ON THE USE OF MACHINE TRANSLATION FOR SPOKEN LANGUAGE UNDERSTANDING PORTABILITY

Christophe Servan¹, Nathalie Camelin¹, Christian Raymond², Frédéric Béchet³, Renato De Mori¹

¹ Université d’Avignon et des Pays de Vaucluse, LIA-CERI, BP 1228, 84911 Avignon Cedex 9 - France
² Université Européenne de Bretagne, IRISA-INSa, UMR 6074, Rennes, France
³ Aix Marseille Universite, LIF-CNRS, Marseille, France

{christophe.servan, nathalie.camelin, renato.demori}@univ-avignon.fr
christian.raymond@irisa.fr
frederic.bechet@lif.univ-mrs.fr

ABSTRACT

Across language portability of a spoken language understanding system (SLU) deals with the possibility of reusing with moderate effort in a new language knowledge and data acquired for another language.

The approach proposed in this paper is motivated by the availability of the fairly large MEDIA corpus carefully transcribed in French and semantically annotated in terms of constituents. A method is proposed for manually translating a portion of the training set for training an automatic machine translation (MT) system to be used for translating the remaining data. As the source language is annotated in terms of concept tags, a solution is presented for automatically transferring these tags to the translated corpus. Experimental results are presented on the accuracy of the translation expressed with the BLEU score as function of the size of the training corpus. It is shown that the process leads to comparable concept error rates in the two languages making the proposed approach suitable for SLU portability across languages.

Index Terms— Spoken Language Understanding, Portability across languages, Dialog Systems

1. INTRODUCTION

Portability of spoken language applications from one language to another is an interesting research problem of great practical importance as documented, for example, by recent papers [1, 2].

Across language portability of a spoken language understanding system (SLU) deals with the possibility of reusing with moderate effort in a new language knowledge and data acquired for another language. Let us define as destination and source, respectively the first and the second language.

Automatic language translation can be applied at different levels. For example in [3] spoken language interpretation is based on word patterns obtained by automatically training semantic classification trees from annotated data. These patterns are manually translated into a new language and used for interpretation of sentences in the new language. Another possibility is the automatic translation of knowledge resources like stochastic grammars from a language to another and then use the translated grammar to perform recognition and interpretation in the destination language [2].

The choice of an approach may depend on theoretical considerations, but also on practical aspects such as the amount, type and accuracy of available manually annotated or validated material.

The approach proposed in this paper is motivated by the availability of a fairly large MEDIA corpus carefully transcribed in French and semantically annotated in terms of constituents [4]. The semantic annotation associates concept tags describing constituents of semantic structures to manually identified word patterns called supports. It appears from the annotations made by linguists that the supports can be single words but also more complex syntactic structures often expressing knowledge chunks like references to a reference in a set.

The availability of automatic translation systems makes it possible to translate an entire corpus into a new language with little human effort. Assuming that the semantic of a domain is independent of the language, it is interesting to investigate if given the supports for concepts in a source language, it is possible to automatically characterize them in the destination language. Notice that this does not imply any grammar translation, but simply the translation of each support in the training set of the source language into a corresponding support in the message translated into the destination language. The difficulty of this operation may depend on the specific languages involved. This paper shows that it is effectively feasible if the source language is French and the target language is Italian. The quality of the semantic annotation obtained in this way depends, among other factors, on the quality of the translation that depends on the amount of manually transcribed data used for training the automatic machine translation (MT) component. The BLEU measure [5] is used for measuring translation quality. The concept error rate (CER) measured with the test set in the destination language should decrease with the increase of the translation quality which increases with the size of the corpus used for training the MT. Detailed experimental results show that this is effectively the case.

In practice, the transcriptions of the entire MEDIA corpus were automatically translated with MOSES [6] from French into Italian. In each message, the support for each concept, manually set in French, was automatically obtained in the Italian translation. The Italian annotated corpus was then used for automatically training the interpretation knowledge source based on Conditional Random Fields (CRF) with exactly the same process and tools as for French (see [7] for details).
This paper is structured as follows: section 2 presents the strategy used for translating the French MEDIA corpus into Italian; in section 3 we describe the Spoken Language Understanding model developed through the LUNA project on the MEDIA corpus and how we adapt it to process the Italian version of MEDIA; finally we present in section 4 an evaluation of our method both on the translation and SLU sides.

2. TRANSLATING THE MEDIA CORPUS

Statistical machine translation is an important paradigm in machine translation research. In this study we use the Moses toolkit [6] to perform automatic translation from French to Italian. Moses can translate ambiguous sentences by using a confusion network decoding and needs a parallel corpus with the source and the target language in order to train the translation models.

All the experiments presented in this paper have been made on the French MEDIA corpus [4]. This corpus contains recorded dialogs between a simulated system (Wizard of Oz protocol) and users. The target domain is hotel booking and tourism inquiries. The corpus is made of 1257 dialogs from 250 speakers and contains about 70 hours of speech. In this study we used about 1000 dialogs as the training corpus, representing about 14K dialog turns, and 208 dialogs as a test corpus, containing 3.5K dialog turns.

In order to obtain a parallel corpus French/Italian on MEDIA we used the following method:

1. We split the MEDIA train corpus into different sections.
2. The first section is used as a bootstrap corpus; we manually translate it from French to Italian and a first translation model is trained on it.
3. The following section is automatically translated thanks to Moses with the translation model obtained. The corpus translated is manually checked, corrected and added to the cluster of sections already translated.
4. A new translation model is trained on all the sections manually translated and checked.
5. We iterate the process in step 3 until half of the whole MEDIA training corpus has been translated.

With this method we obtained an Italian MEDIA training corpus containing half the dialogs of the French one.

The influence of the amount of data used for training the MT component is investigated by considering sections of the training corpus of increasing size as defined in Table 1. The n-th section contains all the data of the sections with lower index.

<table>
<thead>
<tr>
<th>Section Number</th>
<th># of words</th>
<th># of turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5132</td>
<td>828</td>
</tr>
<tr>
<td>2</td>
<td>10484</td>
<td>1656</td>
</tr>
<tr>
<td>3</td>
<td>16126</td>
<td>2484</td>
</tr>
<tr>
<td>4</td>
<td>21404</td>
<td>3312</td>
</tr>
<tr>
<td>5</td>
<td>26986</td>
<td>4140</td>
</tr>
<tr>
<td>6</td>
<td>32529</td>
<td>4968</td>
</tr>
<tr>
<td>7</td>
<td>37991</td>
<td>5796</td>
</tr>
<tr>
<td>8</td>
<td>43383</td>
<td>6628</td>
</tr>
</tbody>
</table>

Table 1. Size of each section used to measure the translation learning curve

For each section we train a Moses system and apply it to the French MEDIA test corpus. This automatically translated Italian corpus is then compared to the manual translation of the same corpus by Italian students from our University. Two measures have been computed: Language Model perplexity and BLEU measure. The perplexity value for the different sections of the training corpus is shown in figure 1. These measures were obtained on the test set by using two Italian Language Models. The parameters of the first model were estimated with the manually translated corpus used for training the MT component. The second one is a mixture of the first LM and the LM estimated with the results of automatic translation on the whole remaining French MEDIA training corpus. Therefore the first LM (referred as manual translation (sections)) however these translations are all manual ones. On the contrary the second model is trained on the whole MEDIA training corpus, but automatically translated (the portion of manually translated corpus increases for each section). As we can see in figure 1 the automatically translated corpus can improve the quality of the Language Model, even if the automatically translated corpus contains errors. For the sake of comparison we have also translated the whole MEDIA corpus with a free open-domain WEB translation tool. The perplexity obtained with a Language Model trained on such a corpus was 75.5 on the MEDIA test corpus. This confirms the need for specialized translation models when dealing with specific language registers and semantic domains.

In order to test the translation quality, we use the BLEU measure proposed in [5].

The results are shown in figure 2 and compare favourably with results reported at the 4th workshop on Statistical Machine Translation [8] in 2009.

3. SPOKEN LANGUAGE UNDERSTANDING PORTABILITY

The SLU system used in this study was developed on the MEDIA corpus in the framework of the LUNA1 European project. The inter-

http://www.ist-luna.eu/
The BLEU measure

Fig. 2. BLEU measured on the test MEDIA corpus with conditions for training the Moses toolkit of manual translation of each section only.

The interpretation process is based on Conditional Random Fields (CRF) as described in [7]. The first level in this SLU process is to extract from an utterance a list of concept tag hypotheses. These concepts correspond to the elementary bricks on which a structured interpretation of an utterance will be obtained. For example Named Entities, application specific keywords and expressions, dialog control expressions and some verbs are used as concepts in our MEDIA SLU system. The structured interpretations are represented by semantic frames as presented in [9] and are obtained by a semantic composition process on the elementary concepts. This second level is mostly language independent as the composition process is based only on the concept tags. Therefore the SLU portability is essentially done at the concept tag level.

Generation of concept tag hypotheses has been described in [7] and is briefly summarized in the following.

Let \( V \) be a vocabulary of words that can be hypothesized by an ASR system.

Let \( C^K_1 = C_1, \ldots, C_K \) with \( C_i \in V_C \) be a sequence of concept tags that can be hypothesized from a sequence of word hypotheses \( W^N_1 \), with \( W_i \in V \) in an initial interpretation step.

To each concept \( C_i \) is associated a sequence of words in \( W^N_1 \) of sequence of words being the support \( S_j \) of \( C_i \) in \( W^N_1 \).

A label is associated to all words \( W_j \) of a support \( S_j \); this label corresponds to the concept tag \( C_i \) and the position of \( W_j \) in \( S_j \).

We use the B (beginning) I (inside) and O (outside) (BIO) model as proposed by [10]. For example, the sentence:

\[
\text{une chambre pour deux personnes dans au Novotel}^3
\]

contains the following concepts:

\[
\begin{align*}
C_1 &= \text{amount} \quad S_1 = \{\text{une}\} \\
C_2 &= \text{room} \quad S_2 = \{\text{chambre pour deux personnes}\} \\
C_3 &= \text{hotel\_brand} \quad S_3 = \{\text{Novotel}\}
\end{align*}
\]

With the BIO model, we represent the sentence by a sequence of pairs \((\text{word}_i, \text{label})\):

\[
(\text{une,amount}\_B) (\text{chambre,room\_type}\_B) (\text{pour,room\_type}\_I) (\text{deux,room\_type}\_I) (\text{personnes,room\_type}\_I) (\text{dans,}\text{NULL}\_O) (\text{au,}\text{NULL}\_O) (\text{Novotel,hotel\_brand}\_B)
\]

The goal of the CRF tagger used in our SLU module is to predict the correct sequence of labels \( L^N_1 \) for a sequence of words \( W^N_1 \).

To train the tagger we translate the MEDIA training corpus into the BIO format and use as features n-grams of words and labels with the CRF++ toolkit\(^4\).

This corpus-based tagging approach only needs a training corpus with concept tags in order to learn a new model, no prior knowledge is required. Therefore adapting the concept tagger to a new language consists in translating the training corpus into the new target language. Such a translation can be made by a statistical machine translation approach, as presented in section 2. In order to obtain semantic annotations in the target language at the same time as the word translation, we use one of the features of the Moses toolkit which allows adding XML tags in the text to translate. These XML tags are projected into the target language without disturbing the translation process.

The portability process of adapting the French MEDIA concept tagger to a new language can be described as follows:

- The French MEDIA training corpus is formatted with XML tags representing the concept tags. On the previous example we obtain:

\[
\begin{align*}
\text{<tag c="amount_room"> une </tag> <tag c="room_type"> chambre pour deux personnes </tag> dans au Novotel </tag> \\
\text{<tag c="hotel_brand"> Novotel </tag> \\
\text{camera per due persone </tag> nel al Novotel </tag>
\end{align*}
\]

- This XML corpus is translated with the Moses models as presented in section 2. We obtain:

\[
\begin{align*}
\text{<tag c="amount_room"> una </tag> <tag c="room_type"> camera per due persone </tag> nel al Novotel </tag>
\end{align*}
\]

- The Italian XML corpus is translated into the BIO format:

\[
\begin{align*}
\text{(una,amount_room) (camera,room_type) (per,room_type) (due,room_type) (personne,room_type) (nel,\text{NULL}) (al,\text{NULL}) (Novotel,hotel\_brand) ,}
\end{align*}
\]

- Finally the CRF tagger is trained on this Italian MEDIA training corpus.

4. EXPERIMENTS

The annotated French test corpus was manually translated into Italian and used for test. It is made of 208 dialogs, 3.5K dialog turns, and contains 8.4K concept occurrences. We evaluate the performance of our SLU language portability process thanks to the Concept Error Rate (CER) measure. CER is defined as the ratio of the sum of deletion, insertion and substitution errors over the total number of concepts.

Two conditions are evaluated:

1. **manual translation**: the CRF tagger is trained only on the manual translation of the French MEDIA training corpus;
2. **manual translation + automatic translation**: a Moses translation model is trained on the manual translations and is applied to the remaining non-translated MEDIA training corpus. The CRF tagger is trained on the concatenation of both manual and automatic translation.

\(^4\)Toolkit CRF++: http://crfpp.sourceforge.net/

\(^5\)The XML tags are automatically projected from the French source corpus to the target Italian one. This is rather straightforward between French and Italian since there is no major word reordering between these two languages.
The CER measure

Figure 3 presents the results of these two conditions for different sizes of the manually translated MEDIA training corpus, corresponding to the sections presented in 2.

As expected for the first condition, the CER decreases when new training data are added to the training corpus. The second condition clearly outperforms the first one. This can be explained by the fact that the training corpus of the CRF tagger is much bigger in the second condition however the quantity of human effort (manual translation) is identical at each point of the two curves. All the remaining training corpus added is done in a fully unsupervised way. This result validates our approach using statistical machine translation in order to automatically augment a bootstrap training corpus manually translated.

It is also very interesting to notice that the lowest CER value of the second condition is nearly reached at the second point in the curve, corresponding to the second section of table 1. At this point only 828 dialogs are manually translated, although the CER obtained with the manual translation + automatic translation condition is lower than the one obtained with more than 6600 manually translated dialogs for the manual translation condition.

5. CONCLUSION

Automatic MT was used for porting an SLU system from French to Italian. Good results were obtained by manually translating a selected portion of the French training corpus for training the MT system with which all the remaining data were translated.

A method is proposed for obtaining in the translated corpora semantic annotations associated to words starting from the word patterns expressing concepts in the source language.

Directions for further research are on the transformation of idiomatic sentences between languages, on the propagation of relations between word patterns expressing annotated concepts in the source language to the destination language and to strategies for progressive manual translation of sentences to be used to better training the MT system.

6. REFERENCES


