Neural Network and Information Theory in Automatic Speech Understanding

Salma Jamoussi, Kamel Smaïli and Jean-Paul Haton

LORIA/INRIA-Lorraine
615 rue du jardin botanique, 54602 Villers-lès-Nancy, FRANCE
Tel.: ++33 (0)3 83 59 20 00 - Fax: ++33 (0)3 83 27 83 19

E-mail : {jamoussi, smaili, jph}@loria.fr

Abstract

In this paper, we present two methods for speech understanding : an artificial neural network and an information theory based method. For both methods we have to index input sentences using semantic classes (or concepts). In the first method, we perform supervised learning and we obtain very good indexing results. In the second one, we propose a new method based on mutual information statistical measure to retrieve concepts, and also to tag each sentence by its concepts. Both methods have been tested on a tourist information corpus. The information theory method yields better recall, whereas the neural network achieves a better precision. Better performance has been obtained by the neural network method (about 4%).

1 Introduction

Language and speech recognition processing become very important research areas and their applications are more and more present in our daily life. Interactive applications must then be able to process users spoken queries, so they have to recognize what has been uttered, extract its meaning and give suitable answers or execute right corresponding commands [1].

In this paper, we present two methods to clean up the speech understanding problem. The first one is based on artificial neural network. The main interests of neural networks are their generalization capacity, their capacity to tolerate errors and moreover they can handle uncertainty and noisy data. For these reasons, neural networks seem to suit very well to our problem, and could achieve good results.

The second method is based on the information theory and more precisely on the mutual information measure. Such a method allows us not only to automatically find semantic classes but also to tag data with statistical measures. Consequently, this method is considered as a data driven clustering and tagging method which need no manual indexing nor a supervised learning step.

The second section of this paper deals with the speech understanding problem, the third and the forth ones are devoted to describe respectively the neural network method and the statistical one. In the fifth section, we introduce the database used for training, development and test steps. We compare the two methods performances in the sixth section and finally we conclude our paper in the seventh and last section.

2 The Speech Understanding Problem

A speech understanding system could be considered as a machine that produces an action as the result of an input sentence. Thus, the understanding problem could be seen as a translation process, it translates a sequence of words into a special form that represents the meaning convoyed by the sentence [3]. The sentence is then labelled by a list of conceptual entities (often called concepts). The result is a useful intermediate representation which will be used in order to interpret semantically the sentence.

Speech understanding problem can be seen then as an association problem, where we have to associate inputs (e.g. speech or text) to their respective meanings represented by a list of concepts. In [6], the authors give a general architecture for the speech understanding systems (see figure 1). They divide the problem into two subproblems. The first, and most important one, amounts to give a semantic representation to an input sentence. This representation must be formulated using an intermediate language which must be simple and representative. The following sections of this paper are devoted to explain and compare two methods for resolving such a problem.

The second step consists of converting the obtained concepts to an action to be done as a final response to the user. In order to achieve such a goal, we have just to convert these concepts into a target formal command (e.g. SQL queries, command language, etc.). This step is not difficult to achieve. In fact, if we have the right concepts, we only need to go back to the input sentence and to find suitable values for the obtained concepts. For example, if in a travel reservation framework we obtain the following concepts *"Reservation, City_Departure, City_Destination, Date"* with the following sentence as an input *"I would like to make a reservation from London to Paris the first of July"*, in the conversion step we have just to affect to each concept its real value. The following SQL request could be generated in order to know the different flying times which make the clause condition true : *"SELECT * FROM table WHERE dep = 'London' AND dest = 'Paris' AND date = '01/07/2002"'*.

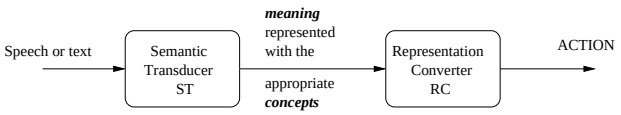

Figure 1: General architecture of a speech understanding system according to [6].

The first step to do will be the definition of the input and the output languages for the semantic transducer bloc. The input are queries formulated in natural language. The input vocabulary will be composed of any natural language word, the only restriction will be the application domain. The output language or also the "semantic language" (as defined above) must be able to give the meaning of the input sentences in an efficient and an easy way. In order to achieve that, we have to gather all the words which share the same semantic features and group them in the same "concept". A concept is then related to a given meaning, it can be substituted to any natural language sequence concerned by the same idea.

3 Neural network method description

In this first method, we choose to use a basic Multi-Layer Perceptron (MLP) with three layers. This network needs a supervised learning step for which we use a French input corpus and its equivalent in terms of concepts in the output. Our corpus is thus made up of pairs, each pair contains a natural language sentence and its corresponding meaning in terms of concepts.

3.1 Vocabulary construction

One of the questions to solve is to determine the vocabulary necessary on which the understanding process is based. The vocabulary used in this method is extracted from a tourist database.

French is highly inflected language, and the number of inflectional words is larger than in English. The use of base-forms allows us to extend the input vocabulary which will contain all the possible inflectional forms of a base-form. Therefore, the size of the vocabulary will be at least five times greater than the basic one. For example, to the base-form "speak" will be associated "speak", "speaks", "spoke" and "spoken". Using this method, we obtain 460 different base-forms.

The same principle is used to find a suitable codification for the output concepts. In this case concepts are independent from the morphological form. 46 hand determined concepts are used as the output of the neural network.

3.2 The neural network design

Our neural network is a MLP with three layers. The number of neurons in the input layer is 460 (total number of the base-forms). For the output layer we use 46 neurons (one neuron for each concept).

As explained before, each neuron in the input layer is associated to an unique word from the learning database. Each word is represented by a number which is the same as the corresponding input neuron number. Thus, if we want to achieve the learning of the sentence "When the music festival will be held" with its corresponding outputs "Date" and "Event", all the input neurons which represent these sentence words will be set at one, all the others will be null. For this example, the output layer of the network has only two neurons set to one (the neurons representing the concepts "Date" and "Event"), others will be set to null (figure 2). So our input and output vectors are binary, and these pairs represent the association existing between words and concepts.

For the hidden layer, we have to decide for the number of neurons which will constitute it. Figure 3 shows the evolution of the concept error rate [1] according to the number of the hidden layer neurons. This experiment allow us to find out the optimal number : 50.

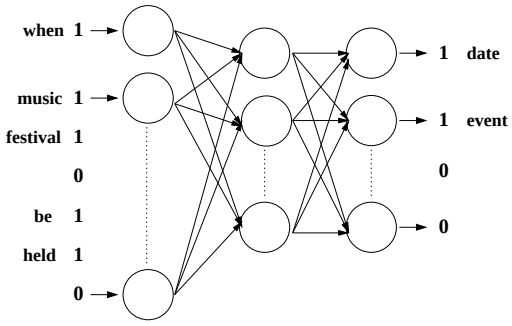

Figure 2: The neural network architecture and its functioning principle.

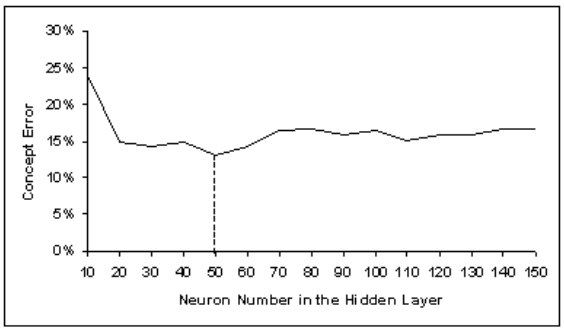

Figure 3: Concept error rate according to the neurons number in the hidden layer.

Finally, our MLP has 460 neurons in the input layer, 50 neurons for the hidden layer and 46 neurons for the output layer.

[1] Concept error = inserted concepts + omitted concepts + substituted concepts

4 Statistical method description

In this method, we use quantitative measures based on the information theory principles. These measures allow us to compute the association degree between two given words and then to make up lists of the most correlated words [2]. And then, these lists participate in the concepts construction. At the end we will use these generated concepts to label data.

4.1 Data clustering

In order to find the list of concepts, we must first of all, clean our corpus. So we need to filter it and to remove all the stop words and the words with a weak occurrence frequency. Like in the neural method, we also replace each word by its base-form and finally we compute the association between any pair of words as in [4] :

$$\begin{aligned} I(A:B) = \quad & P(A,B)\log\frac{P(A|B)}{P(A)P(B)} + \\ & P(A,\overline{B})\log\frac{P(A|\overline{B})}{P(A)P(\overline{B})} + \\ & P(\overline{A},B)\log\frac{P(\overline{A}|B)}{P(\overline{A})P(B)} + \\ & P(\overline{A},\overline{B})\log\frac{P(\overline{A}|\overline{B})}{P(\overline{A})P(\overline{B})} \end{aligned} \quad (1)$$

This formula represents the average mutual information (MI) measure between two words A and B. It allows to decide if the word A is significantly correlated with the word B or not. In fact, if the two words are often together in the same sentences, $I(A:B)$ will have a high value otherwise it will have a small value and it means that the two words are very independent and they don't represent any special meaning.

In the case of a high MI value, A and B form a "trigger pair" [4]. We apply this formula for all the word couples of the corpus to find the list of the trigger pairs. Then, we associate for each word w the list of the most correlated words. We will assume that a word w_k is very correlated with the word w if $I(w:w_k)$ is higher than a threshold $S(w)$ computed as follows :

$$S(w) = \frac{min_{w_i \in V}\{I(w:w_i)\} + max_{w_i \in V}\{I(w:w_i)\}}{2}$$

Where V is the considered vocabulary. Thus we obtain for each word its correlated word list. We can now find the final list of concepts, the idea amounts to group all completely connected words together, it means that if we have (A,B) as a trigger pair and (B,C) an other trigger pair, we can assume that (A,B,C) is a concept only if we have the trigger pair (A,C). We repeat this process for all the obtained triggers and we obtain the final concept list.

In our case, we obtain 64 different concepts which cover almost all the corpus meanings. Each concept contains 2, 3 or 4 words.

4.2 Sentence labelling

The aim of this step is to label sentences with their corresponding concepts. In our case, a concept $C_j = (c_{j1}, c_{j2}, ..., c_{jm})$, is a set of correlated words and a sentence P_i is composed of a list of words $P_i = (p_{i1}, p_{i2}, ..., p_{in})$. Our idea is to test for the sentence P_i if the concept C_j can label it or not, so we have to compute the degree of correlation between a concept and a sentence. The most natural thing to do is to compute their average mutual information quantity and this can be done by computing for each couple of word (p_{ik}, c_{jl}) its correlation degree and by calculating an average IM as :

$$IM(P_i, C_j) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m} I(p_{ik}:c_{jl})}{n \times m}$$

For each sentence, we test its correlation degrees with all the possible concepts and we keep only the most correlated ones. To decide if a concept C_j must be kept to tag a sentence P_i or not, we have to fix a reject threshold. This threshold will differ from a sentence to another because it depends on the correlation degrees found each one. That can be given by :

$$S(P_i) = k \times max_{j=1..64}\{IM(P_i, C_j)\}$$

Figure 4 gives the evolution of the concept error according to the value of k. We associate to each sentence P_i the concepts which give higher correlation degree than $S(P_i)$ and this will finish the tagging step. A comparison of this tagging method with a concept segmentation based on Viterbi algorithm is under work.

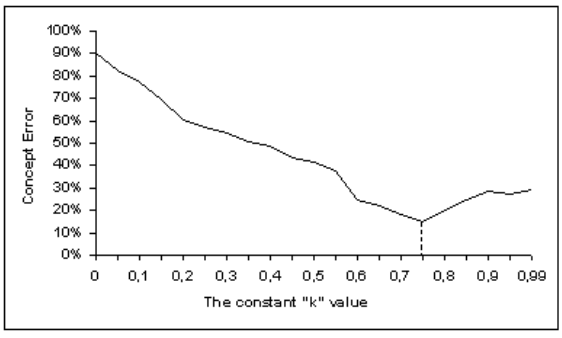

Figure 4: Concept error rate according to the constant k.

5 The Corpus

In our experiments, we transcribe a corpus containing 500 French queries. The proposed application is a tourist office database queries. Our system will be considered as

an interactive terminal where tourists ask for different information. Our corpus is then composed of many such queries. For each query, we associated a concept set which translate its meaning. One third of the corpus contains unknown words noted "UNK". They represent all words that have not been seen in the learning database.

We used 400 sentences for the training, 50 for the development and 50 for the test in each method. Obtained results are given in the next section.

6 Results and discussion

We are based our evaluation on some measures used in the information retrieval domain : "Recall" and "Precision". The recall "R" represents the rate of good answers obtained among the total good answers needed and the precision "P" represents the rate of good answers obtained among all the obtained answers. The system efficiency "E" is then calculated as follows :

$$E = \frac{2 \times R \times P}{R + P}$$

E, also called *F-measure*, represents the harmonic average of the recall R and the precision P [5]. This measure allows us to combine the two measures R and P in only one measure, it also represent a reliable measure because it decreases when only one measure (R or P) decreases and it increases when the both measures (R and P) increase.

In the table 1, we give the obtained results with the both methods on the development corpus and test corpora. These results show that the neural method has a very good precision capacity and that the statistic method has a good recall capacity. The global efficiencies of the two methods are encouraging and they aren't very remote. Nevertheless, the gap between them can be explained with two main arguments. First, the two methods use different concept kind, the MLP uses concepts elaborated manually and the statistic method discovers itself its needed concepts. Second, the neural network achieves a supervised learning step, whereas the second method uses unsupervised mechanisms.

Our statistical method seems to be interesting thanks to its capacity to find the concepts and to label the sentences automatically without human expertise. It can be very efficient method in the several cases where we have no indexed corpus and no established concept lists.

Finally we can notice that the two methods are complementary, and that they can be combined to give one hybrid system with very interesting features. In fact, we have in one hand the neural method which gives very good precision results and on the other hand the statistic method which shows a very good recall capacity. In addition both methods have very different features but that can be combined to improve results.

	Neural method		Statistic method	
	Development	Test	Development	Test
R	85%	65%	**95%**	**83%**
P	**98%**	**88%**	81%	64%
E	**91%**	76%	**87%**	72%

Table 1: Obtained results with both methods on the development and test corpora.

7 Conclusion

Speech understanding can be seen as the process of translating input natural language sentences into output sentences in an appropriate semantic language. Under this point of view, two approaches have been presented in this paper. The first method based on a neural network gave good results and showed a very large precision capacity. The second method is a new one based on the mutual information measure and the concept tagging approach is original. In addition to its very interesting features, it gave also encouraging results and especially a very good recall capacity. Integrating this method into our speech dictation machine MAUD [7] is under work.

8 References

[1] C. Bousquet-Vernhettes and N. Vigouroux and G. Pérennou, "Stochastic Conceptual Model for Spoken Language Understanding," in *Proc. SPECOM'99*, Moscou, 1999.

[2] N. Coccaro and D. Jurafsky. "Towards better integration of semantic predictors in statistical language modeling," In *Proc. ICSLP'98*, Australia, 1998.

[3] R. Pieraccini, E. Levin, and E. Vidal, "Learning how to understand language," in *Proc. EuroSpeech'93*, pp. 1407-1414, Germany, 1993.

[4] R. Rosenfeld, *Adaptive statistical language modeling: maximum entropy approach*, Ph.D. thesis, Pittsburgh, 1994.

[5] C. Van Rijsbergen, "Information retrieval, 2nd edition", Butterworths, 1979.

[6] E. Vidal, R. Pieraccini, and E. Levin, "Learning associations between grammars:a new approach to natural language understanding," in *Proc. EuroSpeech*, Germany, 1993.

[7] I. Zitouni, J.F. Mari, K. Smaïli, and J.P. Haton, "Variable-length sequence language model for large vocabulary continuous dictation machine : the n-seqgram approach," in *Proc. EuroSpeech'99*, Hungary, 1999.