Portability of Semantic Annotations for Fast Development of Dialogue Corpora

Bassam Jabaian\textsuperscript{1,2}, Fabrice Lefevre\textsuperscript{1}, Laurent Besacier\textsuperscript{2}

\textsuperscript{1}LIA, University of Avignon, Avignon, France
\textsuperscript{2}LIG, University of Joseph Fourrier, Grenoble, France

\{bassam.jabaian,fabrice.lefevre\}@univ-avignon.fr, laurent.besacier@imag.fr

Abstract

Generalization of spoken dialogue systems increases the need for fast development of spoken language understanding modules for semantic tagging of speaker’s turns. Statistical methods are performing well for this task but require large corpora to be trained. Collecting such corpora is expensive in time and human expertise. In this paper we propose a semi-automatic annotation process for fast production of dialogue corpora by automatically pre-annotating the corpus before performing manual corrections. For the pre-annotation we propose to port a system initiated from an existing corpus and to adapt it to the new data. The French MEDIA dialogue corpus (hotel reservation) is used as a starting point to produce two new corpora: one for a new language (Italian) and another for a new domain (theatre ticket reservation). We show that the automatic pre-annotation leads to a significant gain in productivity compared to a fully manual annotation and thus allows to derive adaptation data which can be used in turn to further improve the systems.

\textbf{Index Terms:} Spoken Dialogue Systems, Spoken Language Understanding, Language Portability, Statistical Machine Translation

1. Introduction

Human-machine dialogue systems can be designed for several application domains and for different languages. The need for a fast production of such systems for new languages and new domains is still an open problem that needs to be addressed.

Dialogue system components are usually dedicated to a specific task in a specific language. With the growing need for these systems arises the need for rapid development or adaptation of these modules. Some of them like the dialogue act manager (DAM) are mostly language independent, but others like automatic speech recognition (ASR) and spoken language understanding (SLU) depend on the target language and on the application domain at stake. In this work we focus on the fast development of corpora for SLU development in a new language and for a new domain.

The SLU plays an intermediary role between the ASR and the DAM, it is due to extract the semantic information from the (noisy) transcription of the user’s utterance. Several statistical approaches have been proposed recently to train an SLU system (e.g., [15, 8, 3]). Those approaches shift the need of human expertise compared to expert-based approaches since only a semantically annotated corpus is required. However the size of this corpus has a large influence on the system’s quality. Collecting such a corpus is costly in time and human expertise.

As shown by [2] most of the time dedicated to create a dialogue system goes to data collection and annotation. Several works dealt with the fast production of such corpora. In [14] an active learning procedure was proposed to reduce corpus annotation and verification time, while [12] proposed to build a mini corpus to bootstrap a first system further used in an iterative data collection process. These approaches however suppose a deployed system allowing a continuous collection of data from which they can extract an optimal subset to train the SLU. In this paper a semi-automatic annotation is proposed in order to reduce the corpus annotation time based on a fixed dataset.

An automatic pre-annotation is realized by an SLU tagger which needs an annotated corpus to be trained. For the new language tagger we suggest to automatically translate the existing corpus and to port its annotation to the target side. While for the new domain corpus, we propose to use the original (then out-of-domain) corpus to train the SLU tagger. In both cases the pre-annotation is then corrected by human annotators. In the rest of this paper the new language and domain corpus will be referred to as PM-LANG and PM-DOM respectively.

This work\textsuperscript{1} was motivated by the availability of the French MEDIA corpus (hotel reservations and tourist information). We used this corpus to bootstrap an Italian corpus for the same task and a French corpus for the ticket reservation domain. Taggers used for the pre-annotation were trained using conditional random fields (CRF). Productivity gains of the pre-annotation were measured and the quality of the new corpora was evaluated.

This paper is organized as follows: in Section 2 we describe the French MEDIA corpus. Section 3 describes

---

\textsuperscript{1}This work is supported by the ANR funded Port\textsuperscript{3}MEDIA project (ANR 08 CORD 026 01). More information on the project website, www.port-media.org
the new corpora, including details on their collection and their pre-annotation, while Section 4 is dedicated to an experimental evaluation of these corpora for an SLU task.

2. The MEDIA Corpus Description

As described in [1] the MEDIA corpus covers a domain related to the reservation of hotel rooms and touristic information. The corpus is made up of 1257 dialogues from 250 speakers. A subset of the corpus was manually translated into Italian during the LUNA project [13].

This corpus is tagged with 99 concepts giving a precise semantic representation. This semantic annotation basically segments each sentence into various chunks. Each chunk is annotated with the concept name, the value of the concept and its mode.

3. Fast Corpus Annotation

The PM-LANG (Italian) corpus has been recorded and transcribed with the same specifications and configurations as the MEDIA corpus: the same tools, protocols, scenarios and constraints were used to collect the dialogues. So the only difference between the two corpora is the spoken language. Hence the adaptation consisted in translating prompts and scenarios from French into Italian, but no changes were made to the content of the scenarios.

Contrary to MEDIA, the PM-DOM corpus is dedicated to the ticket reservation for the Avignon Festival. We tried to keep up the characteristics and paradigms of MEDIA. Hence the adaptation work meant to create new scenarios for users and to adapt the dialogue manager for the agent.

To reduce the annotation time for the new data, we proposed to automatically pre-annotate the transcribed dialogues and then to manually correct this annotation. The pre-annotation has been carried out by an SLU system trained using CRFs models. CRFs are briefly described in Section 3.1. Those models need a semantic segmentally annotated corpus to be trained. Since we do not have such corpus initially, to train a model in the target language, we suggest to translate the existing MEDIA corpus into the new language and then to infer the annotation of MEDIA into the translated corpus. As far as a new domain is concerned, an adaptation to the task is still needed to perform an efficient pre-annotation. Section 3.2 describes the language portability and the domain adaptation as well as the annotation process.

3.1. CRF-based semantic taggers

To train the SLU system we use a statistical method performing well for sequential tagging: conditional random fields (CRFs) [6]. CRFs represent a log-linear model, normalized at the sentence level. The CRFs model the probability between concepts \( c_1^N \) and words \( w_1^N \) as follows:

\[
P(c_1^N | w_1^N) = \frac{1}{Z} \prod_{n=1}^{N} H(c_{n-1}, c_n, w_{n+2}^n)
\]

with

\[
H(c_{n-1}, c_n, w_{n+2}^n) = \sum_{m=1}^{M} \lambda_m \cdot h_m(c_{n-1}, c_n, w_{n+2}^n)
\]

Log-linear models are based on feature functions \( h_m \) representing the information extracted from the training corpus, \( \lambda \) are estimated during the training process, \( Z \) is a normalization term.

3.2. Annotation Process

Concerning the French PM-DOM corpus, the model used for the pre-annotation was trained directly on MEDIA data (of the same language) and a named entity tagger was added to take into account the new domain specifications. The MEDIA based tagger covers concepts which are shared between the MEDIA and the PM-DOM corpora but cannot cover the domain novelties. The concepts dealing with command, location and payment are generated by the MEDIA tagger, while other new domain concepts (such as artist names, play titles etc.) are still unknown for this tagger. To deal with that, a simple tagger based on regular expressions is applied. This tagger is obtained using a list of named entities consisting mainly of name of authors, play titles and locations extracted from the program of the Avignon Festival. The pre-annotation is the combination between SLU trained on MEDIA and the output of the named entity tagger.

For the PM-LANG corpus, since no training corpus in the target language is available a priori, we proposed to port automatically the French MEDIA corpus into Italian (translation and annotation). Recently, several works (eg [4] or [9]) investigated and compared methods for language portability of an SLU system. For the pre-annotation of the new language corpus we selected the best performing method adapted to our task.

The applied method proposes to translate the source training corpus into the target language and then to infer the corresponding semantic annotations. We use the manually translated subset of the MEDIA corpus as a parallel corpus to train a French into Italian statistical machine translation system (SMT) in order to translate the remaining part of the MEDIA corpus. This system —with a BLEU score of 43.62— is trained using the MOSES toolkit [5]. MOSES provides a state-of-the-art phrase-based translation (PB-SMT) system based on log-linear models.

The annotation transfer was based on the automatic alignment between French and Italian sentences. More precisely, this method consists of concept projection using word to word alignment between parallel source-target corpora. Then since the source corpus is already

INTER SPEECH 2012 215
annotated at the chunk level, we proposed to use the word to word alignment information between the two corpora to match directly the semantic concepts with the chunks in the target corpus. The automatic word alignment of these corpora is generated with the help of the Berkeley aligner toolkit, described in [11]. So we get an annotated corpus in the target language on which we train the SLU tagger to pre-annotate the PM-LANG corpus.

For both corpora, the annotation was performed through three iterations. During the data transcription, a first subset is pre-annotated automatically and corrected manually. The correction is then added to the training data to improve the model used for pre-annotating the following subset. The time required to correct the automatic pre-annotation was measured and compared to the one needed for annotating from scratch. For this, the following protocol was implemented. For each iteration a set of 10 dialogues was annotated twice by two different annotators. The first annotator corrected the pre-annotation while the second one annotated the same 10 dialogues from scratch. The productivity gains are shown in Table 1, significant productivity gains have been achieved for both corpora over the three iterations. The pre-annotation in the third iteration leads to a productivity gain of more than 50% for the PM-LANG corpus and 40% for the PM-DOM corpus (more details are given in [10] for the pre-annotation process applied).

Table 1: Productivity gains over 3 iterations for PM-LANG and PM-DOM semantic annotation.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM-LANG</td>
<td>49.02%</td>
<td>35.65%</td>
<td>53%</td>
</tr>
<tr>
<td>PM-DOM</td>
<td>17.14%</td>
<td>34.52%</td>
<td>43.82%</td>
</tr>
</tbody>
</table>

4. Experiments and Results

The results of this semi-automatic annotation is the PM-LANG corpus of 604 dialogues and the PM-DOM corpus of 700 dialogues. A description of these corpora is given in Table 2. Once corpora were collected and annotated we tried to use them for SLU models improvement. Each corpus is divided into two parts: one for training and another one for testing (200 dialogues to be comparable to MEDIA). The Wapiti toolkit [7] is used to train the CRF models. CER is the evaluation criterion chosen for this study, and is simply defined as the ratio of the sum of deleted, inserted and substituted concepts by the total number of concepts in the reference.

4.1. PM-LANG Corpus

To evaluate the PM-LANG corpus we trained an SLU model on the training data set, and we used two test sets to evaluate the model. The first one is the test manually translated from the MEDIA corpus, and the other one is the PM-LANG test set. We will refer to the model trained on PM-LANG corpus as PM-LANG model.

We also evaluate the SLU model used for the pre-annotation on both test sets. This model is called M-LANG in text and tables. At the end the two corpora are merged to train a new combined model. This model is also evaluated on the two test sets. The results of these evaluations are reported in Table 3.

Results on M-LANG model show the robustness of the model. The performance of this module on the two test sets is very similar (20.5% vs. 20.8%). This model performs well for a new test set built from scratch as well as for a test set translated from another corpus.

The M-LANG model gives a CER of 18.9% on its matching test set, while its performance is lower for the PM-LANG test set. We will refer to the model trained on PM-LANG corpus as PM-LANG model.

The combination of the two corpora gives a model performing better than each one separately (19.5% on the M-LANG corpus and 17.6% on the PM-LANG corpus).

4.2. PM-DOM Corpus

We trained an SLU model on the PM-DOM training set in order to evaluate the corpus. In parallel we evaluated the MEDIA baseline corpus (hotel reservation) on the PM-DOM test set (ticket reservation).

As shown in Table 4, the PM-DOM model gives a CER of 19.1% which is relatively good regarding the size of the training data. The MEDIA model which was used for the pre-annotation gives a CER of 41.2% on the PM-DOM corpus. This low score can be explained by a considerable difference in concept tags between the MEDIA and PM-DOM corpora.

Table 2: MEDIA, PM-LANG and PM-DOM description.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Domain (Lang)</th>
<th>Dialogues (Sent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDIA</td>
<td>Hotel Booking (FR)</td>
<td>1259 (18.8K)</td>
</tr>
<tr>
<td>PM-LANG</td>
<td>Hotel Booking (IT)</td>
<td>600 (11K)</td>
</tr>
<tr>
<td>PM-DOM</td>
<td>Ticket Booking (FR)</td>
<td>700 (8.3K)</td>
</tr>
</tbody>
</table>

Table 3: Evaluation (CER %) of Italian models on two test sets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-LANG</td>
<td>3.1</td>
<td>15.0</td>
<td>2.3</td>
<td>20.5</td>
<td></td>
</tr>
<tr>
<td>PM-LANG</td>
<td>3.8</td>
<td>13.9</td>
<td>3.1</td>
<td>20.8</td>
<td></td>
</tr>
<tr>
<td>combination</td>
<td>4.7</td>
<td>17.4</td>
<td>3.2</td>
<td>25.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.6</td>
<td>12.1</td>
<td>3.3</td>
<td>18.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.8</td>
<td>14.6</td>
<td>2.1</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.9</td>
<td>9.0</td>
<td>4.6</td>
<td>17.6</td>
<td></td>
</tr>
</tbody>
</table>
This difference in concepts between the two corpora has a significant influence on their direct combination. Unlike what was observed for the new language, the concatenation of the MEDIA and PM-DOM corpora for training a new model is not performing well. As shown in Table 4, the combined model gives 37.5% of CER which is much lower than the model trained on PM-DOM alone. This degradation comes from the fact that the combination gives an important amount of insertion in the output hypotheses. Those insertions come from MEDIA corpus concepts which do not exist in the PM-DOM corpus.

In order to have a better combination we propose to filter the MEDIA concepts, eliminating the MEDIA concepts not occurring in the PM-DOM corpus. For corpus filtering we considered two options. The first one consists in eliminating all chunks annotated with irrelevant concepts (not used in the PM-DOM concepts set) and we get a filtered corpus referred to as F-MEDIA (10K sentences vs 13K of MEDIA). The second one consists in eliminating the whole sentences that contain irrelevant concepts, referred to as F'-MEDIA (7K sentences). Each of these filtered corpora is then concatenated with the PM-DOM corpus and a model is trained on the combined corpus.

It is important to mention that the F-MEDIA corpus contained more sentences than F'-MEDIA but some of them are somewhat incorrect. The noise comes from the removal of unsuitable chunks appearing inside the sentence. There is a strong assumption behind this option, that removing a concept chunk in a sentence will generally lead to another acceptable utterance. But obviously the precise nature of the removed concepts has a great impact on the outcome and many sentences can become grammatically incorrect.

Results presented in Table 4 show that the model trained on the combination of PM-DOM with F'-MEDIA (18.9%) is performing better than the combination with F-MEDIA (19.5%). The combination is better with less data than with more noisy data. This combination slightly increases the performance of our method compared to using PM-DOM only (19.1% vs. 18.9%).

5. Conclusions

In this paper we presented the results of portability approaches to semantic annotations for the fast production of corpora for new language and domain for the understanding task. The required time for the new data annotation was reduced by an automatic pre-annotation. This pre-annotation was performed using models based on an existing corpus. The pre-annotation was manually corrected and correction time was compared to the time required for an annotation from scratch. Important productivity gains of roughly 50% are observed for both language and domain corpora.

We also evaluated the performance of the derived corpora and we obtained a CER of 18.9% for the new language and 19.1% for the new domain. Finally we proposed to combine the new corpora with the corpora used to pre-annotate them, and we show that this combination can further improve (significantly for the new language task) the overall performance of the models.

6. References


<table>
<thead>
<tr>
<th>Model</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM-DOM</td>
<td>3.2</td>
<td>13.0</td>
<td>2.9</td>
<td>19.1</td>
</tr>
<tr>
<td>MEDIA</td>
<td>11.2</td>
<td>25.6</td>
<td>4.4</td>
<td>41.2</td>
</tr>
<tr>
<td>PM-DOM + MEDIA</td>
<td>5.8</td>
<td>2.8</td>
<td>28.9</td>
<td>37.5</td>
</tr>
<tr>
<td>PM-DOM + F-MEDIA</td>
<td>3.0</td>
<td>13.2</td>
<td>3.3</td>
<td>19.5</td>
</tr>
<tr>
<td>PM-DOM + F'-MEDIA</td>
<td>3.4</td>
<td>11.8</td>
<td>3.9</td>
<td>18.9</td>
</tr>
</tbody>
</table>

Table 4: Evaluation (CER %) of French models on PM-DOM test.