Using inter-lingual triggers for machine translation

Caroline Lavecchia, Kamel Smaïli, David Langlois, Jean-Paul Haton
Loria BP 239 54506 Vandœuvre Lès-Nancy France
lavecchi@loria.fr, smaili@loria.fr, langlois@loria.fr, jean-paul.haton@loria.fr

Abstract
In this paper, we present the idea of cross-lingual triggers. We exploit this formalism in order to build up a bilingual dictionary for machine translation. We describe the idea of cross-lingual triggers, the way to exploit and to make good use of them in order to produce a bilingual dictionary. We then compare it to ELRA and a free downloaded dictionaries. Finally, our dictionary is evaluated by comparing it to the one achieved by GIZA++ [1] (which is an extension of the program GIZA [2]) into an entire translation decoding process supplied by Pharaoh [3]. The experiments showed that the obtained dictionary is well constructed and is suitable for machine translation. The experiments have been conducted on a parallel corpus of 19 million French words and of 17 million English words. Finally, the encouraging results allow us to put forward the concept of cross-lingual triggers which could have so many applications in machine translation.

1. Introduction
Statistical techniques have been used in several areas of natural language processing: speech recognition, OCR, information retrieval, machine translation, speech-to-speech translation, ... To make these applications working, huge corpora are necessary to learn several model’s parameters. Corpora are also used to build up automatically a dictionary for speech recognition, indexing or OCR. In this paper we investigate how to take advantage from parallel corpora to build up a bi-directional dictionary. For each word $e$ in a source language, we would like to find out the best $n$ words $f_1, f_2, \ldots, f_n$ which are considered as the most likely translations in a target language. Obviously the role of target and source language can be exchanged. A bilingual dictionary may be created by using linguistic knowledge (a human dictionary) [4] or automatically from parallel corpora by using techniques based on EM algorithm [5] as in GIZA++ tools [1]. Himestera [6] used a symmetric EM algorithm to compile a bi-directional dictionary and claimed that his algorithm leads to better estimates of the translation probabilities. Kuman and Hirakawa [7] use both linguistic and statistical information to generate a machine translation dictionary from parallel Japanese and English texts. Smadja et al [8] proposed the tool Champollion which translates a list of given collocations from parallel corpora by using Dice coefficient as a similarity measure.

In the rest of the paper, we give an overview of triggers in section 2. Section 3 presents the notion of cross-lingual triggers which associate to each word its related words in the source and the target language. A description of used corpora and results are provided and discussed in section 4. We end with a conclusion which points out the strength of our method and gives some tracks about future work in our research group.

2. A Brief Remind of Triggers
The concept of triggers has been largely used in statistical language modeling. Triggers improve and generalize the Cache model [9]. The Cache model enhances the probability of a word $w$ when it occurs in its left context. A trigger model goes further and enhances the probability of a list of words which are correlated to $w$ [10]. Triggers are determined by computing mutual information given by:

$$I(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$ (1)

For each dictionary entry the $n$ best correlated words in terms of mutual information are kept. We call a trigger a set made up of a trigger and its triggered words. In language modeling triggers are used as a new language model which is interpolated with a classical n-gram [11].

3. Cross Lingual triggers
Cross lingual triggers have been also used in [12] to enrich resource deficient languages from those which are considered as potentially important.
A cross lingual trigger is henceforth a set made up of a word $e$ in a source language, and its best correlated words in a target language $f_1, f_2, \ldots, f_n$. This will be written as: $T_{\text{cross}}(e) = f_1, f_2, \ldots, f_n$. The method we propose produces cross-language triggers (classical one) and inter-language triggers. That means Source-Source, Target-Target, Source-Target and Target-Source triggers are calculated. In order to find out these triggers, all the pairs of sentences have been concatenated inside the same corpus as in Fig. 1. The triggers in which we are interested are depicted. For a trigger word $e^k$, a partial mutual information (PMI) is calculated over each pair $k$ of sentences and then a global mutual information $MI_G$ is evaluated over all the corpus ($S$ pairs), namely:

$$PMI(e^k, f^k_j) = \log \frac{P(e^k, f^k_j)}{P(e^k)P(f^k_j)}$$ (2)

$$MI_G(e^k, f^k_j) = \frac{1}{S} \sum_{k=1}^{S} PMI(e^k, f^k_j)$$ (3)

(3) is used to retrieve inter-lingual triggers but it is employed also to generate intra-lingual triggers. The above formula looks like the one used in the literature but is not exactly the same. In fact, our objective is to lead to machine translation dictionary without using any external knowledge. That is why the mutual information is calculated inside a window which has the length of a concatenated pair of sentences (for which one is the translation of the other). Clearly, we would like to retrieve the words in a target language.
4. Dictionary production

The experiments presented below have been conducted on the proceedings of the European Parliament [13]. We used the French-English parallel corpus of 598014 sentence pairs. The French side has a total of 19 million words (78431 unique tokens). The English side has a total of 17 millions words (56243 unique tokens). We constructed a unique dictionary including English and French words. The vocabulary is built up from the union of the 26811 most frequent French words and of the English and French words. The vocabulary is built up from unique tokens). We constructed a unique dictionary including English and French words. The vocabulary is built up from the union of the 26811 most frequent French words and of the 19588 most frequent English words\(^1\). For each vocabulary word the source word \(s\) and the target word \(t\) are correlated to a word \(e\) in a source language. Among the set \(F\), we hope to find a subset \(T\) which is made up only by the translations of \(e\).

\( F = f_1, f_2, \ldots, f_n \) which are correlated to a word \( e \) in a source language. Among the set \( F \), we hope to find a subset \( T \) which is made up only by the translations of \( e \).

- **Disciplines**
  - disciplines (0.22)
  - régles (0.07)
  - investissements (0.06)
- **Competences**
  - compétences (1.22)
  - pouvoirs (0.1)
  - institutions (0.09)

Table 2: Examples of English words triggered by French words

<table>
<thead>
<tr>
<th>French trigger word</th>
<th>English trigger word</th>
<th>( M_{IG} \times 10^{-4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coopération</td>
<td>cooperation</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>développement</td>
<td>6</td>
</tr>
<tr>
<td>Coopératives</td>
<td>coopératives</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>femmes</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>associations</td>
<td>0.09</td>
</tr>
<tr>
<td>Difficulté</td>
<td>difficulté</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>difficultés</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>problème</td>
<td>0.3</td>
</tr>
<tr>
<td>Disciplines</td>
<td>disciplines</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>régles</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>investissements</td>
<td>0.06</td>
</tr>
<tr>
<td>Compétences</td>
<td>compétences</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>pouvoirs</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>institutions</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 3: A selection of few entries of English-French dictionary

<table>
<thead>
<tr>
<th>English word</th>
<th>Potential translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish</td>
<td>pêche, poisson, poissons</td>
</tr>
<tr>
<td>Fisherman</td>
<td>pêcheur, pêcheurs, pêche</td>
</tr>
<tr>
<td>Flag</td>
<td>pavillon, drapeau, navires</td>
</tr>
<tr>
<td>Flexible</td>
<td>flexible, souple, travail</td>
</tr>
<tr>
<td>Foods</td>
<td>alimentaires, alimentaire, produits</td>
</tr>
<tr>
<td>Henceforth</td>
<td>désormais, dorenavant, dès</td>
</tr>
</tbody>
</table>

\(^1\)French and English words occurring more than 7 times.
To sum up we can say that the results obtained are very interesting and the recall is probably better than 65%. We have to compare TrigDic to a better reference (a hand-constructed one) to have a precise evaluation.

6. Translation decoding with triggers

In order to evaluate the real contribution of our method, we have to integrate the retrieved dictionary into an entire decoding translation process supplied by Pharaoh4 [3]. To achieve that we assign to each potential word’s translation a probability calculated from \( MI \). In a first experiment, we use the TrigDic dictionary generated in section 4: each word of source and target language is associated with its 10 best inter-lingual triggered words, and each word gets 5 potential translations. Each of these 5 translations is given a probability dependent on \( MI \). The translation probability for other vocabulary words is set to 0. Translation results in terms of Bleu [14] for a subset of the source corpus are given in Table 7, column ‘TrigDic’. The performance is compared to the one obtained with a GIZA++ dictionary using the IBM Model 2 [15].

To improve these results, we investigate two hypotheses. First, the size of the triggers lists (10) and the number of potential translations (5) may be too restrictive. To study this explanation, we extended the dictionary: each word of source and target language is associated with its 50 best inter-lingual triggered words, and each word gets 10 potential translations. The new results are given in Table 7, column ‘extended TrigDic’. These results show a slight improvement.

The decoding based on our vocabulary is less powerful than the one obtained by GIZA++ and these results are robust across corpora with different sizes. In order to improve these results, we investigate two hypotheses. First, the size of the triggers lists (10) and the number of potential translations (5) may be too restrictive. To study this explanation, we extended the dictionary: each word of source and target language is associated with its 50 best inter-lingual triggered words, and each word gets 10 potential translations. The new results are given in Table 7, column ‘extended TrigDic’. These results show a slight improvement.

4The target language model is a trigram model (Good-Turing smoothing, cutoff set to 7 for bigrams and trigrams). The decoding weights are set to: 1 for language model, 1 for translation model, 0 for word penalty, and 1 for distortion model. Decoding is with reordering.
Second, in the two previous experiments, a null probability is assigned to all the words which are not in the top list of the potential translations. In fact, assigning a probability for only the n best translations gives no chance to any other word to be a candidate for translation. That is why probabilities have to be smoothed in an attempt to give more words a chance to be potential translations. In a first step, we propose to assign a not null translation probability to the empty word (a word can be translated to no word in the target sentence). The results are given in Table 7, column ‘with smoothing’ (we used here the extended TrigDic dictionary). This simple solution allows to lead to better performance. To improve the results, we have in the future to define a more efficient smoothing technique.

Moreover, a realistic translation should not be done word by word, that is why we guess that a phrase by phrase translation may achieve better performance. Even if Pharaoh segments the input into phrases, we think they have to be introduced up stream in order to estimate independently the phrase probabilities P(\varepsilon|\phi). In the few last years we developed statistical method to generate phrases [16][17]. In a next work, we will use this method to rewrite the source and target corpora in terms of phrases. Then, we will use cross-lingual triggers on phrase corpora to constitute a more relevant dictionary. This dictionary and the estimation associated to each potential translation should improve the decoding performance.

7. Conclusion and future work

We have presented a method for translating words based on the concept of cross lingual triggers. These triggers have been retrieved from parallel corpora of sentence pairs. The pairs have been concatenated, intra and inter-lingual triggers have been carried out. For each word (French or English) a list of its corresponding triggers has been proposed. An entry of a bilingual dictionary is made up of a source word and its best translations. The best translations are obtained by cross lingual triggers which themselves cross-trigger the source word. The obtained dictionary is relevant. It has been evaluated by comparing it to ELRA and an Internet dicionaries. The results are encouraging (65% in terms of recall) and they are probably higher if they are compared to a better dictionary (a human one). In the near future we will check our dictionary by a human.

First results in Pharaoh are less good than the one obtained by GIZA++. This is due to the fact that, the probabilities associated with the translations are not smoothed. Only the first best translations are considered as potential translations and then have a significant probability. We have to propose more translations and in all cases we have to smooth probabilities. In a near future, phrases will be introduced up stream to make the cross lingual triggers generating a bilingual phrases dictionary.

The idea of using cross lingual triggers seems to be very important, it can be used in several areas in machine translation. For instance, they could be used as a confident measure. Several other utilizations of this method have been imagined and are under-work in our research group.

8. Acknowledgments

This work is supported by EADS (European Aeronautic Defence and Space Company) foundation for the Speech-To-Speech translation Project.

9. References


