Using inter-lingual triggers for Machine translation

Caroline Lavecchia, Kamel Smaïli, David Langlois, Jean-Paul Haton

▶ **To cite this version:**

Caroline Lavecchia, Kamel Smaïli, David Langlois, Jean-Paul Haton. Using inter-lingual triggers for Machine translation. 8th Annual Conference of the International Speech Communication Association - INTERSPEECH 2007, Aug 2007, Antwerp, Belgium. pp.2829-2832. inria-00155791

HAL Id: inria-00155791
https://inria.hal.science/inria-00155791

Submitted on 19 Jun 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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]. Himestra [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. Kumano 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_i 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_i [10]. Triggers are determined by computing mutual information given by:

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

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: $Trig(e) \longrightarrow f_1, f_2, \ldots, f_n$. The method we propose produces intra-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_i^k and a triggered target word f_j^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:

$$PMI(e_i^k, f_j^k) = \log \frac{P(e_i^k, f_j^k)}{P(e_i^k)P(f_j^k)} \tag{2}$$

$$MI_G(e_i, f_j) = \frac{1}{S} \sum_{k=1}^{S} PMI(e_i^k, f_j^k) \tag{3}$$

(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

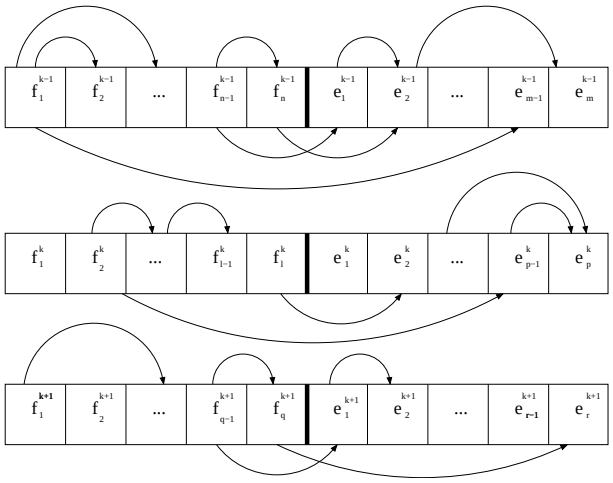

Figure 1: Cross Lingual Triggers

$F = f_1, f_2, \ldots, f_n$ 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.

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 19588 most frequent English words[1]. For each vocabulary word the 10 best triggers are kept. All the "tool" words (small words in English and French as: *or, it, in, thus ...., de, la, le, donc, ...* are put in a bag of words and are excluded from being considered as triggered words: their translations are included by hand into the final dictionary.
Our method leads to remarkable inter-lingual triggers where the triggered words could be considered as potential translations of the trigger or very close in terms of meaning. Table 1 illustrates some examples of the obtained English-French triggers, whereas Table 2 gives some French-English triggers.

The third column indicates the global mutual information associated to each couple (trigger and triggered words). Using inter-lingual triggers is in our opinion the first step for several other applications in machine translation. Our first goal is to provide automatically a bilingual dictionary in multiple languages. In this paper we focus only on French and English. For each dictionary and for each entry we produce p-word translations. The translations of a word e are obtained by selecting all the target triggered words $f_1, f_2, \ldots, f_n$ which trigger the source word e as illustrated in Fig. 2. Namely, an entry e in a dictionary is defined as:

$$e.f_1, f_2, \ldots, f_n \Leftrightarrow \forall j \in [1..n], e \in Trig(f_j) \text{ and } f_j \in Trig(e) \quad (4)$$

Tables 3 and 4 give respectively a view of the obtained English-French and French-English dictionaries. The translations of a word are presented on a decreasing order in terms of IG_M.

[1] French and English words occurring more than 7 times.

Table 1: Examples of French words triggered by English words

English trigger word	**French triggered word**	$MI_G \times 10^{-4}$
Cooperation	coopération	38
	collaboration	7
	développement	6
Cooperatives	coopératives	0.31
	entreprises	0.09
	associations	0.09
Difficulty	difficulté	1.5
	difficultés	0.9
	problème	0.5
Disciplines	disciplines	0.22
	règles	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

French trigger word	**English triggered word**	$MI_G \times 10^{-4}$
Coopération	cooperation	38
	development	6
	countries	5
Coopératives	cooperatives	0.31
	women	0.09
	associations	0.08
Difficulté	difficulty	1.5
	difficulties	0.5
	difficult	0.5
Disciplines	disciplines	0.22
	new	0.05
	different	0.04
Compétences	powers	4
	competences	1.2
	competence	1

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

English word	**Potential translations**		
Fish	pêche	poisson	poissons
Fisherman	pêcheur	pêcheurs	pêche
Flag	pavillon	drapeau	navires
Flexible	flexible	souple	travail
Foods	alimentaires	alimentaire	produits
Henceforth	désormais	dorénavant	dès

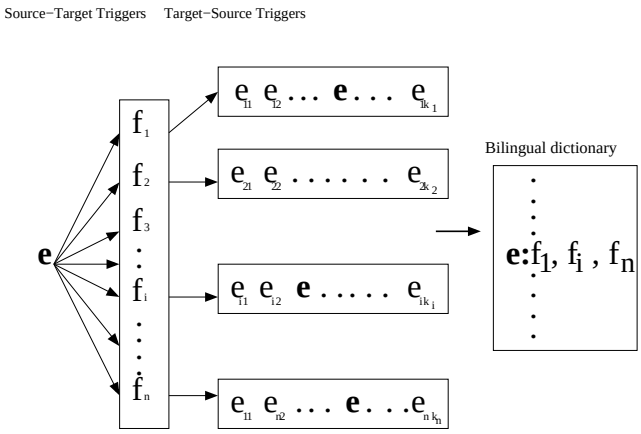

Figure 2: Illustration of dictionary designing

Table 4: An assortment of French-English dictionary

French word	Potential translations		
Humide	wetland	wet	rainforest
Humble	humble	opinion	modest
Mécaniquement	automatically	systematically	necessarily
Méconnaissance	ignorance	lack	knowledge
Sonnette	alarm	sound	bells
Urgence	urgent	urgency	emergency

5. Evaluation and Experiments

To evaluate the pertinence of our dictionary (TrigDic), we compared it with two dictionaries: one distributed by ELRA[2] and a free downloaded one[3]. The comparison is only done on the French-English side. To make the evaluation relevant, we compare only words which exist in TrigDic and in the two other dictionaries. Our dictionary share 10405 words with the ELRA dictionary and 11265 with the Internet dictionary. The evaluation in terms of recall is presented in Table 5.

Table 5: Results in terms of recall

	Rank1	Rank5
ELRA	53%	65%
Internet	41%	52%

The results show that if we consider only the translation given in first position, the recall is 53%, and if we consider the results without taking care about the rank, the recall reaches 65% with ELRA dictionary reference and 52% with the Internet dictionary reference. If we consider the ELRA dictionary as a reference, we can say that our algorithm finds out the pertinent translation of a word in 65% of cases. In a first analysis, we can consider that our algorithm has a failure rate of 35%. A deeper analysis contradicts this assertion. In fact, the failure rate can be explained as follows:

- Into TrigDic we kept only the first five best translations.
- When a potential translation in TrigDic does not exist in ELRA dictionary, we notice that frequently the one we propose is correct and sometimes is very close to the meaning.
- In some cases, the translation proposed by ELRA is less commonly used than ours as shown in Table 6. Then even if the translation we propose is correct, it is not counted as correct.

Table 6: Comparison between ELRA and TrigDic dictionaries

Word	ELRA	TrigDic
Chevaux	Horsefles	Horses, animals, horse
Chimère	Bubble	Illusion, Fantasy, dream, fancy
déléguée	Deputy	Delegated, united, delegate, legislative

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 Pharaoh[4] [3]. To achieve that we assign to each potential word's translation a probability calculated from MI_G. 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_G. 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].

Table 7: Decoding Test Results

Size	GIZA++	TrigDic	extended TrigDic	extended TrigDic with smoothing
59530	0.24	0.18	0.19	0.21
5000	0.24	0.19	0.20	0.21
250	0.25	0.20	0.20	0.22

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.

[2]M0033-3 SCI-FRAN-EURADIC which contains 70832 entries

[3]http://xdxf.revdanica.com/down/index.php which contains 41398 entries

[4]The 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 $P(e|f)$. 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 dictionaries. 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 imaginated 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

[1] F. J. Och and H. Ney, "Improved statistical alignment models," in *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2000, pp. 440–447.

[2] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, I. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky, "Statistical machine translation," in *Final Report, JHU Workshop*, 1999.

[3] P. Koehn, "Pharaoh: A beam search decoder for phrase-based statistical machine translation models," in *6th Conference Of The Association For Machine Translation In The Americas*, Washington, DC, USA, 2004, pp. 115–224.

[4] J. Klavans and E. Tzoukermann, "The bicord system, combining lexical information from bilingual corpora and machine readable dictionaries," in *Proceedings of the 13th Annual Meeting of the Associationof Computational Linguistics*, 1990, pp. 174–179.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[6] D. Hiemstra, "Using statistical methods to create a bilingual dictionary," Ph.D. dissertation, University of Twente, The Netherlands, 1996.

[7] A. Kumano and H. Hirakawa, "Building a MT dictionary from parallel texts based on linguistic and statistical information," in *Proceedings of the 15th COLING*, Japan, 1994, pp. 76–81.

[8] F. Smadja, K. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons: A statistical approach," *Association for Computational Linguistics*, vol. 22, no. 1, pp. 1–38, 1996.

[9] R. Kuhn and R. DeMori, "A cache-based natural language model for speech recognition," *IEEE Trans. PAMI*, vol. 12, no. 6, pp. 570–582, 1990.

[10] C. Tillmann and H. Ney, *Selection criteria for word trigger pairs in language modeling*. Lecture Notes in Artificial Intelligence 1147, Springer Verlag, 1996, pp. 98–106.

[11] ——, "Word trigger and the EM algorithm," in *Proceedings of the Conference on Computational Natural Language Learning*, Madrid, Spain, 1997, pp. 117–124.

[12] W. Kim and S. Khudanpur, "Lexical triggers and latent semantic analysis for cross-lingual language model adaptation," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 2, pp. 94–112, 2004.

[13] P. Koehn, "Europarl: A multilingual corpus for evaluation of machine translation," in *MT Summit*, Thailand, 2005.

[14] K. Papineni and al., "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual of the Association for Computational linguistics*, Philadelphia, USA, 2001, pp. 311–318.

[15] P. F. Brown and al., "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, pp. 263–311, 1993.

[16] I. Zitouni, K. Smaïli, and J.-P. Haton, "Statistical language modeling based on variable-length sequences," *Computer Speech and Language*, vol. 17, pp. 27–41, 2003.

[17] D. Langlois, K. Smaïli, and J.-P. Haton, "Retrieving phrases by selecting the history: application to automatic speech recognition," in *Proc. ICSLP*, Denver, 2002, pp. 721–724.