A train-on-target strategy for Multilingual Spoken Language Understanding*

Fernando García-Granada, Encarna Segarra, Carlos Millán, Emilio Sanchis, Lluís-F. Hurtado

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

Abstract. There are two main strategies to adapt a Spoken Language Understanding system to deal with languages different from the original (source) language: test-on-source and train-on-target. In the train-on-target approach, a new understanding model is trained in the target language, which is the language in which the test utterances are pronounced. To do this, a segmented and semantically labeled training set for each new language is needed. In this work, we use several general-purpose translators to obtain the translation of the training set and we apply an alignment process to automatically segment the training sentences. We have applied this train-on-target approach to estimate the understanding module of a Spoken Dialog System for the DIHANA task, which consists of an information system about train timetables and fares in Spanish. We present an evaluation of our train-on-target multilingual approach for two target languages, French and English.

Keywords: Spoken Language Understanding, Language Portability, Corpora Alignment, Train-on-Target

1 Introduction

Spoken Language Understanding (SLU) is an important challenge in human-machine interaction systems either oral or written [17, 7]. Although the semantic interpretation of a text in a semantically unrestricted universe is still far from being solved, there are SLU systems developed for tasks semantically restricted that provide reasonable results. One of the areas of application of SLU systems is Spoken Dialogue Systems for limited domains. In a large number of those systems it is necessary to obtain a template with the information to make a query to an information system. This is done over several dialog turns, so that for each turn it is necessary to obtain the semantic information provided by the user, i.e., the specific data that have been provided as well as the information on the intention behind the turn.

* This work has been partially funded by the projects ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics (MEC TIN2014-54288-C4-3-R).

Generally, the aim of a SLU system is to provide a semantic interpretation of the input sentence in terms of some semantic units (or concepts), and to identify the relevant information (or values) that are attached to each of them. The semantic units are defined beforehand according to the nature of the task and represent both the user intention and the types of pieces of information that are expected to be provided to the system.

For the construction of SLU systems different statistical approaches can be found. Some of these approaches are based on Markov Models or Stochastic Grammars [16, 14, 9, 5, 8, 15]. They are also approaches based on discriminative models such as Support Vector Machines (SVM) or Conditional Random Fields (CRF) [4, 13, 12]. In all these approaches one of the main problems that must be addressed is the segmentation of the input sentence, since the goal is not only to obtain one or more classes associated to a sentence but also the text segment that corresponds to each semantic meaning found, considering the context of the whole sentence. This is the main reason why the CRF are the best solution among the discriminative models, since in its decision the whole sentence participates jointly and associates each word/segment with a meaning.

The process of segmenting and labeling the training corpus, in most cases manually done, is a very time-consuming task which makes the adaptation of SLU systems to different tasks or languages difficult and expensive. 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.

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 years [3, 10, 2, 6, 16] in both test-on-source and train-on-target approaches. An essential aspect to ensure the viability of this kind of SLU systems is the performance of the translation process. If we use Statistical Machine Translation (SMT) systems, such as MOSES [11], 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.

The work presented in this paper addresses the problem of developing a multilingual SLU system that translates the training corpus to learn models in the target language, that is, the work presents a train-on-target strategy. Applying this strategy involves the estimation of SLU models in the target language, and to do this, a training set in each new language is needed. In this work we have used several general-purpose translators. Due to the good performance of the CRF-based SLU systems, we estimate CRF SLU models from the translated training set.

We have applied this train-on-target approach to the SLU module of a Spoken Dialog System for the DIHANA task [1], which consists of an information system about train timetables and fares in Spanish. To evaluate the multilingual approach, we have acquired a French and an English corpus for testing, which consists of written and spoken sentences. In a previous work [3], we applied a test-on-source approach to the same task.

2 The DIHANA corpus

The Spanish DIHANA corpus is a set of 900 dialogs in Spanish in a telephone-based information service for trains. The corpus was acquired using the Wizard of Oz technique, it contains therefore many phenomena of spontaneous speech. 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, and the third scenario was analogous to the second one, but considering a round trip. The corpus has a total of 10.8 hours of speech recordings and 225 speakers.

In order to use this corpus for SLU tasks, a semantic labeling was performed. A total amount of 30 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 segment per semantic label). For example, the sentence in Spanish "Me podría decir los horarios para Barcelona este jueves?" (Could you tell me the timetables to go 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 : courtesy
los horarios de trenes: <time>
para Barcelona : destination_city
este jueves: date
```

Some characteristics of the semantically labeled corpus are shown in the following table.

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

3 Spoken Language Understanding

The Spoken Language Understanding problem can be approached as the search of the concept sequence that represents the meaning of the sentence. Each concept represents the meaning of a sequence of words (a segment) of the sentence, as it is shown in the example of Section 2. The output of the understanding system is a sequence of (*segment*, *concept*) pairs.

In Figure 1 a scheme of the understanding process is presented, including both training and test processes. For the training process the input sentences should be segmented and labeled in terms of concepts. In the test process, given an input sentence $w = w_1, w_2, \ldots w_\mathrm{N}$, the understanding process provide a sequence of (*segment*, *concept*) pairs $(w_1 \ldots w_j, c_1)$, $(w_{j+1} \ldots w_t, c_2)$, $\ldots$, $(w_{k+1} \ldots w_N, c_n)$.

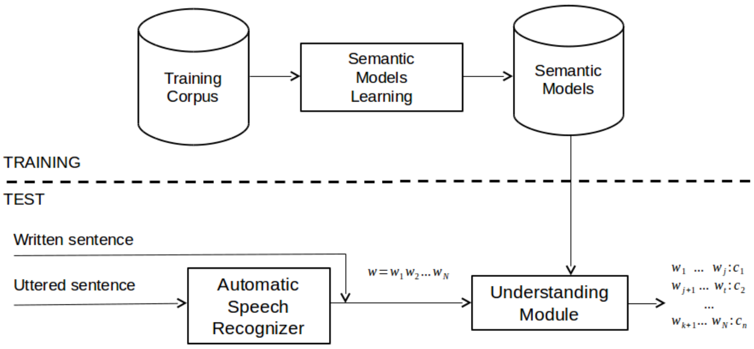

Fig. 1. Understanding process scheme

Given a training set of segmented and labeled sentences, the understanding models should be learned. As mentioned in Section 1, there are different ways to model the lexical, syntactic and semantic constraints. In this work we present a CRF-based approach.

4 Segmentation and labeling of the translated training corpus

For a train-on-target approach to the SLU problem it is necessary to translate the training corpus, as well as to provide it with a semantic labeling and a segmentation. The most straightforward technique consists on translating separately each one of the segments associated to the semantic labels. This way the sequence of semantic labels is directly translated to the new language. To obtain the complete sentence in the new language it is necessary to concatenate the segments in the order provided by the sequence of semantic labels. This technique presents some drawbacks, the segment concatenation can generates non correct sentences, especially when short segments are translated, because they are translated without considering the context.

Since the translation of very short segments can generate many errors, because the automatic translations take into account the context of the words in the sentence, we have explored another approach that is based on a complete translation of the sentences and a posterior segmentation. Figure 2 shows the scheme of the proposed translation approach for the training corpus.

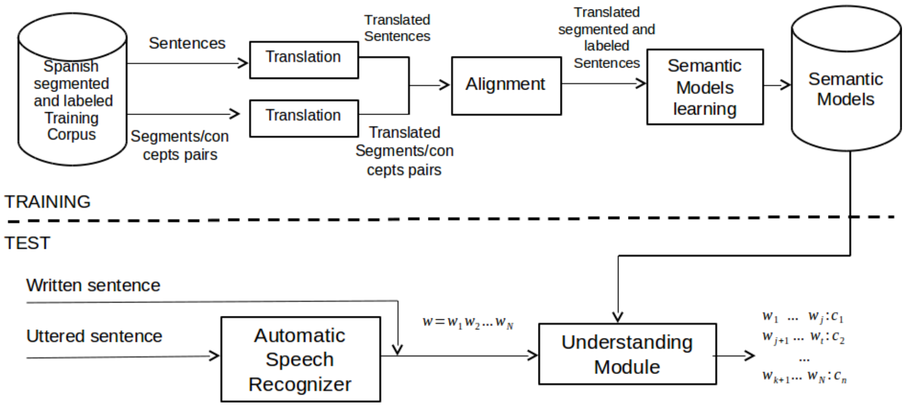

Fig. 2. Translation and understanding processes scheme

In a first phase a complete translation of the training sentences is performed, as well as the translation of the segments associated to concepts. In a second phase a segmentation and labeling of the complete translated sentences is performed by means of an alignment of that sentences with sentences built by the concatenation of the corresponding translated segments. This alignment is performed by minimizing the Levenshtein distance. This way a segmentation is induced in the complete translated sentence, and the semantic labels can be associated to the obtained segments. In this approach we assumed that the sequence of semantic units is the same in both languages. Figure 3 shows the translation and alignment of the Spanish sentence ”*Quiero conocer el precio de los trenes hacia Ávila*”.

Spanish:	Quiero conocer	el precio de los trenes	hacia Ávila
English:	I want to know	the price of the train	to Ávila
French:	Je veux savoir	le prix du train	à Ávila
Semantic labels:	query	<price>	destination_city

Fig. 3. Example of translation and alignment

5 Experimental work

In order to study the correctness of our train-on-target proposal for Multilingual Spoken Language Understanding, some experimental work was carried out. The source language was Spanish and the target languages were English and French.

The DIHANA corpus was used both to learn the models and to do the testing of the system. Particularly, 4,887 turns were used as training set and 1,000 turns were used as test set.

DIHANA corpus contains only sentences in Spanish. To get test sentences in English and French a manual translation process was performed. The test set was translated into English by six native speakers. In addition, they also uttered the 1,000 turns. In a similar way for French, the test turns were translated into French by four native speakers. But, for various reasons, only 500 of the 1,000 sentences were uttered. The test sentences uttered by the native speakers were recognized using the Automatic Speech Recognizer (ASR) of Google. The Word Error Rate obtained was 20.0 for English and 19.5 for French.

Four types of experiments were performed: English and French as target language and text input (correct transliteration) or speech input (ASR by Google). In all cases, except for French-Audio, 1,000 samples were used for test.

It is necessary to translate the training corpus into the target language to apply the train-on-target approach. Five free general-purpose on-line translators (T1:Apertium, T2:Bing, T3:Google, T4:Lucy, T5:Systranet) were selected to do this translation. The technique used to segment the translated training sentences was based on minimizing Levenshtein distance, as discussed in the Section 4. In regard to the formalism used to learn the understanding models, it should be noted that, in all the experiments, Conditional Random Fields formalism has been used. Two previous and two subsequent words were considered as context.

The first series of experiments consisted of a comparison of the performance of the understanding system depending on the translator used to translate the training set.

As comparison measure, the Concept Error Rate (CER) was selected. CER is a well known and used measure to evaluate understanding systems. It can be seen as the equivalent to WER when, instead of words, semantic labels are considered. Furthermore, to perform a better comparison, the confidence intervals at 95% of all experiments were also computed. The values for the confidence intervals were around ±1.5 for text.

Table 2 shows the results obtained for English as target language, both for text input (Text column) and speech input (Speech column). Each row shows the performance of the system when the training corpus is translated using each one of the considered translators. For reference purposes, the CER results for monolingual Spanish SLU were 9.6 for text and 16.8 for speech.

Table 2. Results for English and the five considered translators in terms of CER

Translator	Text	Speech
T1	30.5	37.6
T2	24.8	30.4
T3	**23.0**	**28.1**
T4	34.8	40.4
T5	31.1	38.6

Significant differences can be observed depending on the translator considered. The best results are obtained when the T3 is used for both input text and audio. It should be noted that, although the system obtains worse results when speech input is used, this deterioration is not as one could expect (from 23.0 to 28.1) considering that the WER of the recognition process was 20.0. This may be because many of the misrecognized words did not have important semantic information.

The same experimentation was repeated for French as target language. Table 3 shows the results obtained. It must be remembered that, in the case of French and speech input, only 500 turns were used for testing.

Comparing the two tables (Table 2 and Table 3) it can be seen that better results are obtained for French, but the overall behavior is quite similar: there are significant differences among translators and speech input produces worse results than text input, but not as bad as it could be expected.

To study the complementarity of the translators, a second series of experiments was carried out. Instead of learning the understanding model with the training sentences translated by a single translator, in these experiments, we

Table 3. Results for French and the five considered translators in terms of CER

Translator	Text	Speech
T1	35.9	38.8
T2	24.5	25.9
T3	**21.7**	**25.1**
T4	27.4	32.4
T5	26.9	31.7

learned the models with the union of the sentences translated by two, three, four, and even five (all) translators.

To simplify the display of the combinations in the tables they have been coded in a binary form. As there are five translators, we used a sequence of 5 bits where each bit indicates whether or not the corresponding translator has been used in a combination.

Although all combinations have been tested, because of space problems, only the best performing combinations are shown below. Table 4 shows the results of the best combinations of translations for English as target language for both text and speech input. Each block of rows shows the best results for a specific number of translators, from a single translator (first block) to all translators together (last block).

Table 4. Results for English and the best combinations of the five considered translators in terms of CER

Combination	Text	Speech
00100	23.0	28.1
01100	22.5	27.6
10100	21.9	27.6
10101	**21.5**	27.1
10110	21.6	**27.0**
01111	22.2	27.2
10111	**21.5**	27.1
11111 (all)	21.9	**27.0**

Analyzing the results, some conclusions can be drawn: not always to use more translators produces better results; in all the best combinations the best single (00100) is used, this is not true for the second best (01000); translators with individually bad results appear in combinations with good results, probably this is due to their complementarity with the best translators. Unfortunately, differences in the results are not statistically significant at 95% and the conclusions could not be entirely correct.

The same experimentation was repeated for French as target language. Table 5 shows the results of the best combinations.

Table 5. Results for French and the best combinations of the five considered translators in terms of CER

Combination	Text	Speech
00100	21.7	25.1
00101	20.1	22.0
00110	20.3	23.1
00111	20.0	**21.7**
10101	19.9	22.1
11100	19.9	23.4
01111	**19.8**	22.3
10111	20.1	22.0
11101	**19.8**	22.7
11111 (all)	**19.8**	22.2

As happened in the case of individual translators, the results for French are slightly better than the results for English (perhaps because Spanish and French are closer languages). But still, the conclusions are similar in both languages.

6 Conclusions and future works

In this paper, we have applied a train-on-target approach to the SLU module of a Spoken Dialog System for the DIHANA task. Significant differences can be observed depending on the general-purpose translator used to translate the training set. It can be observed that, in general, the French results are lightly better than the English ones, as we expected due to the fact that French and Spanish are closer than English and Spanish. We can also observe that not always the use of more translators provided better results, in fact, the use of all the translators is not the best combination of them, but is very close to the best. Anyway, it can be concluded that the use of multiple translators improves the results of each one separately.

Although the results slightly improved those obtained with the approach test-on-source [3], the complete process in a train-on-target approach for multilingual SLU is more complex because we have not only to translate a larger set, the training set is usually at least 5 times the test set, but also we have to obtain the segmentation and labeling of the training translated sentences. In counterpart, the use of a train-on-target approach is more efficient because an on-line translation is not necessary during the real use.

As future works, we proposed to explore other alignments strategies allowing, for example, a reordering in the sequence of semantic segments.

References

1. 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: LREC 2006. pp. 1636–1639 (2006)

2. Calvo, M., Hurtado, L.F., García, F., Sanchis, E.: A Multilingual SLU System Based on Semantic Decoding of Graphs of Words. In: Advances in Speech and Language Technologies for Iberian Languages, pp. 158–167. Springer (2012)
3. Calvo, M., Hurtado, L.F., Garca, F., Sanchis, E., Segarra, E.: Multilingual spoken language understanding using graphs and multiple translations. Computer Speech and Language 38, 86–103 (2016)
4. Dinarelli, M., Moschitti, A., Riccardi, G.: Concept Segmentation And Labeling For Conversational Speech. In: Interspeech. Brighton, U.K. (2009)
5. Esteve, Y., Raymond, C., Bechet, F., Mori, R.D.: Conceptual Decoding for Spoken Dialog systems. In: Proc. of EuroSpeech'03. pp. 617–620 (2003)
6. García, F., Hurtado, L., Segarra, E., Sanchis, E., Riccardi, G.: Combining multiple translation systems for Spoken Language Understanding portability. In: Proc. of IEEE Workshop on Spoken Language Technology (SLT). pp. 282–289 (2012)
7. 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), 1569–1583 (2010)
8. He, Y., Young, S.: A data-driven spoken language understanding system. In: Proc. of ASRU'03. pp. 583–588 (2003)
9. Hurtado, L., Segarra, E., García, F., Sanchis, E.: Language understanding using n-multigram models. In: Advances in Natural Language Processing, Proceedings of 4th International Conference EsTAL, Lecture Notes in Computer Science, vol. 3230, pp. 207–219. Springer-Verlag (2004)
10. Jabaian, B., Besacier, L., Lefèvre, F.: Comparison and Combination of Lightly Supervised Approaches for Language Portability of a Spoken Language Understanding System. pp. 636–648 (2013)
11. Koehn, P., et al.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proc. of ACL demonstration session. pp. 177–180 (2007)
12. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning. pp. 282–289. Citeseer (2001)
13. Lefèvre, F.: 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, pp. 13–16. IEEE (2007)
14. 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. pp. 1836–1839. Makuhari, Chiba, Japan (2010)
15. Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L.: Extracting Semantic Information Through Automatic Learning Techniques. IJPRAI 16(3), 301–307 (2002)
16. Servan, C., Camelin, N., Raymond, C., Bchet, F., Mori, R.D.: On the use of Machine Translation for Spoken Language Understanding portability. In: Procs. of ICASSP'10. pp. 5330–5333 (2010)
17. Tür, G., Mori, R.D.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Wiley, 1 edn. (2011)