanding process is done in two phases. The first phase consists of a transduction of the input sentence in terms of an intermediate semantic language. As this intermediate semantic language that we defined is sequential with the input, some sequential transduction techniques can be used. 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 in order to construct the frame. An example of the actions in this second phase is the conversion of relative dates and hours into absolute values. For example ”next Monday” into ”mm/dd/yr” or ”in the morning” into ”interval hours (from 5 to 12)”. The first phase of the understanding process is based on automatically learnt stochastic models. We have defined a set of 53 semantic units, which are a kind of semantic categorization. Each semantic unit represents the meaning of words (or sequences of words) in the sentences. For example, the semantic unit ”consult” 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. An example is shown:

Spanish

w_1: por favor	v_1: consulta
w_2: a que hora salen los trenes?	v_2: <hora_salida>
w_3: hacia	v_3: marcador_destino
w_4: Alicante	v_4: ciudad_destino

English

w_1: please	v_1: query
w_2: what is the railway timetable?	v_2: <Departure_time>
w_3: to	v_3: destination_marker
w_4: Alicante	v_4: destination_city

The semantic sentence V for the semantic language model training is :

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

The stochastic modelization is divided into two levels. The higer level (semantic model) represents the sequences of semantic units allowed. In the lower level (semantic-unit models) the language (sequences of words) associated to each semantic unit is modelled. An annotated training corpus is used to obtain the stochastic models. Each sentence of the training set consists of the sequence of words, the sequence of semantic units and the associated segmentation.

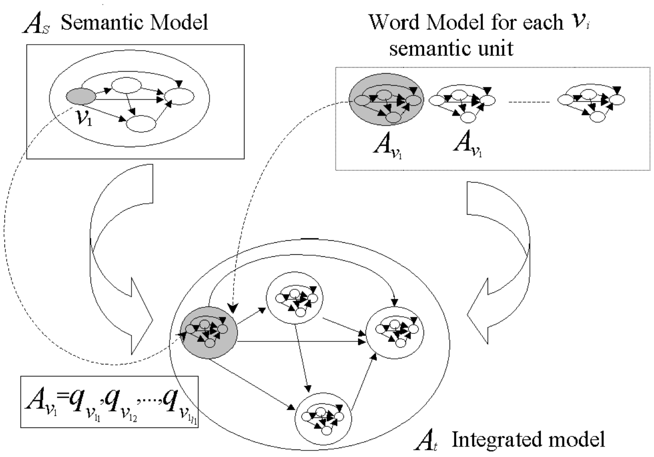
A_S Semantic Model
Word Model for each v_i semantic unit
v_1
A_{v_1}
A_{v_1}
$A_{v_1}=q_{v_{11}},q_{v_{12}},...,q_{v_{1|1}}$
A_t Integrated model

Fig. 2. *Scheme of the two-level approach.*

Let T be a training set of pairs (W,V) of sequences of words and semantic units. We learn a model $A_{\rm s}$ for the semantic language $L_{\rm s} \subseteq \mathcal{V}^*$, and a set of models (one for each semantic unit $v_{\rm i} \in \mathcal{V}$, where $\mathcal{V}$ is the set of semantic units). The regular $A_{\rm s}$ for the semantic language $L_{\rm s}$ is estimated from the semantic sentences $V \in \mathcal{V}^*$ of the training sample. The regular model $A_{\rm v_i}$ for each semantic unit $v_{\rm i} \in \mathcal{V}$ is estimated from the set of segments, $w_{\rm i}$, of the training sample associated to the semantic unit $v_{\rm i}$. These estimations are made through automatic learning techniques.

In this work, a classical bigram model was estimated for the semantic model $A_{\rm s}$, and the models for the semantic units, $A_{\rm v_i}$, were also estimated as classical bigrams. However, other techniques can be used for the estimation of both models. This is the case of grammatical inference techniques such as the ECGI, MGGI, or k-testable in strict sense [9][10].

In our approach the understanding process is as follows: Given the sequence of words of the input sequence $W = w_1w_2\dots w_n$, the process consists of finding the sequence of semantic units $V = v_1v_2\dots v_k$ which maximizes the probability:

$$\widehat{V} = \operatorname*{argmax}_V P(W|V)P(V)$$

The term $P(W|V)$ is the probability of the sequence of words W given the sequence of semantic units V. We approach this probability, following the Viterbi criterium, as the maximum for all posible segmentations of W in $|V|$ segments.

$$P(W|V) = \max_{\forall l_1, l_2, \ldots l_{k-1}} \{P(w_1, ..., w_{l_1}|v_1) \cdot P(w_{l_1+1}, ..., w_{l_2}|v_2) \cdot \ldots \cdot P(w_{l_{k-1}+1}, ..., w_n|v_k)\}$$

where the probability of each segment is done by means of bigram probabilities of words given the associated semantic unit:

$$P(w_i, ..., w_j|v_s) = P(w_i|v_s) \prod_{k=i+1}^{j} P(w_k|w_{k-1}, v_s)$$

The term $P(V)$ is the bigram probability of the sequence V.

$$P(V) = P(v_1) \prod_{i=2}^{k} P(v_i|v_{i-1})$$

The understanding process is done through the Viterbi algorithm, which supplies the best path in the integrated model (Fig 2.). This path gives not only the sequence of sematic units but also the segmentation associated to it.

In our approach, a specific model for each dialogue act is obtained. That is, we split the training samples into 6 subsets, corresponding to the labels: *Opening, Confirmation, Waiting, New_query, Not_understood, Answer.* Each subset is associated to a first-level dialogue label, and it contains the user turns which occur after its label. For example, the set *Opening* contains all the user turns which have been uttered after the system has generated the *Opening* dialogue act. In this way, we hope to have a more specific modelization of the user turns.

We only apply the specific modelization of the higher level (semantic model A_s). This is because this level represents the semantics of the sentence, while the lower level represents the specific instantiation of this semantics in terms of sequences of words. Therefore, we can take advantage of all the samples to learn the semantic unit models (A_{v_i}).

The understanding model selected by the Dialogue Manager in this decoding process is only the specific model, which is associated to the last dialogue act produced by the system.

7 Experimental results and conclusions

In order to study the appropriateness of Specific Language Understanding models (SLU), a preliminary experimentation on the BASURDE task was conducted. The obtained results were then compared with the results for the same task but using a unique Language Understanding model (LU). In this section, we present the results of these two types of models for the same task. At the light of these results, some conclusions will be drawn.

The corpus consisted of the orthographic transcription of a set of 215 dialogs, which were obtained using the Wizard of Oz technique. These dialogues contained 1,460 user turns which were our experimental set. For this set, a cross-validation procedure was used to evaluate the performance of our models. The experimental set was randomly split into five subsets of 292 turns. One of our experiments then consisted of five trials, each of which had a different combination of one subset taken from the five subsets as the test set, with the remaining 1,168 turns being used as the training set.

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 frames (%cf); i.e., the percentage of resulting frames that are exactly the same as the corresponding reference frame.
- the percentage of correct frame slots (frame name and its attributes) (%cfs).

The experimental results (%cssu, %csu, %cf and %cfs) obtained by using LU and SLU models are shown in Table 1.

Table 1. Experimental Results

	%cssu	%csu	%cf	%cfs
LU	68.1	87.6	80.84	89.1
SLU	69.6	88.1	81.9	89.3

From Table 1, it can be concluded that using SLU models helps to slightly improve the understanding process.

In the light of the results obtained, we can conclude that using Specific Language Understanding Models is a good way to implement the required feedback between the dialogue manager and understanding modules.

More works should be done to improve feedback. For instance, SLU and LU model can be interpolated in order to avoid the lack of coverage of specific models. It can also be defined new sets of specific models.

References

1. Cmu communicator spoken dialog toolkit (csdtk). http://www.speech.cs.cmu.edu/communicator/.
2. A. Bonafonte, P. Aibar, N. Castell, E. Lleida, J.B. Mariño, E. Sanchis, and M.I. Torres. Desarrollo de un sistema de diálogo oral en dominios restringidos. In *I Jornadas en Tecnología del Habla, Sevilla (Spain)*, 2000.
3. H. Bonneau-Maynard and F. Lefèvre. Investigating stochastic speech understanding. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2001.

4. J. Glass and E. Weinstein. Speech builder: facilitating spoken dialogue system development. In *Proc. in EUROSPEECH*, volume 1, pages 1335–1338, 2001.
5. K Hacioglu and W. Ward. Dialog-context dependent language modeling combining n-grams and stochastic context-free grammars. In *Proc. of ICASSP*, 2001.
6. L. Lamel, S. Rosset, J.L. Gauvain, Bennacef S., Garnier-Rizet M., and B. Prouts. The limsi arise system. *Speech Communication*.
7. C. Martinez, E. Sanchis, F. García, and P. Aibar. A labeling proposal to annotate dialogues. In *Proc. of third International Conference on Language Resources and Evaluation (LREC))*.
8. F. Pla, A. Molina, Sanchís E., and García F. Language understanding using two-level stochastic models with pos and semantic units. In Lecture Notes, editor, *Proc in 4th International Conference TSD*, number 2166, pages 403–409, 2001.
9. E. Segarra and L. Hurtado. Construction of Language Models using Morfic Generator Grammatical Inference MGGI Methodology. In *Proc. of EUROSPEECH*, pages 2695–2698, 1997.
10. E. Segarra, E. Sanchis, F. García, and L.F. Hurtado. Extracting semantic information through automatic learning. In *Proc. of IX Spanish Symposium on Pattern Recognition and Image Analysis (AERFAI)*, 2001.