A Statistical Segment-Based Approach for Spoken Language Understanding

Lucía Ortega, Isabel Galiano, Lluís-F. Hurtado, Emilio Sanchis, Encarna Segarra

Department de Sistemes Informàtics i Computació, Universitat Politècnica de València, València, Spain
{lortega, mgaliano, lhurtado, esanchis, esegarra}@dsic.upv.es

Abstract

In this paper we propose an algorithm to learn statistical language understanding models from a corpus of unaligned pairs of sentences and their corresponding semantic representation. Specifically, it allows to automatically map variable-length word segments with their corresponding semantic units and thus, the decoding of user utterances to their corresponding meanings. In this way we avoid the time consuming work of manually associate semantic labels to words, process which is needed by almost all the corpus-based approaches. We use the algorithm to learn the understanding component of a Spoken Dialog System for railway information retrieval in Spanish. Experiments show that the results obtained with the proposed method are very promising, whereas the effort employed to obtain the models is not comparable with that of manually segment the training corpus.

Index Terms: Spoken Language Understanding, Semantic Classification, Statistical modelization

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. In the last years, there have been many achievements in the development of dialog systems. Generally, most of these systems are designed for tasks that are semantically restricted, that have a medium-size vocabulary, and that run in a mixed initiative framework where the interaction is done in spontaneous speech.

Corpus-based approaches that are based on the use of statistical models have widely used for Language Understanding (LU) [1]. Some examples of the use of Hidden Markov Models to model the understanding process can be seen in [2] and [3]. There are also other statistical approaches based on classification, transduction and grammatical inference techniques [4], [5], [6], [7], and [8].

In most of the mentioned works, the meaning of the sentences is represented by frames consisting of concepts and attribute-value pairs; although there are some works using a more hierarchical semantical representation [9].

The advantages of using statistical semantic models for LU are well-known: they can represent the variability of expressing meanings in terms of segments of words, and they can be automatically learnt through the analysis of a large training corpus. However, at the same time, to collect and prepare such a large and representative corpus is perhaps the major drawback of this approach: it requires the definition of the semantic units (concepts, markers, or attributes) that will be used to represent the semantic meaning of the sentences and, later on, a manual or semi-automatic process of semantic labeling in which several human experts associate word segments to their corresponding semantic meanings (or in some cases other morpho-syntactic labels). Furthermore, the manual labeling requires the participation of different labelers, and this implies the possibility of applying different criteria in similar situations. This possibility of heterogeneous criteria can be a source of errors in the posterior processes of learning and decoding. It is also possible that the manual analysis could not detect the best segments that can be associated to the semantic units.

In this paper we present an approach that tries to minimize these problems by automatically learn a map among variable-length word segments and their corresponding semantic units and thus, the decoding of word sentences into their corresponding meanings. In order to do so, it applies an iterative process of classification. The training corpus consist of sentences and their corresponding semantic representation, that is a sequence of semantic units (not necessary sequential with the sentence). Therefore, the corpus does not include the association of words to semantic units; in fact at the beginning of the process each word of a sentence can be associated to each one of the semantic units in the semantic representation of the sentence. The proposed method starts from this elemental mapping and, through the statistical analysis of the correlations among words and semantic units detects the word segments of variable length that can be representatives of each one of the semantic units. This approach has been applied to the understanding module of a dialog system for railway information retrieval.

In Section 2 the semantic modelization and the algorithms used for learning and decoding are presented. Section 3 presents the evaluation of the algorithm on the Semantic Corpus of Dihana, a Spoken Dialog System to access a railway information system using Spontaneous Speech in Spanish. Finally in Section 4 the Conclusions and future works.

2. The statistical LU model

The semantic decoding process can be viewed as follows:

Given a sentence, \( w = w_1^{N} \), uttered by a user the model predicts its intended meaning, represented as a sequence of semantic units (we will call concepts, from now on) \( c = c_1^M \), by maximizing the conditional probability \( P(c|w) \). By using the well-known Bayes rule we have:

\[
\hat{c} = \arg \max_{c} P(w|c) \cdot P(c)
\]

where \( P(c) \) is the priori probability of the sequence of concepts \( c \) and \( P(w|c) \) is the probability of the sequence of words \( w \) given the sequence of concepts \( c \), a kind of mapping function among variable-length segments of words and concepts.
Therefore, our statistical LU model requires a method for computing \( P(c) \) and \( P(w|c) \) probabilities, and a search algorithm which finds \( c \) over all possible sequences of concepts \( c \).

### 2.1. The statistical modelization proposal

The main characteristics of our statistical modelization proposal are the following:

- The model is estimated from an unaligned set of pairs: word sequence \( w \), concept sequence \( c \) associated with \( w \). This means that, apart from the semantic representation of each sentence, no labeling process at all has been done in order to associate words or word segments to their corresponding concepts.

- Moreover, if the semantics of the sentence is given in a canonical form, then the concept sequence is not necessarily sequential with the sentence.

- Given that the order in which the information is supplied in sentences is not a relevant factor in determining its meaning, we approach the priori probability \( P(c) \) using the unigram probability of the concepts estimated from the training corpus (\( P(c_i) \)), that is \( P(c^N_i) = \prod_{i=1}^{N} P(c_i) \); in other words, two different sequences of concepts are equivalent if they contain the same set of concepts.

- Because of the lack of alignment between the training pairs, we can not estimate by counting the probabilities \( P(w|c) \). To do this, we use an iterative classification process in which variable-length segments of words appearing in a sentence are assigned to the concept that, from an statistical and semantic point of view, represents them in the best way. Therefore, \( P(w|c) \) can be seen as the likelihood of one of the possible segmentations of \( w \) in a sequence of word-segments of variable length. This probability can be approached as the maximum for all possible segmentations of \( w^N \) in \( M \) segments.

\[
P(w^N|c^M) = \max_{l_1, l_2, \ldots, l_{m-1}} \{ P(w_1, \ldots, w_1|c_1) \cdot P(w_{l_1+1}, \ldots, w_{l_2}|c_2) \cdot \ldots \cdot P(w_{l_{m-1}+1}, \ldots, w_N|c_M) \}
\]

- From the definition of this objective function, the semantic decoding process can be done through a classical dynamic programing algorithm able to deal with variable-length segments.

### 2.2. The learning process

The learning corpus consists of an unaligned set of pairs: word sequence, concept sequence. An example of the semantic representation of the corpus is shown in Figure 1. The order in which the information is supplied by the user is not a relevant factor because it is possible to say the same in a different order. An example of this is shown in Figure 2. Examples are extracted from DIHANA corpus (see Section 3.1) where semantic representation (frames) is done in a predefined canonical order.

The learning process consists of an iterative procedure that obtains the sequences of words of lengths from 1 to \( l_{\text{max}} \) that can be associated to the concepts. This association is done using discriminative criteria.

Four main steps are performed in each iteration of the learning algorithm. The iterative process starts considering segments of length 1 (words) and increase the length by one until \( l_{\text{max}} \), \( \forall l \in \{1, \ldots, l_{\text{max}}\} \):

1. As there is not information about explicit correlation between segments and concepts, in this first step each segment is assigned to all the concepts that appear in the semantic representation of the sentence. From this information a set of word segments associated to each concept is obtained.

2. In order to increase the capability of the segments to discriminate among concepts, a process of refinement of this first association is done. A segment is considered relevant for a concept if it appears frequently with this concept and it do not appears frequently with others concepts. Therefore, a pruning of the initial sets is done by using a threshold: only the segments that have a high probability in a concept and low probability in other concepts are maintained in the set of segments associated to each concept, and are deleted from the other sets.

   To do that, \( P(c_i|s_i) \) is calculated \( \forall s_i, c_i \), where \( c_i \) is a concept, \( s_i \) is a segment of length \( l \) and \( P(c_i|s_i) \) is the probability of associating the concept \( c_i \) to segment \( s_i \). From these probabilities, segments \( s_i \) for which \( P(c_i|s_i) < \text{\textit{threshold}} \) are removed from the set associated with concept \( c_i \).

3. In this step a process of categorization based on linguistic criteria can be done in order to increase the coverage of the model. In this way, some lexical realizations of certain well-known categories can be considered although they did not appear in the training corpus. Examples of this kind of categorization are months, numbers,
or cities. Another kind of linguistic categorization can be done by substituting words by lemmas. In all this cases, it is necessary to control that this categorization process does not generate any ambiguity. That is, a word relevant to a concept can be substitute by its category only if there are not other words in the same category that are relevant to another concept.

4. In order to avoid ambiguity when increasing the length of the segments a pruning procedure is applied. For each concept, those segments that are composed by shorter segments that belong to other concepts are pruned.

The goal of this discriminative procedure is to find appropriate segments of different lengths that characterize the semantic concepts. It should be noted that, when increasing the length of the segments we can better discriminate between words that are semantically ambiguous if they are considered in an isolated way. This is for example the case of the word "Valencia" that can be associated to Origin and Destination, but when considering segments of length 2 the sequence "to Valencia" will be clearly assigned to Destination.

### 3. Experimental results

In this section, we describe the results of the evaluation of our understanding system with the DIHANA corpus [10].

#### 3.1. The DIHANA corpus

The DIHANA task consists of queries by telephone about railway timetables and prices in Spanish.

A set of 900 dialogs was acquired in the DIHANA project by means of the Wizard of Oz technique. Three types of scenarios were defined:

- Timetables for a one-way trip.
- Timetables for a two-way trip.
- Timetables, prices, and services.

The number of users was 225 with 4 dialogs per user. The characteristics of the transcribed corpus are shown in Table 1.

<table>
<thead>
<tr>
<th>Number of user turns</th>
<th>6 226</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of words</td>
<td>47 222</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>811</td>
</tr>
<tr>
<td>Average number of words per user turn</td>
<td>7.6</td>
</tr>
<tr>
<td>Duration of the recording (hours)</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of the transcribed corpus.

Once the corpus was transcribed, the set of semantic concepts was defined, and the corpus was labeled in terms of frames (sequence of concepts and attribute-value pairs in a canonical order). It should be noted that in this work we have called concepts both frame concepts and frame attributes. In previous works of LU [4] a segmentation and word labeling was done, but in this work only the frame representation is needed.

We defined 17 concepts for the DIHANA task grouped in three sets: concepts that are general for any dialog system (GENERAL), concepts that represent queries to the information system (QUERY), and concepts that have associated values and that represent constraints to the query (ATTRIBUTE). The concepts are presented in Table 2.

<table>
<thead>
<tr>
<th>GENERAL</th>
<th>QUERY</th>
<th>ATTRIBUTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affirmation</td>
<td>Hour</td>
<td>Trip-Type</td>
</tr>
<tr>
<td>Negation</td>
<td>Train-Type</td>
<td>Origin</td>
</tr>
<tr>
<td>Not-Understood</td>
<td>Duration</td>
<td>Destination</td>
</tr>
<tr>
<td></td>
<td>Price</td>
<td>City</td>
</tr>
<tr>
<td></td>
<td>Date</td>
<td>Date</td>
</tr>
<tr>
<td></td>
<td>Departure</td>
<td>Arrival</td>
</tr>
<tr>
<td></td>
<td>Class</td>
<td>Class</td>
</tr>
<tr>
<td></td>
<td>Order-Number</td>
<td>Order-Number</td>
</tr>
<tr>
<td></td>
<td>Services</td>
<td>Services</td>
</tr>
</tbody>
</table>

Table 2: Concept list for the DIHANA task.

#### 3.2. Results

A training and a test sets were defined for the experiments. The 80% of the sentences of the corpus were used for training and the other 20% were used for test.

The results of the experiments are shown in the Table 3. The results are presented in terms of Concept Accuracy (which is the equivalent to Word Accuracy using concepts as units).

We have done two experiments. In the first experiment (Words in the Table) we used the corpus without applying any pre-processing. In order to improve the coverage of the models, in the second experiment the categorization process has been done (Categories in the Table). This categorization consisted of:

- Categorization: task-dependent categories were defined; for example, cities, services, numbers, days, months, time-periods. Some of these categories can include not only single words but also segments of two or three words. For example “in the morning” as time-period.

Figure 3 shows an example of categorization.

- Lemmatization: nouns, verbs, ...

Original sentence: “I would like to know the timetables from Valencia to Madrid this Sunday”

Categorized sentence: “I would like to know the timetables from [city-name] to [city-name] this [day-of-week]”

Figure 3: Example of categorization

In order to compare with the understanding results when explicit segmentation and annotation of the training corpus exist, other experiment has been done. In this case we trained stochastic models from the segmented and labeled corpus, as in [4]. The corresponding results of the decoding process is also presented in Table 3 (Manually segmented in the Table).

For all these experiments, we present two type of results: the first one using the correctly transcribed test sentences as input for the semantic decoder (Transcription in the Table), and the second one using the recognized test sentences (Speech in the Table). We used the CMU-Sphinx2 as recognizer and we obtained a Word Accuracy of 82%.

The results show a better performance of the model when the correct transcriptions were used, as expected.

Comparing first and second experiments it can be appreciated that the use of categorization outperforms the behavior of the system when correct transcriptions are used. However, the results are not better when the recognized input is considered.
It could be due to the fact that the recognition process generates errors that are propagated to the categorization process. On the other hand, results with models that have been learnt from segmented training corpus outperform our approach that uses less information in the training. However, our approach have the advantage of reduction of the effort in the preparation of the corpus.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Concept Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words Speech</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>0.82</td>
</tr>
<tr>
<td>Categories Speech</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.88</td>
</tr>
<tr>
<td>Manually segmented</td>
<td>Speech Transcription</td>
</tr>
<tr>
<td></td>
<td>Transcription</td>
</tr>
</tbody>
</table>

Table 3: Experiment Results.

4. Conclusions and future work

We have presented in this paper an approach to the development of the understanding module of a spoken dialog system. The semantic models are automatically learnt from a training corpus where the process of labeling has been simplified because only the global annotation of the sentence is necessary, instead of a detailed word to concept labeling.

It is also an interesting characteristic of our approach the capability of automatically find appropriate segments of words that can be associated to the different concepts. Experiments shows that this approach gives good results, requiring few labeling effort.

As future work, we can explore the possibility of applying other classification and clustering methods. It would be also convenient to apply this methodology to more complex tasks, where the lexical realizations of concepts present more overlapping and ambiguity.

5. Acknowledgements

Work partially supported by the Spanish MEC and FEDER under contract TIN2008-06856-C05-02

6. References