Multilingual Spoken Language Understanding using graphs and multiple translations

Marcos Calvo, Lluís-Felip Hurtado, Fernando Garcia, Emilio Sanchis, Encarna Segarra

Departament de Sistemes Informatics i Computacio,
Universitat Politecnica de Valencia
Cami de Vera s/n, 46020, Valencia, Spain

Abstract

In this paper, we present an approach to multilingual Spoken Language Understanding based on a process of generalization of multiple translations, followed by a specific methodology to perform a semantic parsing of these combined translations. A statistical semantic model, which is learned from a segmented and labeled corpus, is used to represent the semantics of the task in a language. Our goal is to allow the users to interact with the system using other languages different from the one used to train the semantic models, avoiding the cost of segmenting and labeling a training corpus for each language. In order to reduce the effect of translation errors and to increase the coverage, we propose an algorithm to generate graphs of words from different translations. We also propose an algorithm to parse graphs of words with the statistical semantic model. The experimental results confirm the good behavior of this approach using French and English as input languages in a spoken language understanding task that was developed for Spanish.

Keywords:
Multilingual Language Understanding, Graph of Words, Graph of Concepts, Statistical Semantic Models

Email addresses: mcalvo@dsic.upv.es (Marcos Calvo), lhurtado@dsic.upv.es (Lluís-Felip Hurtado), fgarcia@dsic.upv.es (Fernando Garcia), esanchis@dsic.upv.es (Emilio Sanchis), esegarra@dsic.upv.es (Encarna Segarra)

1. Introduction

Spoken Language Understanding (SLU) is an important challenge in human-machine interaction systems (Hahn et al., 2010; Raymond and Riccardi, 2007; Tür and Mori, 2011). In particular, SLU is a key component of Spoken Dialog Systems. Although many of the SLU systems are rule-based, there has been a growing interest in statistical modelization, which has provided good results (Maynard and Lefèvre, 2001; Segarra et al., 2002; He and Young, 2006; Lefèvre, 2007; De Mori et al., 2008). Statistical models have the advantages that they can be automatically learned from training samples and can accurately modelize the variability of semantics in spontaneous speech. Unfortunately, however, it is necessary to have a segmented and labeled training corpus that in most cases must be manually generated. This is a very time-consuming task which makes the adaptation of SLU systems to different tasks or languages difficult and expensive. In order to address this problem, many efforts have been made to develop semi-supervised and unsupervised learning techniques for semantic modelization (Tür et al., 2005; Ortega et al., 2010). These techniques can help to learn models from unlabeled corpora, in some cases taking advantage of the large amount of data that can be extracted from the web and other linguistic resources (Tür et al., 2011; Heck and Hakkani-Tür, 2012).

When the problem is to adapt a SLU system that was developed for one language to another language, it would be desirable to take advantage of the effort made for the original language and not have to replicate the work for the other language. This issue has also been addressed in other areas of Natural Language Processing such as Part-Of-Speech (POS) tagging (Täckströ et al., 2013). In this work, a projection of POS annotation from English to other resource-poor languages is done.

The multilingual approaches to SLU can be grouped in two classes, so-called test-on-source and train-on-target. In the test-on-source approach, there is a SLU system developed for a source language and the test are utterances in another language. The process consists of translating the test sentence into a sentence in the source language and performing the SLU of this translated sentence by using the SLU system in the source language. In the train-on-target approach, a new SLU model is trained in the target language, which is the language in which the test utterances are pronounced. To do this, it is necessary to translate the training corpus from the original language to this new language and to learn the corresponding SLU models.

It must be noted that the translation of the training corpus not only consists of the translation of the sentences but also in the segmentation and semantic labeling of the training sentences into this new language. Once we have a model in this target language, the understanding process can be solved as in the monolingual SLU because the input utterance and the models are in the same language. Some works that focus on the adaptation of SLU systems to other languages have been presented in the last few years (Servan et al., 2010; Lefèvre et al., 2010; Jabaian et al., 2013; Calvo et al., 2012; García et al., 2012). The work presented in this paper addresses the problem of developing a multilingual SLU system that avoids the effort of manually relabeling the corpus in other languages. In a previous work, we studied the possibility of translating the training corpus to learn models in the target language (García et al., 2012), which is a train-on-target strategy.

In this work, we propose an approach to multilingual Spoken Language Understanding where there is a SLU system in a source language and the user turns are translated from the target language into this source language, which is a test-on-source strategy. An essential aspect to ensure the viability of systems of this kind is the performance of the translation process. If we use Statistical Machine Translation (SMT) systems, such as MOSES (Koehn and et al., 2007), it is necessary to have a parallel corpus in both languages that must be specifically designed for the domain, and this corpus is not always easy to obtain. On the other hand, we could use general-purpose translators that can be found on the web. The problem is that these translators often generate many errors; however, by using different translators and combining these translations, we may be able to correct the errors as well as improve the coverage. This is why we have focused our work on obtaining good mechanisms to combine different translations and on determining how to process them in the semantic module of a multilingual Spoken Dialog System. The SLU system presented here is based on a decoupled architecture in which there is a first phase consisting of the translation process and a second phase that corresponds to the semantic decoding process. In order to be able to recover errors generated in the translation phase, multiple hypotheses are conveyed to the second phase by means of a graph.

We have developed an algorithm to obtain a graph of translations from the sentences generated by the translation process in the first phase. This graph represents a finite language that is a generalization of the translated sentences. In other words, the language represented by the graph not only contains all the sentences generated by the translation process, but it also

contains other additional sentences. For the second phase, the semantic decoding, we have developed a graph-parsing algorithm that supplies the best path in the graph of translations according to the stochastic semantic model.

In summary, our work proposes a test-on-source strategy to adapt SLU systems developed for one language to another language with the following main contributions:

- The use of two alternative translation processes to obtain different translations of the user utterance: we propose the use of general-purpose web translators and also the use of a SMT system (MOSES) estimated from a parallel corpus.
- The construction of a graph of words from these different translations using a proposed graph-construction algorithm.
- The application of a graph-parsing algorithm that supplies the best path in the graph of words according to the understanding model.

We have applied this approach to the SLU module of a Spoken Dialog System for the DIHANA task (Benedí et al., 2006), which consists of an information system about train timetables and fares in Spanish. To evaluate the multilingual approach, we have acquired a French and English corpus for testing, which consists of written and spoken sentences.

2. Related work

2.1. Using graphs of words for monolingual SLU

The most extended approach to Spoken Language Understanding consists of an in-cascade architecture in which the SLU module is fed with the output of the Automatic Speech Recognizer (ASR). Most of the SLU methods are not designed to handle uncertainty in the input; they use just the 1-best output provided by the ASR. However, there has been growing interest in better exploiting the information that can be extracted from the ASR as well as its n-best hypotheses, either in the form of a Word Lattice (WL) or in the form of a Word Confusion Network (WCN).

Both WLs and WCNs are graph structures (with words attached to their arcs) that represent the uncertainty of the recognition process. The difference between them lies in their structure: while WLs can have different

topologies, in a WCN all the arcs must go from one node to the following one. Nevertheless, it is possible to convert a WL into a WCN, following the algorithm explained in Hakkani-Tür et al. (2006).

It is also possible to build a Word Lattice from a set of n-best sentences. For example, an automaton (which in this case is equivalent to a WL) that represents the finite language that includes the same n sentences, is presented in Daciuk et al. (2000). However, it might be interesting to generalize this language according to the structures that appear in the n-best sentences. Following this idea, a Grammatical Inference algorithm can be used to generate an automaton that represents an extra-language induced from the original set of n sentences.

An approach to SLU that is based on finite state transducers is presented in Raymond et al. (2006). A combination of a word-to-concept transducer and a lattice of words (generated by the ASR) is proposed. The output of the system is a weighted n-best list of hypotheses. The weights of the hypotheses are obtained from the acoustic, linguistic, and semantic confidence measures. The experimentation shows that this approach, which involves dealing with multiple input hypotheses as well as with variability in the lexical realization of concepts, provides good performance.

In Hakkani-Tür et al. (2006), a postprocess of the lattices generated by the ASR is performed in order to obtain a better representation of the variability of the different hypotheses supplied by the ASR. This is done by constructing WCNs from a combination of sub-lattices. This way, an accurate generalization of the input hypotheses to the understanding process as well as a reduction in the size of the input networks provide a better performance of the SLU system.

Recently, some works that adapt statistical discriminative approaches to deal with weighted networks as input have been presented. In Henderson et al. (2012), an application of Support Vector Machine (SVM) classifiers to directly process WCNs in a single pass is proposed. The authors show that this approach outperforms the results obtained with one input hypothesis and with n-best hypotheses.

The good results obtained by using Conditional Random Fields (CRFs) for SLU have prompted some authors to extend the CRF techniques to accept WCNs as input (Tur et al., 2013). They use the WCNs for both CRF training and decoding. This approach is based on the idea that the WCNs have a special topology according to which every arc must go from one node to the following one. Hence, each set of arcs between two nodes can be seen

as a "bin", the size of which can be limited beforehand to avoid WCNs that are too large. Then, the linear prediction functions of the CRF model can be expressed in terms of all the possible n-grams induced by neighboring bins. Moreover, the use of WCNs for training allowed the authors to modelize the uncertainty incorporated by the ASR process and led them to an improvement in the experimental results.

Another interesting work (Deoras et al., 2013) is around joint decoding of words and semantic tags on word lattices. They demonstrated significant improvements in both recognition and semantic tagging accuracy over cascade approach.

2.2. Multilingual SLU

In Jabaian et al. (2013), different approaches based on both test-on-source and train-on-target strategies are proposed and applied to the MEDIA corpus. MEDIA (Devillers et al., 2004) is a corpus for hotel reservation and tourist information in French. In this work the source language is French and the target language is Italian. They study the possibility of infering stochastic translators from a small subset of aligned sentences, which also permits translating the semantic labeling. For the SLU process, they use CRFs and Statistical Machine Translation models. They provide similar results by considering test-on-source and train-on-target.

In García et al. (2012), a train-on-target approach was applied to the MEDIA corpus, with Spanish as the target language. In this work, different general-purpose web translators are used for the translation of the training sentences; that is, no specific translation models are learned for the task. Using two kinds of semantic modelization (CRFs, and a two-level statistical model), the authors show that the reduction in system performance from monolingual to multilingual is lower than 15%. In Misu et al. (2012) a train on target approach is presented in which a set of training samples for training the SLU models in the target language are collected and selected in a semi-supervised way.

In He et al. (2013), an adaptive training to address the problem of mismatching between training/testing conditions is proposed. This mismatch is due to the fact that the training corpus into the original language is usually composed of clean data, while the input sentences (after recognition and translation) are noisy data. The authors propose an approach that does not need any training data in the other language. It consists of first translating the training corpus into the second language and then translating the

sentences back to the original language. This way, the new training corpus is enriched by the distortions generated by the translation process. In their experiments with the ATIS corpus in English and Chinese, the authors show that this method improves the results of systems trained with only clean data.

Recently, Stepanov et al. (2013) presented a study of the effect of using corpora of domains that are different from those of the SLU system domain (e.g news versus conversations) for training the translation models. This happens because it is difficult to have in-domain corpora when we have to learn models for new tasks, and it is usual to search for corpora on the web or in other repositories. The authors propose some adaptation techniques to obtain more accurate translations taking into account the in-domain characteristics. They work with the LUNA corpus (Italian-Spanish and Italian-Turkish) (Dinarelli et al., 2009), and they use general-purpose web translators (such as Google translator) and stochastic translators that are specifically learned for the task, using MOSES.

3. The proposed architecture for multilingual SLU

To address the multilingual SLU problem, we propose a sequential architecture in which the communication between its different modules is performed by means of graphs that represent multiple hypotheses. This way, mistakes made in an earlier module of the system (where the knowledge represented in the other modules is not taken into account) can be recovered afterwards when more information is available. Our approach for minimizing these errors consists of providing more than a single hypothesis among the different parts of the process and weighting them in a convenient form. A compact and homogeneous way of transmitting this information is via graphs of linguistic units, which will vary based on the information available at each moment.

The SLU problem can be approached by finding the sequence of concepts $\hat{C}$ that corresponds to the meaning of a given utterance A. Considering a stochastic modelization of the semantics, this can be expressed as:

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

Let u be the user's language, and let s be the language in which the original SLU system was trained. Let W_s be the sequence of words in the

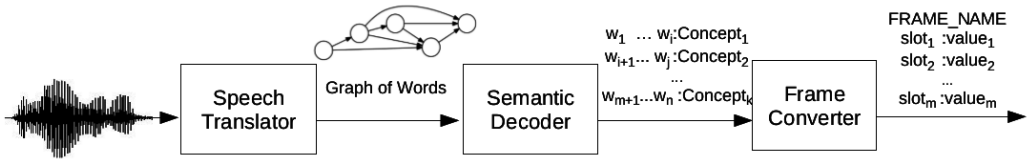

Figure 1: General architecture of the multilingual SLU system.

language s corresponding to the utterance A. By considering this variable, Equation 1 can be rewritten as:

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

Applying the Bayes' Rule, the probability of this equation can be rewritten as:

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

Making a reasonable assumption about the independence of the variables A and C, and taking into account that the maximization is independent from $p(A)$, Equation 3 can be rewritten as:

$$\hat{C} = \underset{C}{\operatorname{argmax}} \max_{W_s} p(A|W_s) \cdot p(W_s, C) \tag{4}$$

To obtain the best sequence of concepts $\hat{C}$, we adopt the test-on-source approach, following a decoupled architecture which sequentially applies the different knowledge sources (Figure 1). First, the input provided by the user in the language u is recognized and translated into the language s by means of a Speech Translation process. In order to minimize the effects of errors made in this stage, the output of the speech translation process is a graph of words that represents a set of possible translations of the input sentence and its probability $p(A|W_s)$. Then, this graph is processed by a Semantic Decoding process which works in the language s and is able to deal with graph representations. The output of the Semantic Decoder is a semantic interpretation of a translation of the original sentence based on the probability $p(W_s, C)$.

3.1. Speech translation

In this work, we propose two methods for implementing the architecture described. Both methods differ in the way that the Speech Translation process is performed. The first one is based on the combination of outputs from a set of general-purpose web translators (Figure 2); in the second one, a Statistical Machine Translation system (MOSES) is trained using an automatically collected task-dependent parallel corpus (Figure 3). However, both implementations share the SLU system in the system's language s, which is able to accept graph representations as input and processes them in order to obtain a semantic representation of a translation of the user sentence.

In both methods, the Speech Translation process starts with the recognition of the input utterance by means of an ASR. Then the output of the ASR is translated into the the system's language s by using one of the following strategies.

The first option, which uses general-purpose web translators, is shown in Figure 2. We use a set of translators because we are interested in transmitting the variability generated by different translations of the user sentence, thereby increasing the coverage of the system and minimizing the effect of the errors made in this step. The second option for the translation process consists of training a task-dependent SMT system (Figure 3). The problem in this case is how to collect the parallel corpus needed to train this system. The process for obtaining this parallel corpus is discussed in Section 4. As it happened with the first option, it is convenient to provide more than a single translation of the user sentence in order to reduce the impact of translation errors in the forthcoming modules. Therefore, when the second option is performed, the n-best translations are obtained from the SMT system.

A third module concludes the Speech Translation process in both methods: the Graph of Words Builder module. This module brings together the multiple outputs provided by the previous module and builds a graph of words. This graph not only represents the sentences supplied by the translators, it also represents a reasonable generalization of them. Hence, building the graph of words this way constitutes a Grammatical Inference process in which the syntactic structures of the original sentences are generalized. Every string W_s that belongs to the language of the graph is weighted with its probability of being a translation of the original utterance A, which is the first term of Equation 4. As a result of the generalization process, the semantic decoder can consider and analyze some sentences that were not in the

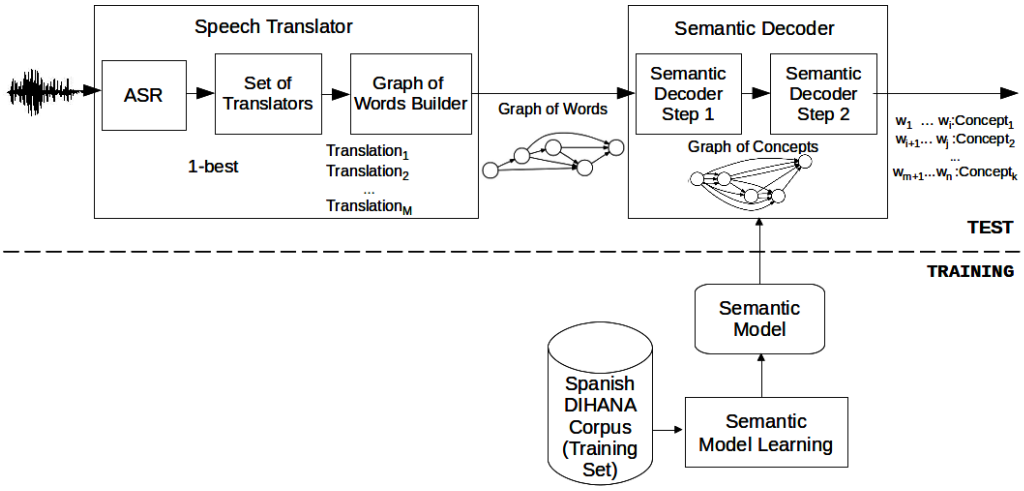

Figure 2: Architecture based on the combination of the outputs of general-purpose web translators.

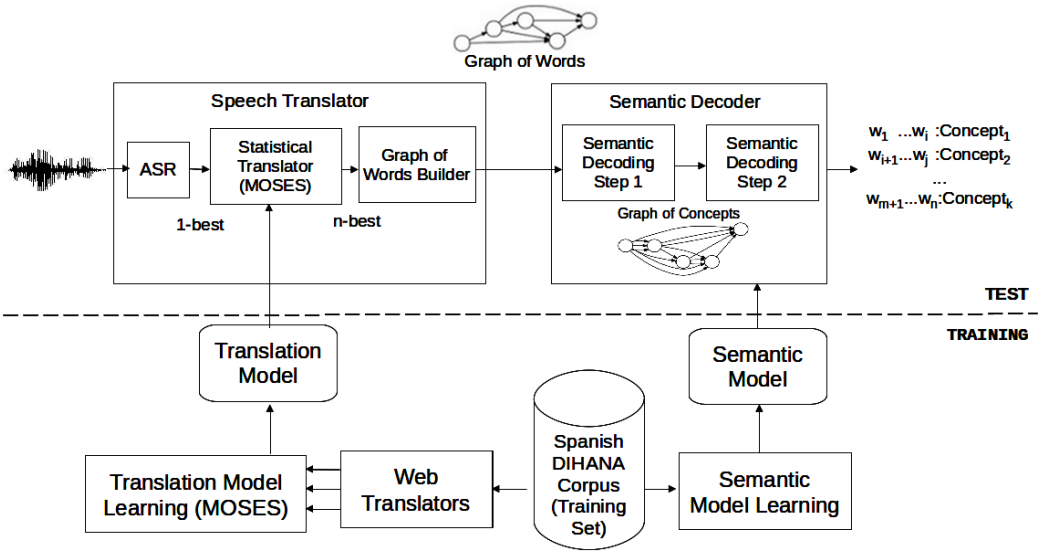

Figure 3: Architecture based on a MOSES translator.

original set of translations but that are made of pieces of them. In the example in Figure 4, the translation *"puede repetir a qué hora sale el primero"* is the most correct for the input French sentence *"pouvez-vous répéter à quelle*

heure part le premier" and can be obtained from the graph of words.

3.2. Semantic Decoding

The Semantic Decoding process is common to both implementations of the Speech Translation process (see Figure 2 and Figure 3). It provides a semantic interpretation of the input utterance, which is chosen from the language represented by the graph of words. In other words, it takes a weighted graph of words as input and provides a sequence of words and its semantic interpretation as output. This sequence of words is one of the paths in the input graph, and the semantic interpretation is represented as a segmentation and labeling in terms of concepts of this sequence of words.

This Semantic Decoding process works in two steps, which are performed taking into account the information of a stochastic semantic model that is trained in the system language s. Let W_s be the translated sentence, the probability $p(W_s, C)$ is calculated by considering the stochastic semantic model:

$$p(W_s, C) = p(W_s|C) \cdot p(C) \tag{5}$$

The stochastic semantic model provides a language model of concepts, which allows to compute $p(C)$, and a model that associates word sequences to sequences of concepts, $p(W_s|C)$. The first step of the Semantic Decoding process builds a graph of concepts from the input graph of words. We call it a graph of concepts because its arcs are labeled with pairs (sequence of words, concept). To obtain these arcs, a Dynamic Programming (DP) algorithm is performed.

Let $\mathscr{C}$ be the set of concepts, and let $c \in \mathscr{C}$ be a concept. This DP algorithm takes into account both the probability $p(A|W_s)$ provided by the graph of words, and the probability of a sequence of words between positions m and n, $W_s^{m,n}$, given a concept $p(W_s^{m,n}|c)$, which is supplied by the stochastic semantic model. It is worth noting that, in this case, the conditional probability of a sequence of words attached to a concept c given that concept, $p(W_s^{m,n}|c)$ is equivalent to the conditional probability of this sequence of words given the full sequence of concepts C, since in the final result there will not be any overlapping between the sequence of words associated to the concepts. Therefore, every word of the final result will be only attached to one concept. This graph of concepts is a concise representation of the possible semantic interpretations of the sentences encoded in a graph of words.

The second step of the Semantic Decoding process takes a graph of concepts and analyzes it in order to find the best sequence of concepts $\hat{C}$. In order to fulfill Equation 4, the probabilities represented in the graph of concepts are combined with the probability of the sequence of concepts $p(C)$. Hence, $\hat{C}$ can be obtained by searching for the path in the graph of concepts that maximizes Equation 4. The result of this step is not only the best sequence of concepts, it is also a translation of the input utterance and a segmentation of it in terms of concepts.

Finally, the segmentation obtained in this second step is converted into a standard frame representation by discarding semantically irrelevant segments and reordering the rest of the segments in a canonical way. This process is done by the Frame Converter module shown in Figure 1.

4. Automatically learning a task-dependent translation system

The use of SMT systems requires the availability of a large enough amount of parallel training data to adequately train the parameters of the translation models. However, obtaining task-specific training data by translating the original data by hand is very expensive and time-consuming. A solution to this problem is to use several general-purpose web translators to automatically translate the task-specific training sentences into another language.

In (García et al., 2014) we presented an approach that attempts to take advantage of these translation resources in order to train a task-dependent translation system. In other words, given the training sentences in the SLU system language s, these sentences are translated to the user language u by using several general-purpose web translators. This way, we build a parallel corpus where each sentence has different translations associated to it. From this parallel corpus, we train a SMT model (using MOSES) that is specific for the task. It should be noted that by means of this process, the learned translator can represent and modelize the variability generated by the different general-purpose web translators; however it could include some translation errors.

Due to the difficulty of the problem, we cannot guarantee that the best translation obtained by this SMT model is consistent with the meaning of the original sentence. Therefore, it is convenient for the SMT system to supply more than one hypothesis to the SLU module. Moreover, we do not think that separately processing the n-best translated sentences (for each input sentence) generated by the translator is the best solution; it would be

better to adequately combine segments of different translations. Thus, we have developed an algorithm to build a graph of words from the set of n-best translated sentences as described in the following section.

5. Generation of the graphs of words

The process for generating a graph of words in the system language from a set of translations is as follows:

1. *Alignment of the alternative translations.* The set of translated sentences is processed using a Multiple Sequence Alignment (MSA) algorithm resulting in an alignment matrix.
2. *Construction of the graph of words.* From the alignment matrix, a graph of words that represents a generalization of the multiple translated sentences is obtained.

A Multiple Sequence Alignment (MSA) algorithm generates an alignment of sequences of symbols that minimizes the global number of edit operations among all the sequences. Although some of the MSA algorithms were originally developed for the alignment of sequences of biological elements, they can be adapted to other tasks like speech recognition or translation (Sim et al., 2007; Bangalore et al., 2001). The approach presented in this paper uses this kind of algorithm to obtain an alignment of the set of translated sentences, and this alignment is post-processed to generate the graph of words. Specifically, we have adapted the ClustalW MSA algorithm (Larkin et al., 2007), considering that the set of symbols are the words and that all symbol substitutions are equiprobable.

An example of an alignment matrix generated by our adaptation of the ClustalW MSA algorithm is shown in Figure 4. Each row in the matrix represents an aligned sentence, and the columns represent the synchronization points. The special symbol '-' represents that no word of the sentence has been aligned in this position with any other word of the other sentences. From this alignment, we build a graph of words representing not only the translated sentences but also an extra-language that includes the regularities or common segments of words as well as alternative translations of some segments. In the example in Figure 4, the language represented by the graph of words includes the sentence *"puede repetir a qué hora sale el primero"*, which is not in the set of translated sentences

source sentence | *pouvez-vous répéter à quelle heure part le premier*
(could you repeat what time the first one departs at)

(1) **multiple translations**

poded a qué hora sale el premier
puede repetir a qué hora marcha el primero
poded a qué hora sale el premier
puede repetirse a qué hora parte el primero
pueda repetirse a qué hora parte el primero
puede me repetir a qué hora marcha el primero

(2) **alignment matrix**

-	-	poded	a	qué	hora	sale	el	premier
puede	-	repetir	a	qué	hora	marcha	el	primero
-	-	poded	a	qué	hora	sale	el	premier
puede	-	repetirse	a	qué	hora	parte	el	primero
pueda	-	repetirse	a	qué	hora	parte	el	primero
puede	me	repetir	a	qué	hora	marcha	el	primero

(3)

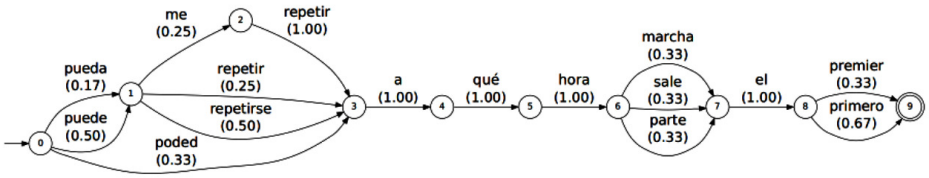

Figure 4: Steps for obtaining the graph of words from the original sentence *"pouvez-vous répéter à quelle heure part le premier"*, (*"could you repeat what time the first one departs at"*).

To build this graph of words from the alignment matrix, the following algorithm is used:

1. A set of nodes N (corresponding to the columns of the matrix plus one for the initial node) is created.
2. If it does not already exist, an arc is created for each cell in each row that contains a word. The origin node of this arc is the one that represents the column of the previous word in the same sentence. The destination node of the arc is the one that represents the column to which the cell belongs. The arc is labeled with the word in the cell and its frequency is set to 1. If the arc already exists, its frequency is increased by 1.
3. The frequencies of the arcs are normalized to represent probabilities.

Thus, the product of the weights of all the arcs in a full path representing the sentence W_s in the system language may be considered to be the probability $p(A|W_s)$.

6. Semantic decoding

As stated in Section 3.2, the Semantic Decoding is completed in two steps.

6.1. Step 1

Since every arc in the graph of words is labeled with a word, any path between any pair of nodes represents a segment of words. Some of these segments $W_s^{m,n}$ are semantically relevant to one or more concepts of the set of concepts $\mathscr{C}$ defined for the task. Therefore, it is possible to build a new graph in which the nodes are the same as the graph of words and the arcs are labeled with sequences of words and the concept to which they are associated. Also, each of these arcs can be weighted with a combination of the probability of the path in the graph of words, $p(A|W_s^{m,n})$, and the probability of the segment given a concept, $p(W_s^{m,n}|c)$.

To compute the probability of a sequence of words given a concept $p(W_s^{m,n}|c)$, a stochastic model of the lexical structures in the target language associated to each concept is needed. Training an n-gram language model (LM) for each concept (i.e. the set $\{LM_c \mid c \in \mathscr{C}\}$) fulfills this requirement. In order to estimate these LMs, it is necessary to have a segmented and labeled training corpus. For our semantic decoding algorithm, we have represented these LMs as stochastic finite state automata.

Given the acyclic nature of the graphs of words, it is possible to establish a topological order between their nodes. This topological order allows these graphs to be processed from left to right.

Formally, we define a graph of concepts $GC = (N_{GC}, E_{GC})$ built from a graph of words $GW = (N_{GW}, E_{GW})$ and the set of LM_c for each $c \in \mathscr{C}$ as:

- $N_{GC} = N_{GW}$
- $E_{GC} = \{(i, j, W_s^{i,j}, c, wgt) \mid wgt > wgt' \quad \forall (i, j, W_s^{'i,j}, c, wgt')\}$,

  where the 5-tuple (i, j, $W_s^{i,j}$, c, wgt) represents an arc in the graph of concepts, where i and j are respectively the source and the ending nodes of the arc ($i < j$); $W_s^{i,j}$ is a sequence of words associated to a path from node i to node j in the graph of words; c is the concept to

which the sequence of words is associated; and wgt is the weight of the arc, defined as $wgt = p(A|W_s^{i,j}) \cdot p(W_s^{i,j}|c)$.

We can build this set of arcs E_{GC} without any loss of information in the following way. For each concept $c \in \mathscr{C}$ and each state q_c in the automaton representing LM_c, the path in the graph of words that maximizes $p(A|W_s^{i,j}) \cdot p(W_s^{i,j}|c)$ for any pair of nodes i, j with $i < j$, arriving to the state q_c, must be found. This can be done by means of the following Dynamic Programming (DP) algorithm:

$$M(i,j,q_c) =$$

$$\begin{cases} 1 & \text{if } i = j \wedge q_c \text{ is the initial state of } LM_c \\ 0 & \text{if } i = j \wedge q_c \text{ is not the initial state of } LM_c \\ 0 & \text{if } i > j \\ \max\limits_{\substack{\forall a \in E_{GW} : \mathrm{dest}(a) = j \\ \forall (q'_c, \mathrm{wd}(a), q_c) \in LM_c}} M(i, \mathrm{src}(a), q'_c) \cdot p(q'_c, \mathrm{wd}(a), q_c) \cdot \mathrm{wt}(a) & \\ & \text{otherwise} \end{cases} \tag{6}$$

Where a is an arc in the graph of words and $\mathrm{dest}(a)$ refers to the destination node of the arc a in the graph, $\mathrm{src}(a)$ refers to its source node, and $\mathrm{wd}(a)$ and $\mathrm{wt}(a)$ refer to the word and the weight attached to the arc, respectively. Also, $(q'_c, \mathrm{wd}(a), q_c)$ represents a transition from the state q'_c to the state q_c labeled with $\mathrm{wd}(a)$ in the automaton LM_c. Thus, $M(i,j,q_c)$ represents the best path in the graph of words GW that starts in the node i, ends in the node j, and its underlying sequence of words reaches the state q_c of the model LM_c.

To compute Equation 6, a DP matrix must be filled. It is important for the DP algorithm to keep track of the words that constitute the paths that maximize the expression.

Once the DP matrix has been filled, the set of arcs of the graph of concepts can be obtained as follows:

$$\begin{aligned} &\forall i, j, i < j \in N_{GC}, \forall c \in \mathscr{C} : \\ &(i, j, W_s^{i,j}, c, wgt) \in E_{GC} \text{ if } wgt = \max_{q_c \in LM_c} M(i,j,q_c) \end{aligned} \tag{7}$$

where $W_s^{i,j}$ represents the sequence of words underlying the path in the graph of words that satisfies the maximization given i, j, and c. This definition allows us to find the set of paths between each pair of nodes i, j in the graph

of words that maximizes the combined probability of the translation and semantic models for each of the concepts of the task. Figure 5 shows a piece of a graph of concepts $GC = (N_{GC}, E_{GC})$ for the DIHANA task.

6.2. Step 2

The final step for the semantic decoding consists of finding the best sequence of concepts $\hat{C}$. This is now equivalent to finding the full path in the graph of concepts that maximizes the combined probability of the path and the sequence of concepts, $p(C)$, associated to the path. The probability $p(C)$ can be computed by means of a LM of sequences of concepts. This best path in the graph not only provides the best sequence of concepts $\hat{C}$, it also provides the sentence associated to it and a semantic segmentation of this sentence.

It is worth noting that no data in the user language is required to perform this semantic decoding procedure since the semantic model is trained only with data in the target language. This allows this methodology to be easily ported to many languages.

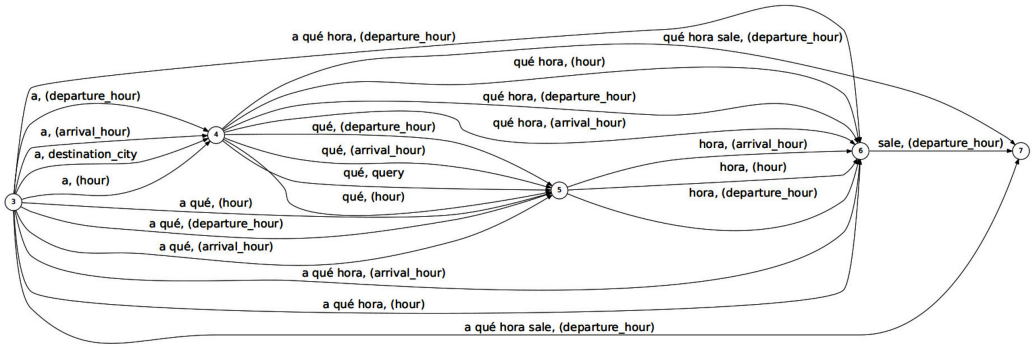

Figure 5: A piece of a graph of concepts (probabilities omitted for clarity).

7. The corpus

7.1. DIHANA corpus

We have evaluated this methodology in the framework of the DIHANA task (Benedí et al., 2006). The goal of the DIHANA task is to access a Spoken Dialog System by phone to ask for information about railway timetables

and fares. For training and testing purposes, a corpus of 900 dialogs in Spanish acquired by 225 speakers using the Wizard of Oz technique was generated, with a total of 6,229 user turns (10.8 hours of speech uttered). The vocabulary has a size of 811 words, and the average number of words per turn is 7.6. Three scenarios were defined and posed to the speakers:

- In the first scenario, the aim of the user is to obtain the timetables for a one-way trip.
- In the second scenario, the users were told to obtain the price of the tickets (and optionally the timetables) of one-way trains.
- The third scenario was analogous to the second one but for a round trip.

In order to use this corpus for SLU tasks, a semantic labeling was performed. Thirty semantic labels were defined, and all the user turns were manually and completely segmented and labeled in terms of these labels. The labeling process as well as the definition of the set of semantic labels itself were developed in such a way that each sentence is associated to a sequence of semantic labels and a segmentation of it in terms of these labels (one semantic label per segment).

For example, the sentence in Spanish *"Me podría decir los horarios de trenes para Barcelona este jueves?" ("Could you tell me the train timetables to Barcelona next Thursday?")* would be segmented this way (the special symbols <> denote a question about the concept that is between the symbols):

```
me podría decir (Could you tell me):  courtesy
los horarios de trenes (train timetables):  <time>
para Barcelona (to Barcelona):  destination_city
este jueves (next Thursday):  date
```

Then, a set of rules translates this intermediate representation in terms of frames, which consist of a set of concepts and their associated attribute-value pairs. This is performed by the Frame Converter module in Figure 1. Since the intermediate language is close to the frame representation, this module only requires a small set of rules to construct the frame. This process consists of the following: the deletion of irrelevant segments of the input sentence; the reordering of the relevant concepts and attributes that appeared in the user

sentence following an order which has been defined a priori; the automatic instantiation of certain task-dependent values, etc.

For example, the sentence *"Me podría decir los horarios de trenes para Barcelona este jueves?" ("Could you tell me the train timetables to Barcelona next Thursday?")* is translated as follows:

(*HOUR*)
Destination-City: Barcelona
Departure-Date: [Thursday-2014-11-12]

Table 1 shows some characteristics of the semantically labeled corpus.

Table 1: Characteristics of the semantically labeled corpus.

Number of user turns:	6,229
Total number of words:	47,222
Vocabulary size:	811
Average number of words per user turn:	7.6
Total number of semantic segments:	18,588
Average number of words per semantic segment:	2.5
Average number of segments per user turn:	3.0
Average number of samples per semantic unit:	599.6

7.2. Multilingual DIHANA corpus

In the test-on-source approach to Multilingual Spoken Language Understanding, the input test sentences are translated to the original language of the SLU system. This means that the SLU module should be fed by a translation module that translates the input utterances. One option is to develop a good-performance task-specific Machine Translation system, since mistakes during the translation process can produce many errors in the SLU output (see Section 4).

This task-specific Machine Translation system must be estimated from a parallel corpus. In (García et al., 2014), we presented a method to obtain a parallel corpus using general-purpose web translators; specifically, we presented how to obtain two parallel French-Spanish and English-Spanish corpora, from the original corpus in Spanish, the DIHANA corpus.

Since these translators usually make many errors, our proposal was based on the combination of several of them. Because several hypotheses are generated by different translators, there are more possibilities for the correct translation to appear in one of the translated sentences or in a combination of them. Therefore, the input test sentences are translated into the original language of the SLU system using the different general-purpose web translators. This way we have several hypotheses for each sentence in the parallel corpus. An example is shown in Figure 6.

Spanish:
Quisiera horarios para el próximo sábado a Barcelona por la mañana

French:
Serait planifier pour le samedi matin à Barcelona
J'aimerais planifier pour samedi prochain à Barcelona le matin
Il voulait horaires pour le prochain samedi à Barcelona par le matin
Il voudrait des horaires pour samedi prochain à Barcelona le matin

English:
I would like schedule for Saturday morning to Barcelona
I would like to schedule for next Saturday to Barcelona in the morning
It wanted schedules for next Saturday to Barcelona in the morning
Would want schedules for next Saturday to Barcelona in the morning

Figure 6: Example of multiple translations.

8. Experiments and results

The Spanish corpus was split into a training set of 4,889 turns, a development set of 340 turns and a test set of 1,000 turns. Both the development and test sentences were manually translated into English and French. All of the English development and test sentences were uttered by English native speakers. For the French dataset, 500 of the test sentences were uttered by native French speakers. The names of Spanish cities were changed for names of cities in English and French, respectively. The word error rates obtained by the Google speech recognizer were 17.6 for Spanish, 20.0 for English and 19.5 for French.

We performed a series of experiments under the test-on-source strategy for Multilingual Spoken Language Understanding, where the stochastic semantic

model was learned using the training corpus in Spanish, and the English test set and the French test set were used for testing. For each test set, three kinds of experiments were performed: a first series of experiments where the Speech Translator was based on the combination of outputs from a set of four general-purpose web translators (Bing, Google, Lucy, Opentrad), Figure 2; a second series of experiments where the Speech Translator was based on a Statistical Machine Translation system (MOSES), Figure 3, that was estimated using an automatically collected task-dependent parallel corpus (see Section 4); and, a third series where the Speech Translator was based on the combination of the general-purpose translators and MOSES.

For all the experimentation, we used Witten-Bell bi-grams for both language models: the models for the sequences of words for each concept and the model for the sequences of concepts.

Some meausures were defined in order to evaluate the system:

- CER (Concept Error Rate) is measured as the minimum edit distance (insertions, deletions, and substitutions) between the sequence of concepts in the hypothesis and the sequence of concepts in the reference, normalized by the length of the concepts at the reference. This measure is calculated from the output of the Semantic Decoder, Figure 1.

- FSER (Frame-Slot Error Rate) is measured as the minimum edit distance (insertions, deletions, and substitutions) between the slots of the frames in the hypothesis and the slots of the frames in the reference, normalized by the number of slots at the reference. Slots refer to concepts and their associated attribute-value pairs. This measure is calculated from the output of the Frame Converter, Figure 1.

As we explained at the end of Section 3.2, the Frame Converter takes the output of the Semantic Decoder and generates the understanding output in terms of frames. This process is done by discarding semantically irrelevant segments and reordering the rest of the segments in a canonical way.

For reference purposes, the monolingual Spanish SLU results for the output of the Google speech recognizer were 14.2 of CER and 12.8 of FSER.

In order to evaluate the quality of the translations from English and French into Spanish using semantic knowledge, we show results using the WER and BLEU standard scores. Both measures were taken from the sentence obtained after the understanding process (the output of the Semantic Decoder).

8.1. Experiments with general-purpose web translators

Table 2 shows the results of the first series of experiments for English (general-purpose web translators) considering the correct transcription of the utterances and the output of the ASR.

Table 2: Results for English using general-purpose web translators

	Text				Speech			
Translators	CER	FSER	WER	BLEU	CER	FSER	WER	BLEU
1000	24.1	15.2	45.6	35.2	35.9	28.8	55.0	25.2
0100	24.5	17.5	47.0	37.0	39.7	33.0	56.5	28.6
0010	32.3	19.6	52.3	25.6	40.8	33.1	58.4	21.7
0001	31.9	26.5	54.2	20.6	40.4	38.3	59.8	18.3
best-SemLM	26.4	18.6	50.9	25.3	35.3	30.3	56.7	21.2
1100	20.7	11.9	43.2	40.5	32.2	25.9	52.4	30.9
1010	21.7	12.9	43.5	38.4	32.1	27.1	51.9	28.9
1001	22.1	14.4	43.5	36.8	32.7	27.6	52.7	27.3
0110	20.8	11.8	44.1	38.8	33.5	27.2	52.6	30.8
0101	24.0	18.1	45.3	37.9	34.4	30.1	53.7	30.0
0011	28.1	19.6	48.2	29.0	37.0	30.6	55.7	24.3
1110	**19.2**	**10.6**	42.5	41.1	**29.9**	**25.0**	51.6	31.9
1101	21.0	12.8	42.5	40.6	32.6	26.5	51.6	31.4
1011	21.5	13.9	43.5	38.2	31.4	26.2	51.8	30.0
0111	19.9	12.2	42.8	39.6	30.7	25.2	51.2	31.4
1111	20.8	12.9	43.6	39.2	31.8	26.3	53.1	30.9

The first column (labeled as Translators) represents the different combinations of the four translators used. For example, the last row (labeled as 1111) represents the use of the four translators for generating the graph of words that was used as input for the semantic decoding module. The five first rows can be considered as baseline for the rest of experiments because only one translator is used; the fifth row (best-SemLM) shows the results when the best translator, according to the semantic LM, is selected for each sentence.

Table 2 shows the different measures (CER, FSER, WER, and BLEU) both for correct transcription (text) and for the output of the ASR (speech). It can be observed that in general, an increase in the number of translators

(i.e. an increase in the number of hypotheses) provided better results in terms of CER and FSER, as expected. It is clear that some translators were more appropriate for the task than others, but, in general, the use of more than one translator provided better results than those considered as baseline. That is true even for the results of best-SemLM, where the best translator for each sentence according to the semantic LM is selected. However, the use of the combination of all translators did not provide the best results, Actually, in the experiments the best results corresponded to the combination 1110. This can be explained by the fact that some errors in the translation process become redundant when they are produced by two or more translators, giving high probability to these erroneous hypotheses.

It should be noted that there were better results in FSER than in CER. This is because, in CER, there are some concepts that are not semantically relevant in this task, such as "courtesy" and "null", which are an important source of errors. These concepts are ignored in the case of the frame representation (measured as FSER), which only represents the relevant semantics of the task. On the other hand, the BLEU and WER measures represent how far the output sentence (after the translation and semantic decoding process) is from the expected correct Spanish sentence. Although figures are low, the important point is that many errors in words have no influence on the semantics of the sentence. It is worth noting that when using more translators the results in terms of BLEU and WER improve, which means that the final translation is better.

As Table 2 also shows, when considering the output of the ASR, there were, logically, more errors than when the correct transcription of the utterance was considered. The tendency of the results depending on the number of translators were similar to those with the correct transcriptions. That is, an increase in the number of translators provided better results in terms of CER and FSER.

Table 3 shows the results of the experiments with French as the input language. The behavior of the system in terms of the number of translators is similar to the experiments with English. In contrast to the English case, the best combination in the French case was the use of all of the translators (four translators, row 1111). The confidence intervals at the 95% confidence level for the best results in Tables 2 and 3 are around ± 1.5 for text and ± 1.8 for speech. For all the experiments shown in this paper the confidence intervals are similar to these.

Table 3: Results for French using general-purpose web translators

	Text				Speech			
Translators	CER	FSER	WER	BLEU	CER	FSER	WER	BLEU
1000	30.7	23.1	52.4	30.1	35.3	30.7	58.7	23.1
0100	33.0	25.4	54.6	27.8	37.3	33.5	61.2	22.7
0010	27.5	18.5	50.6	30.1	30.4	26.1	58.4	21.1
0001	30.6	20.0	54.3	24.5	32.4	27.3	61.5	17.7
best-SemLM	28.9	20.6	52.1	28.1	32.3	29.3	62.2	19.1
1100	26.3	18.9	50.1	34.0	31.1	26.9	55.9	28.2
1010	22.7	15.4	46.5	38.3	27.7	24.4	53.8	28.9
1001	23.9	16.6	49.3	33.8	28.5	24.9	56.2	27.0
0110	24.9	18.0	47.9	36.5	29.7	26.3	55.8	28.5
0101	23.3	16.7	48.9	34.4	28.8	24.8	55.9	27.1
0011	25.0	16.5	48.7	33.5	28.7	23.9	56.8	24.4
1110	21.6	14.4	45.9	39.7	26.2	21.7	53.1	32.3
1101	21.7	14.8	47.4	37.3	26.5	22.6	53.9	30.7
1011	21.2	13.7	45.8	38.7	26.3	22.0	53.3	30.7
0111	21.2	14.4	45.7	39.2	26.7	23.0	53.3	31.0
1111	**20.2**	**13.2**	45.7	39.9	**24.8**	**20.8**	52.3	33.1

8.2. Experiments with MOSES translator

The second series of experiments was done by using a MOSES machine translator, that was learned as explained in Section 4. Although the MOSES translator includes some of the variability generated by the four translators used to obtain the training samples, we studied the system performance considering the n-best hypotheses as the output of MOSES. Table 4 shows the results for English (for correctly transcribed sentences, and for the ASR output) and Table 5 shows the same experiments for French.

The results using this approach were similar or slightly worse than the previous results. It is possible that the generalization learned by MOSES from the training samples (there are not many samples) was not enough to deal with some variability or errors in the test utterances.

In some cases, a slight improvement in the CER, when more than 1-best hypothesis are used, can be observed. This improvement was not always

Table 4: Results for English using MOSES

	Text				Speech			
n-best	CER	FSER	WER	BLEU	CER	FSER	WER	BLEU
1	21.9	**12.8**	44.1	38.5	31.0	**24.2**	53.1	27.9
2	**21.2**	13.3	43.5	39.0	**30.6**	25.0	52.7	28.1
3	22.7	14.8	43.5	38.8	31.5	26.0	52.9	27.9
4	21.4	13.4	43.5	38.8	30.9	25.4	52.8	28.1
5	22.5	14.7	43.2	39.3	31.5	25.7	52.4	28.3

Table 5: Results for French using MOSES

	Text				Speech			
n-best	CER	FSER	WER	BLEU	CER	FSER	WER	BLEU
1	21.9	15.2	45.8	37.4	27.4	23.1	51.5	31.0
2	21.8	15.1	45.8	37.5	26.8	22.1	51.4	31.4
3	**21.6**	**14.7**	45.8	37.6	26.3	21.8	51.2	31.4
4	21.6	15.0	45.6	37.8	26.2	21.7	51.1	31.6
5	21.6	14.8	45.4	38.1	**26.0**	**21.7**	51.0	31.7

translated to FSER, possibly because the variability introduced by the n-best is not relevant for the frame representation, because it does not affect attribute values. In any case, the approach using MOSES in the Speech Translation process also provided successful results.

8.3. Experiments combining general-purpose web translators and MOSES

With the intention of studying the complementarity of the two previous proposals, we performed a new experimentation combining the output of all the general-purpose web translators and the output of MOSES translator. Table 6 shows the results of this combination for English and French.

These results outperform the results achieved in the two previous series of experiments (subsections 8.1 and 8.2). Our approach for SLU, based on the use of graphs, allows the combination of different kinds of translators, taking advantage of the complementary information provided by each translator.

Table 6: Results combining general-purpose web translators and MOSES

	Text				Speech			
	CER	FSER	WER	BLEU	CER	FSER	WER	BLEU
English	19.3	11.2	41.2	42.8	28.8	23.3	50.2	33.7
French	19.5	12.1	44.8	41.1	22.9	19.0	51.6	35.3

8.4. Comparing results with CRFs

In order to compare the results obtained with other SLU approaches we performed some experiments using a discriminative approach, Conditional Random Fields. CRFs have been successfully used for SLU tasks (Hahn et al., 2009). We defined a standard set of features that includes lexical information, setting a window that incorporates the two previous and the two following words. In this experimentation, the Speech Translation process is done with the two approaches: using general-purpose web translators and using MOSES. After the Speech Translation process, the understanding process is done using CRFs.

Table 7 shows the results for English and French languages using general-purpose web translators for the Speech Translation process.

Table 7: Results for text and speech using CRFs in both languages for the general-purpose web translators

		Text		Speech	
	Translators	CER	FSER	CER	FSER
English	1000	24.5	17.4	**33.7**	**29.3**
	0100	**23.9**	**15.8**	34.7	29.7
	0010	31.2	22.2	37.3	31.0
	0001	33.1	26.6	40.2	35.8
French	1000	30.1	24.3	33.6	31.3
	0100	32.7	27.4	37.5	34.5
	0010	**26.2**	**22.2**	**30.4**	**31.1**
	0001	30.8	22.6	34.8	32.6

The best results of the Table 7 for each of the four experimentations (English-text, English-speech, French-text and French-speech) slightly outperform our results using a single translator, the first 5 rows in Tables 2,

and 3. On the other hand, these best results of the Table 7 do not improve the best results of the corresponding experiments of the approach presented in this work, see Table 6. In terms of CER measure, the best results are 23.9, 33.7, 26.2, and 30.4 for CRFs and 19.3, 28.8, 19.5, and 22.9 for our SLU approach. These results confirm our hypothesis that the use of various translators, appropriately combined, improves the use of only one of them separately even when CRFs are used for the understanding system, which have become the most widely used method for SLU.

In order to better compare CRFs with the other approaches that accept graph of words as input, we have performed another series of experiments where CRFs are provided with multiple hypotheses, in spirit to (Deoras et al., 2013). Table 8 shows the best results among all the possible combinations of the four general-purpose translators.

Table 8: Results for CRFs provided with multiple hypotheses

		best comb.	CER	FSER
English	Text	1100	22.1	16.3
	Speech	1110	32.2	28.6
French	Text	0011	23.1	18.5
	Speech	1011	28.9	28.6

Results show that using graph of words as input to the SLU system improves the use of only one hypothesis also for CRFs, as expected. Even so, the results are still lower than those of our SLU system (Table 6).

Finally, we have studied the behavior of the CRFs when the input is provided by the 1-best of MOSES translator. Table 9 shows the results of this approach for the English and the French languages.

Table 9: Results for text and speech using CRFs in both languages for the 1-best MOSES

		CER	FSER
English	Text	19.3	13.3
	Speech	26.8	23.1
French	Text	21.8	16.1
	Speech	24.7	23.5

Attending the CER measure, our approach outperforms the CRF results for French although they are lower for English-speech.

9. Conclusions

In this work, we propose a multilingual stochastic SLU system that takes advantage of multiple translators. We have designed an architecture where the communication between different modules is done by means of graphs. This way, the intermediate modules have the capability of generalizing from different hypotheses, which makes it possible to recover errors generated in previous phases. To do this, a specific semantic-parsing algorithm has been developed. We have proposed two test-on-source approaches for language portability. Both methods differ in the way that the Speech Translation process is performed. The first one is based on the combination of outputs from a set of general-purpose web translators; in the second one, a Statistical Machine Translation system is trained using an automatically collected task-dependent parallel corpus. We have also studied the complementarity of these two proposals by combining the output of all the general-purpose web translators and the output of the Statistical Machine Translation system.

In all cases the manual effort to adapt the system to the new languages is avoided, which was one of the objectives of the work. Consequently, an advantage of this methodology is that it can be easily ported to many languages.

The results show that the performance of the systems is good enough, taking into account that a translation process (usually a source of errors) is embedded between the input sentence and the understanding module. The results also show that the generalization inferred from the multiple translations and the inference algorithm to combine them permits some errors generated in previous phases to be recovered. This is clear in the first group of experiments, where the use of a combination of translations was better than the use of each translator separately, which only considered one hypothesis. On the other hand, although the BLEU and WER measured after the translation and semantic decoding processes indicated a low performance in terms of translation, the CER and FSER were satisfactory showing that many errors in words have no influence on the final semantic decoding.

Acknowledgments

This work is partially supported by the Spanish MEC under contract TIN2014-54288-C4-3-R and by the Spanish MICINN under FPU Grant AP2010-4193.

References

Bangalore, S., Bordel, G., Riccardi, G., 2001. Computing Consensus Translation from Multiple Machine Translation Systems. In: In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001. pp. 351–354.

Benedí, J.-M., Lleida, E., Varona, A., Castro, M.-J., Galiano, I., Justo, R., López de Letona, I., Miguel, A., May 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proceedings of LREC 2006. Genoa (Italy), pp. 1636–1639.

Calvo, M., Hurtado, L.-F., García, F., Sanchis, E., 2012. A Multilingual SLU System Based on Semantic Decoding of Graphs of Words. In: Advances in Speech and Language Technologies for Iberian Languages. Springer, pp. 158–167.

Daciuk, J., Mihov, S., Watson, B. W., Watson, R. E., 2000. Incremental construction of minimal acyclic finite-state automata. Computational linguistics 26 (1), 3–16.

De Mori, R., Bechet, F., Hakkani-Tür, D., McTear, M., Riccardi, G., Tür, G., 2008. Spoken language understanding: A survey. IEEE Signal Processing magazine 25 (3), 50–58.

Deoras, A., Tur, G., Sarikaya, R., Hakkani-Tur, D., 2013. Joint Discriminative Decoding of Word and Semantic Tags for Spoken Language Understanding. IEEE Transactions on Audio, Speech, and Language Processing.

Devillers, L., Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, K., Charnay, L., Bousquet, C., Vigouroux, N., et al., 2004. The french media/evalda project: the evaluation of the understanding capability of spoken language dialogue systems. In: LREC. Citeseer, pp. 2131–2134.

Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A., Riccardi, G., 2009. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In: Proceedings of the 2nd Workshop on Semantic Representation of Spoken Language. Association for Computational Linguistics, pp. 34–41.

García, F., Calvo, M., Sanchis, E., Hurtado, L.-F., Segarra, E., 2014. Obtaining parallel corpora for Multilingual Spoken Language Understanding tasks. In: In Proceedings of the Iberspeech. Las Palmas de Gran Canaria, pp. 208–215.

García, F., Hurtado, L., Segarra, E., Sanchis, E., Riccardi, G., 2012. Combining multiple translation systems for Spoken Language Understanding portability. In: Proc. of IEEE Workshop on Spoken Language Technology (SLT 2012). Miami, pp. 282–289.

Hahn, S., Dinarelli, M., Raymond, C., Lefèvre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., Riccardi, G., 2010. Comparing stochastic approaches to spoken language understanding in multiple languages. Audio, Speech, and Language Processing, IEEE Transactions on 6 (99), 1569–1583.

Hahn, S., Lehnen, P., Heigold, G., Ney, H., 2009. Optimizing CRFs for SLU Tasks in Various Languages Using Modified Training Criteria. In: Interspeech. Brighton, U.K., pp. 2727–2730.

Hakkani-Tür, D., Béchet, F., Riccardi, G., Tur, G., 2006. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. Computer Speech & Language 20 (4), 495–514.

He, X., Deng, L., Hakkani-Tur, D., Tur, G., 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, pp. 8342–8346.

He, Y., Young, S., 2006. Spoken language understanding using the hidden vector state model. Speech Communication 48, 262–275.

Heck, L., Hakkani-Tür, D., 2012. Exploiting the semantic web for unsupervised Spoken Language Understanding. In: Proc. of IEEE Workshop on Spoken Language Technology (SLT 2012). Miami, pp. 228–233.

Henderson, M., Gasic, M., Thomson, B., Tsiakoulis, P., Yu, K., Young, S., 2012. Discriminative spoken language understanding using word confusion networks. In: SLT. pp. 176–181.

Jabaian, B., Besacier, L., Lefèvre, F., 2013. Comparison and Combination of Lightly Supervised Approaches for Language Portability of a Spoken Language Understanding System. Audio, Speech, and Language Processing, IEEE Transactions on 21 (3), 636–648.

Koehn, P., et al., 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In: Proc. of ACL demonstration session. pp. 177–180.

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., Nov. 2007. ClustalW and ClustalX version 2.0. Bioinformatics 23 (21), 2947–2948.

Lefèvre, F., 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4. IEEE, pp. 13–16.

Lefèvre, F., Mairesse, F., Young, S., 2010. Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In: INTERSPEECH. pp. 78–81.

Maynard, H. B., Lefèvre, F., 2001. Investigating Stochastic Speech Understanding. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 260–263.

Misu, T., Mizukami, E., Kashioka, H., Nakamura, S., Li, H., 2012. A bootstrapping approach for slu portability to a new language by inducting unannotated user queries. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, pp. 4961–4964.

Ortega, L., Galiano, I., Hurtado, L.-F., Sanchis, E., Segarra, E., 2010. A statistical segment-based approach for spoken language understanding. In: Proc. of InterSpeech 2010. Makuhari, Chiba, Japan, pp. 1836–1839.

Raymond, C., Bechet, F., De Mori, R., Damnati, G., 2006. On the use of finite state transducers for semantic interpretation. Speech Communication 48, 288–304.

Raymond, C., Riccardi, G., 2007. Generative and discriminative algorithms for spoken language understanding. Proceedings of Interspeech 2007, 1605–1608.

Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L., 2002. Extracting Semantic Information Through Automatic Learning Techniques. IJPRAI 16 (3), 301–307.

Servan, C., Camelin, N., Raymond, C., Béchet, F., Mori, R. D., 2010. On the use of Machine Translation for Spoken Language Understanding portability. In: Proc. of ICASSP. pp. 5330–5333.

Sim, K. C., Byrne, W. J., Gales, M. J. F., Sahbi, H., Woodland, P. C., 2007. Consensus network decoding for statistical machine translation system combination. IEEE Int. Conference on Acoustics, Speech, and Signal Processing IV, 105–108.

Stepanov, E. A., Kashkarev, I., Bayer, A. O., Riccardi, G., Ghosh, A., 2013. Language style and domain adaptation for cross-language slu porting. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, pp. 144–149.

Täckströ, O., Das, D., Petrov, S., McDonald, R. T., Nivre, J., 2013. Token and type constraints for cross-lingual part-of-speech tagging. Transactions of the Association for Computational Linguistics 1, 1–12.

Tur, G., Deoras, A., Hakkani-Tur, D., 2013. Semantic parsing using word confusion networks with conditional random fields. In: Proc. of the INTERSPEECH. pp. 2579–2583.

Tür, G., Hakkani-Tür, D., Schapire, R. E., 2005. Combining active and semi-supervised learning for spoken language understanding. In: Speech Communication. Vol. 45. pp. 171–186.

Tür, G., Hakkani-Tür, D. Z., Hillard, D., Çelikyilmaz, A., 2011. Towards Unsupervised Spoken Language Understanding: Exploiting Query Click Logs for Slot Filling. In: Proc. of INTERSPEECH 2011. pp. 1293–1296.

Tür, G., Mori, R. D., 2011. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, 1st Edition. Wiley.

Vitae

Marcos Calvo received his M.Sc. in Artificial Intelligence from the Universitat Politècnica de València in 2011. He is currently a Ph.D. student in the Departament de Sistemes Informàtics i Computació. He joined the Natural Language Engineering and Pattern Recognition (ELiRF) research group in 2009. His current research is mainly focused on Spoken Language Understanding.

Lluís F. Hurtado received his Ph.D. degree in Computer Science from the Universitat Politècnica de València in 2004. He is currently an Associate Professor in the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València. He is member of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. He has published over 60 papers being involved in many research projects. His research interests cover many areas within speech processing and natural language processing, including: spoken dialog systems, voice-activated question answering, and sentiment analysis.

Fernando García received his Ph.D. degree in Computer Science from the Universitat Politècnica de València in 2003. He is currently an Associate Professor in the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València. He is member of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. He has published over 40 papers being involved in many research projects. His research interests cover many areas within speech processing and natural language processing, including: Spoken Language Understanding, Dialogue Systems, Language Understanding Portability.

Emilio Sanchis received his Ph.D. degree in Computer Science from the Universitat Politècnica de València in 1994. He is currently a Full Professor in the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València. He is the head of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. He

has published over 100 papers in different conferences, workshops and journals being involved in many research projects. His main research interests are focused on dialog systems, question answering and automatic learning.

Encarna Segarra received her Ph.D. degree in Computer Science from the Universitat Politècnica de València, in 1993. In 1986 she joined the Departament de Sistemes Informàtics i Computació of the Universitat Politècnica de València, where she is now Full Professor. She is an active member of the Natural Language Engineering and Pattern Recognition (ELiRF) research group at the same institution. She has published over 80 papers in different conferences, workshops and journals being involved in many research projects. Her current research interests are mainly focused on the automatic learning of language models and its application to spoken dialog systems and lexical disambiguation.