Calvo Lance, M.; Hurtado Oliver, LF.; García Granada, F.; Sanchís Arnal, E. (2012). A multilingual SLU system based on semantic decoding of graphs of words. En Advances in Speech and Language Technologies for Iberian Languages. Springer Verlag (Germany). 328:158-167. doi:10.1007/978-3-642-35292-8_17.

# A multilingual SLU system based on semantic decoding of graphs of words

Marcos Calvo, Lluís-F. Hurtado, Fernando García, and Emilio Sanchis

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{mcalvo,lhurtado,fgarcia,esanchis}@dsic.upv.es

**Abstract.** In this paper, we present a statistical approach to Language Understanding that allows to avoid the effort of obtaining new semantic models when changing the language. This way, it is not necessary to acquire and label new training corpora in the new language. Our approach consists of learning all the semantic models in a target language and to do the semantic decoding of the sentences pronounced in the source language after a translation process. In order to deal with the errors and the lack of coverage of the translations, a mechanism to generalize the result of several translators is proposed. The graph of words generated in this phase is the input to the semantic decoding algorithm specifically designed to combine statistical models and graphs of words. Some experiments that show the good behavior of the proposed approach are also presented.

**Keywords:** Multilingual Language Understanding, Graph of Words

## 1 Introduction

In the last few years, different approaches have been developed for the problem of Spoken Language Understanding (SLU). There are many types of applications for SLU, and one of the most interesting is its use in limited-domain spoken dialog systems. Some characteristics of this kind of systems are that they have to deal with spontaneous speech, the size of the vocabulary is small or medium, and the semantic labels involved in the understanding process are strongly related to some specific words or segments of words present in the user turns. In the recent literature, a variety of approaches for automatic SLU have been proposed, like those explained in [1–3].

As in other speech areas, statistical modelization is one of the successfully approaches that have been used in SLU [4–7]. One of the advantages of these approaches is that the models can be automatically learned from labeled training samples and they can represent the variability of sequences of words and concepts. Due to the limited number of training samples, and the limitations of the learning methods, not all the variability of the speech messages can be correctly represented in the models, and some errors generated in previous phases can not be recovered in the following phases. This is the reason why the use of $n$-best or

weighted graphs of linguistic units are interesting approaches to communicate information between the different modules [8, 9].

The use of some kind of graph of words as the input of the decoding module makes this task more difficult, as the search space becomes larger, and it is necessary to combine the different weights representing the accuracy of the words in the graph and the corresponding probabilities of the model of the decoding process (in our case the semantic models). On the other hand, the advantage of using graphs is that there is more information that could help to find the correct semantic interpretation, rather than just taking the best sentence given by the Automatic Speech Recognizer (ASR), or other module that provides the input sentence.

The methodology proposed in this paper is based on Stochastic Finite State Transducers (SFST). This is a generative approach that composes several transducers containing acoustic, lexical and semantic knowledge. Our method performs this composition obtaining as a result a graph of concepts, where semantic information is associated to segments of words. To carry out this step, we use a statistical modelization of the lexicalisation of concepts; that is, the sequences of words associated to the concepts, and also a statistical model of the sequences of concepts. All these probabilities are automatically learned from a training corpus segmented and labeled in terms of concepts.

One of the problems of the statistical modelization of semantics is that the training process needs a segmented and labeled corpus. In most cases it is necessary a very time-consuming work to label the training corpus. This is the reason why many works oriented to avoid this work have been developed, such as semi-supervised learning techniques, or active learning methods [10, 11]. These techniques are also used to facilitate the adaptation of the system to different tasks or new languages. In particular, when the problem is to translate a previously obtained SLU system into another language, some approaches can be used: to translate the corpus and to do a new labeling; to automatically obtain the translated system and labeling; or to process the sentences in the new language (after translating them) with the original models. The latter approach is the one that we have developed in this paper. That is, we obtained the semantic models for Spanish by using a labeled training corpus, and we used this system to interact with users of other language (French in this work) by translating their sentences into the target language and decoding these translated sentences. However, it must be considered that the quality of the general purpose translators is quite insufficient. This is the reason why it is necessary to supply the maximum information of the original sentences to the semantic decoding process, in order to better tackle the errors generated in the translation process.

In the proposed system, the sentence uttered in the source language is translated into a graph of words in the target language, by means of an adequate combination of the translations generated by several web translators. This way, we obtain a generalization of the translations that allows the semantic decoder to recover some of the errors generated in the translation phase.

This paper is organized as follows. In Section 2, the general framework of the system is presented. In Section 3, the process of generating the graph of words from the different sentences obtained by the translators is described. In Section 4, the algorithm of semantic decoding of the graph of words is presented. Section 5 shows some experimental results over the DIHANA task, and finally, in Section 6 some conclusions and future work are presented.

## 2 The SFST approach for multilingual SLU

The SLU problem can be expressed as stated in Equation 1, where $C$ represents a sequence of concepts or semantic labels and $A$ is the utterance that constitutes the input to the system.

$$\hat{C} = \underset{C}{\operatorname{argmax}}\, p(C|A) \tag{1}$$

The task we are addressing is multilingual SLU, which in our case means that the speaker utters a sentence in one language $s$, but our models are trained in another language $t$. Thus, a translation process between the source and target languages is needed. Taking into account the underlying sequence of words $W_s$ uttered by the speaker in the source language and its translation into the target language $W_t$, this equation can be rewritten as follows.

$$\hat{C} = \underset{C}{\operatorname{argmax}} \max_{W_s, W_t} p(C, W_s, W_t|A) \tag{2}$$

Applying the Bayes' Rule and making some reasonable assumptions about the independence of these variables, the probability of this equation can be decomposed into several factors as shown in Equation 3.

$$\hat{C} = \underset{C}{\operatorname{argmax}} \max_{W_s, W_t} p(A|W_s) \cdot p(W_s|W_t) \cdot p(W_t|C) \cdot p(C) \tag{3}$$

This equation can be seen as the composition of 4 SFST, which are:

- A SFST $\lambda_G$ generated by the ASR module where the acoustic probabilities $p(A|W_s)$ are represented.
- A SFST $\lambda_{W_s 2 W_t}$ that expresses the translation process between the source and target languages.
- A SFST $\lambda_{W_t 2 C}$ that represents the probability that a sequence of words in the language $t$ corresponds to a concept $C$. Thus, it provides the probability distribution $p(W_t|C)$.
- A SFST $\lambda_{SLM}$ which corresponds to a language model of the sequences of concepts. Thus, it modelizes the probability of a sequence of concepts $p(C)$.

It is possible to compose these four transducers as shown in Equation 4.

$$\lambda_{MSLU} = \lambda_G \circ \lambda_{W_s 2 W_t} \circ \lambda_{W_t 2 C} \circ \lambda_{SLM} \tag{4}$$

In consequence, finding the best path in the resulting transducer $\lambda_{MSLU}$ provides as a result the best sequence of concepts $\hat{C}$, a translation $\tilde{W}_t$ of the transcription of the input utterance as well as a segmentation of $\tilde{W}_t$ according to $\hat{C}$.

In this work, our goal is to build and evaluate a multilingual understanding system that receives a sentence in one language and, passing this sentence through a translation process, is able to use understanding models trained in another language. If the input to the system were utterances, the recognition process would add some error to the output of the understanding module. Thus, in order to evaluate the performance of the understanding system without any other external factors, the input to our system will be correct written sentences, which is equivalent to assume that we have a "perfect" ASR. In terms of probabilities, this implies that $p(A|W_s) = 1$. For this reason, we will not use the $\lambda_G$ transducer from Equation 4.

Moreover, $p(W_s|W_t)$ can be rewritten as $\frac{p(W_t|W_s)\cdot p(W_s)}{p(W_t)}$. Taking the written sentence as the input to the system, means that the whole sentence that is going to be translated is known[1]. From this known sentence in the source language, we will obtain a set of possible translations and represent them as a graph of words. If we consider that the probability $p(W_t)$ of any sentence of the set of possible translations is the same, then it is not necessary to take into account this probability in the maximization process. Considering these two simplifications, we can rewrite Equation 3 as:

$$\hat{C} = \underset{C}{\operatorname{argmax}} \max_{W_t} p(W_t|W_s) \cdot p(W_t|C) \cdot p(C) \tag{5}$$

Thus, the $\lambda_{W_s 2 W_t}$ transducer will represent the probability $p(W_t|W_s)$ that a sentence $W_t$ in the target language is a translation of $W_s$.

## 3 Graph of words generation

In this section, the process of obtaining the word-graph in the target language from a sentence in the source language is be explained. This process is divided into three steps:

1. the source sentence (in French in this work) is translated to the target language using several free-available web translators. As a result, a set of sentences in the target language (Spanish in this work) that represent different possible translations of the source sentence is obtained.
2. this set of sentences are aligned using a multiple sequence alignment algorithm.
3. the aligned sentences are used to obtain the word-graph that will be the input to the graph-based understanding module.

[1] It would be the same if we took the 1-best from an ASR.

A Multiple Sequence Alignment (MSA) is a sentence alignment process that allows the alignment of three or more sentences that minimize the number of substitutions, insertions and deletions among all the sentences. Although the original use of MSA is the alignment of biological sequences, MSA algorithms can align sequences of symbols of any kind. Within the frame work of Natural Language Processing (NLP), MSA has been mainly used in automatic translation tasks [12, 13]. All these approaches –including the one presented in this paper– coincide in the creation of a graph of words from the result of the MSA. However, they differ in how the graph is generated and what it is used for.

In this work, a modification of the well-known ClustalW [14] Multiple Sequence Alignment software has been used. These modifications consist basically of: i) it allows the alignment of sentences with any symbol, originally ClustalW only allows symbols representing protein, DNA, and RNA; ii) all weight matrices have been replaced by 0s and 1s (where 1 is the score for symbol matches and 0 is the score for symbol mismatches). That is, the same probability is assigned to all symbol substitutions.

### 3.1 From alignment matrix to graph of words

The result of the Multiple Sequence Alignment process is a MSA alignment matrix. Each row in the matrix represents a different aligned sentence and the columns are synchronization points. In other words, a column of the matrix indicates which symbols of the sentences are aligned at each point. Unless all the sentences are the same, there will be several non-alignment points. These non-alignment points are represented in the alignment matrix by the special symbol '-'.

From the MSA alignment matrix, a directed acyclic weighted graph of words is created. In order to build this graph of words, the following algorithm is used:

1. as many nodes as the number of columns in the alignment matrix plus one additional node to be used as the initial node are created.
2. for each matrix cell containing a symbol other than '-' –that is, a cell that represents a real word of an aligned sentence– an arc in the graph will be created. The destination node of the arc will be the one representing the column to which the cell belongs and the origin node will be the one representing the column of the previous word in the same sentence (or the initial node if the cell contains the first word of the sentence). The arc is labeled with the word in the cell and its weight is set to 1. If the arc already exists (because it has been previously added), its weight is incremented by 1.
3. the weights of the arcs are normalized to represent probabilities.

Figure 1 shows a real example –extracted from the test set– of the full process of obtaining the graph of words in the target language (Spanish) from a sentence in the source language (French). Firstly, the original sentence *pouvez vous répéter à quelle heure sort le premier* is translated using 6 different free-available web translators. Secondly, the 6 translations are aligned using a Multiple Sequence

Alignment algorithm. Finally, the directed acyclic weighted graph of words is created from the MSA alignment matrix.

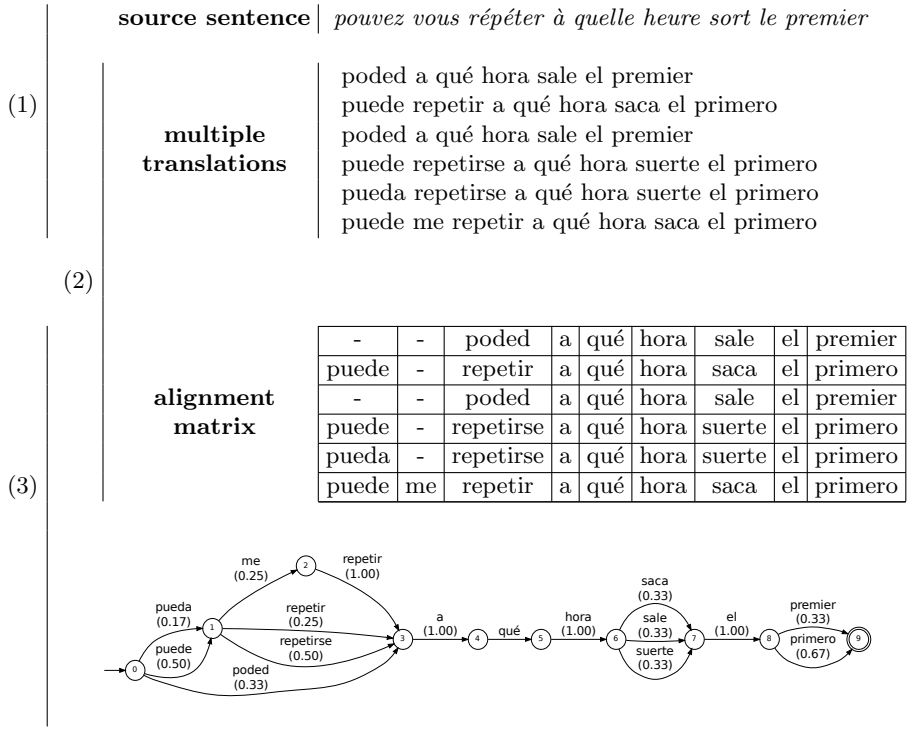


**Fig. 1.** Steps in the process of obtaining the graph of words from the original sentence *pouvez vous répéter à quelle heure sort le premier*.

The obtained graph of words represents a language which is a generalization of the individual translations of the original sentence. A full path –from the initial node to the final node– over the graph may be seen as an alternative translation of the original sentence. In addition, the product of the weights of all the arcs in a full path may be considered as the probability of the string represented by the path $W_t$ (in the target language) to be the translation of the original string $W_s$ (in the source language); that is, $p(W_t|W_s)$.

## 4 Understanding the translated graphs of words

As every arc in the graph of words is labeled with a word, any path between any pair of nodes represents a segment of words that can be related to one or more concepts. Consequently, it is possible to build a new graph with the same set of nodes but where each arc is labeled with a segment of words and a concept

associated to it. Every arc of this new graph can be weighted with a combination of the original graph probability and the probability that the segment belongs to the concept.

To build this graph of concepts, for each pair of nodes $i, j$ and each concept $c$, an arc that represents the most probable path associated to concept $c$ between these nodes is created. We define the most probable path as the one that maximizes the expression $p(W_t^{i,j}|W_s) \cdot p(W_t^{i,j}|c)$ given a concept $c$, where $W_t^{i,j}$ denotes a segment of words induced by a path from node $i$ to node $j$. The resulting arc will be labeled with the concept $c$, and the sequence $W_t^{i,j}$ that maximizes the former probability. The arc will be weighted with the value of this expression for $W_t^{i,j}$ and $c$.

The last formula introduces the probability of a sequence of words given a concept. To estimate it, a model of the lexical structures associated to the concepts is needed. One way to estimate this is to train a language model for each concept, using the segments of words associated to each of them. Thus, the probability $p(W_t^{i,j}|W_s)$ is represented in the graph of words obtained from the translation and generalization process, and $p(W_t^{i,j}|c)$ is provided by the language model specific to each concept.

Finally, finding the best path in the graph of concepts between the initial and the final nodes, provides the best sequence of concepts and the sentence associated to it, as well as a segmentation of this sentence. To find this best path, a language model of the sequences of concepts may be used, in order to properly model their concatenation.

This way of obtaining the best sequence of concepts fulfills what was stated in Equation 4, as $\lambda_{Wd2C}$ is composed by the set of the language models that provide the probability that a sequence of words belongs to a concept and $\lambda_{SLM}$ is the language model of the sequence of concepts that helps to find the best path in the graph of concepts.

## 5 Experiments and results

For the experimentation, we used the DIHANA corpus [15]. This is a corpus in Spanish composed by 900 dialogs acquired by 225 speakers using the Wizard of Oz technique, with a total of 6,229 user turns. The DIHANA task consists of acceding by phone to a spoken dialog system to ask for information about railway timetables and fares using spontaneous speech in Spanish. The experiments reported here were performed using the 5,369 user turns of the DIHANA corpus, splitting them into a set of 480 turns for test and the remaining 4,889 turns for training. Some interesting statistics about the DIHANA corpus are shown in Table 1.

In order to perform multilingual experiments using the DIHANA task, the 480 test sentences were correctly written in French. Then, we used 6 general-purpose free-available web translators to translate them into Spanish. The 6 translations of each sentence were combined using the algorithm explained in

**Table 1.** Characteristics of the DIHANA corpus.

| | |
|---|---|
| Number of words | 47,222 |
| Vocabulary size | 811 |
| Average number of words per user turn | 7.6 |
| Number of concepts | 30 |
| Average number of words per segment | 2.5 |
| Average number of segments per turn | 3.0 |
| Average number of samples per concept | 599.6 |

Section 3, obtaining as a result the graph of words which is the input for the decoding algorithm.

In order to learn all the semantic models, the Spanish training sentences were used. The transcriptions of the user sentences of the DIHANA training corpus were segmented and labeled in terms of concepts. This segmentation were used to learn a language model for each concept. In addition, a semantic model was also leaned using the sequences of concepts. All the language models involved in this experimentation were bigram models trained using Witten-Bell smoothing and linear interpolation.

For the evaluation, we have used three measures: the Translation Word Error Rate (TWER), the BLEU measure, and the Concept Error Rate (CER). The TWER represents the WER of the best path in the graph of words; that is, the path that has been generated by the semantic decoding process. The BLEU is a standard measure used to evaluate automatic translation systems. The CER is the rate of incorrectly understood concepts, considering that the reference sequence of concepts is the same in both languages (which means that, in some cases, some correct sequences in French can be counted as errors). The TWER and BLEU measures represent not only the quality of the composition of transducers but also the behavior of the search algorithm guided by the semantics.

Table 2 shows the results obtained in the experiments, both for the combination of translators and for each one of them, individually. It also shows the results considering the correct sentences in Spanish as input. This result gives an idea of the best CER that could be achieved with our semantic modelization if no error were introduced in the translation and generalization processes.

These results show that the combination of the translators obtains better results that considering them individually. That is, the increasing of the coverage given by the use of several translators, and the adequate combination in the graph of words outperforms the behavior of each isolated translator. These better performances are observed in terms of TWER, BLEU, and CER. In addition, the CER obtained for the combination of the translators is less than 2.7 points higher than the one achieved using the reference Spanish sentences. This means that, although the translation process introduces some syntactic errors (which can be seen in the TWER score), most of the semantic meaning is kept.

**Table 2.** TWER, BLEU, and CER obtained from the combination of translators and each of them individually, as well as for the reference sentences in Spanish.

| Input graphs of words | TWER | BLEU | CER |
|---|---|---|---|
| Reference sentences | – | – | 9.09 |
| Translator 1 | 30.74 | 50.37 | 15.77 |
| Translator 2 | 27.49 | 52.00 | 16.67 |
| Translator 3 | 30.50 | 50.71 | 15.22 |
| Translator 4 | 24.04 | 61.35 | 13.09 |
| Translator 5 | 23.85 | 59.79 | 14.60 |
| Translator 6 | 27.38 | 50.82 | 19.35 |
| Combination of all the translators | **18.68** | **67.40** | **11.78** |

## 6 Conclusions and future work

We have presented an approach to multilingual language understanding. One of the advantages of this approach is that it is not necessary to estimate different models depending on the language. The modelization of the semantics of the task is done by statistical models. The way to represent the variability of the translation process is done by the construction of graphs of words. We have developed a search algorithm to generate graphs of concepts form the graphs of words and the semantic models. Experiments show that the proposed approach achieves good results. It would be interesting, as future work, to adapt the system to other languages –like English– that have syntactic structures different from Latin languages.

**Acknowledgments.** This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, by the Vic. d'Investigació of the UPV under contract 20110897, and by the Spanish MICINN under FPU Grant AP2010-4193.

## References

1. Hahn, S., Dinarelli, M., Raymond, C., Lefèvre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., Riccardi, G.: Comparing stochastic approaches to spoken language understanding in multiple languages. Audio, Speech, and Language Processing, IEEE Transactions on **6**(99) (2010) 1569–1583
2. Raymond, C., Riccardi, G.: Generative and discriminative algorithms for spoken language understanding. Proceedings of Interspeech 2007 (2007) 1605–1608
3. Tur, G., Mori, R.D.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. 1 edn. Wiley (2011)
4. Maynard, H.B., Lefèvre, F.: Investigating Stochastic Speech Understanding. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). (2001)

5. Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L.: Extracting Semantic Information Through Automatic Learning Techniques. IJPRAI **16**(3) (2002) 301–307
6. He, Y., Young, S.: Spoken language understanding using the hidden vector state model. Speech Communication **48** (2006) 262–275
7. De Mori, R., Bechet, F., Hakkani-Tur, D., McTear, M., Riccardi, G., Tur, G.: Spoken language understanding: A survey. IEEE Signal Processing magazine **25**(3) (2008) 50–58
8. Hakkani-Tür, D., Béchet, F., Riccardi, G., Tur, G.: Beyond ASR 1-best: Using word confusion networks in spoken language understanding. Computer Speech & Language **20**(4) (2006) 495–514
9. Tur, G., Wright, J., Gorin, A., Riccardi, G., Hakkani-Tür, D.: Improving spoken language understanding using word confusion networks. In: Proceedings of the ICSLP, Citeseer (2002)
10. Tur, G., Hakkani-Tr, D., Schapire, R.E.: Combining active and semi-supervised learning for spoken language understanding. In: Speech Communication. Volume 45. (2005) 171–186
11. Ortega, L., Galiano, I., Hurtado, L.F., Sanchis, E., Segarra, E.: A statistical segment-based approach for spoken language understanding. In: Proc. of InterSpeech 2010, Makuhari, Chiba, Japan (2010) 1836–1839
12. Sim, K.C., Byrne, W.J., Gales, M.J.F., Sahbi, H., Woodland, P.C.: Consensus network decoding for statistical machine translation system combination. In: IEEE Int. Conference on Acoustics, Speech, and Signal Processing. (2007)
13. Bangalore, S., Bordel, G., Riccardi, G.: Computing Consensus Translation from Multiple Machine Translation Systems. In: In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001. (2001) 351–354
14. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: ClustalW and ClustalX version 2.0. Bioinformatics **23**(21) (November 2007) 2947–2948
15. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López de Letona, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proceedings of LREC 2006, Genoa (Italy) (May 2006) 1636–1639