Obtaining parallel corpora for Multilingual Spoken Language Understanding tasks

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

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

Abstract. Many Spoken Language Understanding systems are based on statistical methods like Stochastic Finite State Automata and Classification Techniques. Although many efforts have been made to develop semi-supervised and unsupervised learning techniques for semantic modeling, almost all of the statistical methods are supervised, hence, it is necessary to have a semantically labeled training corpus in order to learn the models. The multilingual approaches to Spoken Language Understanding can be grouped in two classes, train-on-target and test-on-source. In both approaches the translation of the corpus from the original language to other languages is needed. In this work, we present how to obtain a translated corpus from the original one to estimate the new Spoken Language Understanding system and how to obtain a parallel corpus to estimate a task-specific machine translation system. We also present a methodology to translate not only the sentences, but also the semantic labels and the segmentation of the corpus. Finally we present some preliminary experiments using both approaches.

Keywords: Multilingual Spoken Language Understanding, Parallel corpora, Machine translation.

1 Introduction

Nowadays, automatic language portability of different speech-based systems is an important challenge. Finding a efficient solution to this problem would imply saving a lot of time, money and effort in translating and labeling corpora, as well as adapting and re-training the systems that already work for some language. Also, having multilingual systems available, allows the study of the behavior and robustness of the speech-based systems when the language changes. This would make it possible to identify the strengths and weaknesses of the models and the systems depending on the input language, which would probably lead to some interesting language-dependent improvements. Although the goal of making speech-driven systems for many languages in a totally unsupervised manner seems to be still quite far, Machine Translation (MT) technology can help us to address this problem in a semi-supervised way.

One kind of these speech-based systems is limited-domain spoken dialog systems, which in the last few years has received many efforts from the research community. They try to provide a fluent speech-based interaction between a human and a computer, in the context of a well defined task. An important part of these systems is the Spoken Language Understanding (SLU) component. Its aim 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 one of them. The semantic units are defined beforehand according to the nature of the task and represent both the user intention and the different types of pieces of information that are expected to be provided to the system. Many SLU systems are based on statistical methods like Stochastic Finite State Automata and Classification Techniques [7, 9, 5, 6, 3]. Although many efforts have been made to develop semi-supervised and unsupervised learning techniques for semantic modelization [10, 8], almost all the statistical methods are supervised, hence, it is necessary to have a semantically labeled training corpus in order to learn the models.

The multilingual approaches to SLU can be grouped in two classes, so-called train-on-target and test-on-source. In the train-on-target approach a new SLU model is trained in the target language, that is the language in which the user 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. 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. This hand-based translating process requires a lot of effort, which makes it very interesting to develop a methodology to perform this step automatically.

In the test-on source approach the input sentences are translated into the original language of the corpus, therefore, the understanding process is performed always in this language. This means that the SLU module should be fed by a translation module that translates the input utterances, which implies that the SLU models are not needed to be in more than one language. Thus, it is very important here to develop a good-performance task-specific MT system, as mistakes during the translation process can produce many errors in the SLU output.

In this work, we present how to obtain two parallel corpora French-Spanish and English-Spanish, from an original corpus in Spanish. We also present a methodology to translate not only the sentences, but also the semantic labels and the segmentation of the corpus. Finally, we present some experimental results that show the behavior of the techniques of translating the corpus in both cases.

2 The multilingual DIHANA corpus.

The multilingual DIHANA corpus that we have developed is based on the original Spanish DIHANA corpus. This 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. 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 considering a round trip.

The corpus has a total of 10.8 hours of speech uttered by 225 different speakers.

In order to use this corpus for SLU tasks, a semantic labeling was performed. 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 semantic label per segment). 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 Table 1.

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

The corpus was split into a training set of 4,887 turns and a test set of 1,340 turns.

3 Translating the original corpus.

We have translated this corpus into English and French following different procedures for the training and the test sets. For the training set, the process was based on the combination of the output from several web open-domain translators. This decision was made since for SLU purposes some errors in the translations can dramatically spoil the behavior of the system. For example, mistranslating any keyword strongly related to the semantic meaning, or even any polysemic word that could be translated erroneously by the translator using any other of its meanings, can severely damage the whole meaning of the sentence. If several hypotheses are generated by different translators there are more possibilities that the correct translation appear in one of the translated sentences.

Furthermore, as our aim is to work with limited-domain tasks, it would be desirable to have a task-dependent translation system that makes the minimum semantically important mistakes. Unfortunately, we are not able to train a task-dependent machine translation system, as the original corpus is just monolingual. Therefore, our option was to use web open-domain translators. However, as open-domain translators usually make many errors, our proposal is to obtain several outputs from several translators, and provide them all in our new multilingual corpus. Then, statistical SLU models using this variability can be trained, and a proper combination of these translations may improve the quality of the individual translations [1]. Hence, an improvement of the coverage and the overall quality of the multilingual SLU system could be achieved through this combination.

The training corpus translated can be used in different ways, being two of them the following ones:

- To learn a new model in the target language, from the translated training corpus and their segmentation/semantic-labeling associated.
- To learn an in-domain machine translation system from the pairs of sentences generated by the web open-domain translators.

To obtain these multilingual training corpus we have developed three kinds of corpus translations, all of them based on several open-domain web translators (4 translators were used for French, and another set of 4 web translators for English) as it is shown in Figure 1.

- First corpus translation: In this case the sentences are translated to the other language using the different translators. This way we have several hypotheses for each sentence. For example:

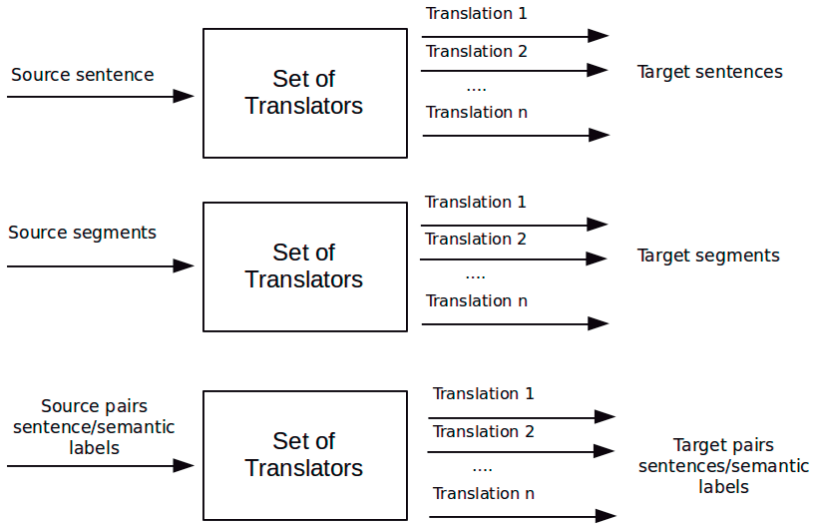

Fig. 1. Translations with open-domain web translators.

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 voult 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`

– Second corpus translation: Also each semantically labeled sequence of words was translated on its own. This way, we also obtained the translations of each semantically relevant piece (the lexicalization of each concept), as well as the segmentation and labeling of the corpus in the new language, without any manual effort. All this information makes it possible to learn a statistical semantic model for the task in the new language. For example: `por la tarde, à l'après-midi, in the afternoon ... o cuánto cuesta, quel est le prix, how much does it cost, ....`
– Third corpus translation: Considering that the translation of short segments can generate more errors than translations of complete sentences due to the

fact that they have not information about context, we have performed a translation of the complete sentences adapting the segmentation and semantic labeling to the translated sentences.
The adaptation of the semantic segmentation and labelling is performed in the following way: as we know the sequence of semantic units we can build a sentence as the concatenation of segments of words corresponding to the semantic units. Then, we obtain the best alignment between this artificial sentence and the sentence generated by the translator. This alignment associates a segmentation to the translated sentence. In this approach we assumed that the sequence of semantic units is the same in both languages.[1]

Quisiera horarios (Departure-Time)	**para el próximo sábado** (Date)	**a Barcelona** (Destination-City)	**por la mañana** (Hour)
Serait planifier (Departure-Time)	**pour le samedi** (Date)	**matin** (Hour)	**à Barcelona** (Destination-City)
I would like schedule (Departure-Time)	**for Saturday** (Date)	**morning** (Hour)	**to Barcelona** (Destination-City)

Regarding the test set, it was manually translated into French and English and uttered by native speakers. This way, we simulate a real scenario in which the native speaker interacts with the system using their own language. The test set in French is composed by 1,277 user turns, 500 of which were uttered by 4 native speakers. The test set in English, obtained in the same way, consists of 1,336 turns, and all of them were uttered by a total of 6 native speakers.

4 Experimental results for SLU

In order to study the quality of this approach, we have performed some preliminary experiments with a statistical SLU model using a test-on-source approach [2]. These experiments were performed considering as test set the French utterances. We trained a Statistical Machine Translator (MOSES) using the parallel French - Spanish training part of this new multilingual corpus (Figure 2). Then, the n-best translations for the recognized utterances were obtained by using this translator. Weighted graphs of words were built from these hypotheses, and a specific SLU decoding method for these structures was developed. The result in terms of Concept Error Rate was 22.40%, which is not too far from the result for the same test set considering the original Spanish utterances (17.72%).

We also studied the behavior of the systems when translating the training corpus to learn models in the new language, using a train-on-target approach. This study was done before the acquisition of this multilingual DIHANA corpus, and it was evaluated by using the French MEDIA corpus. In this previous work [4], we also explored the possibility of using both Conditional Random Fields (CRFs) and Stochastic Finite State Automata (SFSA) for the semantic modelization using a translated training corpus. The Concept Error Rate

[1] If this was not true, the forced alignment could be done between the sentence and an ergodic model of concatenation of the segments.

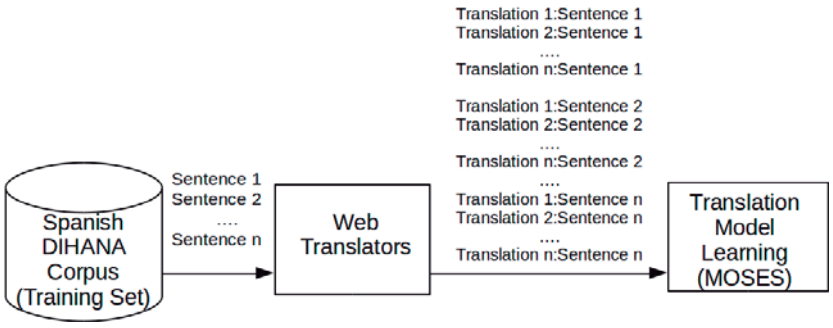

Fig. 2. The creation of a parallel corpus to estimate the MT system.

obtained from the experiments was 23,4% for CRFs and 25,2% for the SFSA. Results showed that it is possible to obtain accurate models by using this kind of translation process.

5 Conclusions

We have presented an approach to translate semantically labeled corpora to other languages, in order to build a SLU system in different languages. It includes a method to translate not only the sentences, but also the semantic labels and the segmentation of the corpus. We have also presented an approach to obtain a parallel corpus to estimate a task-specific machine translation system. Some baseline experiments are presented in order to show the capability of this approach to obtain accurate multilingual SLU systems, in both cases, in a test-on-source approach and in a train-on-target approach.

Acknowledgements. This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, and under FPU Grant AP2010-4193.

References

1. Bangalore, S., Bordel, G., Riccardi, G.: Computing Consensus Translation from Multiple Machine Translation Systems. In: In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001. pp. 351–354 (2001)
2. Calvo, M., Garcıa, F., Hurtado, L.F., Jiménez, S., Sanchis, E.: Exploiting multiple hypotheses for multilingual spoken language understanding. CoNLL-2013 pp. 193–201 (2013)
3. De Mori, R., Bechet, F., Hakkani-Tür, D., McTear, M., Riccardi, G., Tür, G.: Spoken language understanding: A survey. IEEE Signal Processing magazine 25(3), 50–58 (2008)

4. 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 2012). pp. 282–289. Miami (2012)
5. He, Y., Young, S.: Spoken language understanding using the hidden vector state model. Speech Communication 48, 262–275 (2006)
6. 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)
7. Maynard, H.B., Lefèvre, F.: Investigating Stochastic Speech Understanding. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2001)
8. 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)
9. Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L.: Extracting Semantic Information Through Automatic Learning Techniques. IJPRAI 16(3), 301–307 (2002)
10. Tür, G., Hakkani-Tür, D., Schapire, R.E.: Combining active and semi-supervised learning for spoken language understanding. In: Speech Communication. vol. 45, pp. 171–186 (2005)