Training phrase-based SMT without explicit word aligment

Cyrine Nasri, Kamel Smaïli, Chiraz Latiri

Training phrase-based SMT without explicit word alignment

Cyrine Nasri, Kamel Smaili and Chiraz Latiri

SMa^rT, LORIA, Campus scientifique,
BP 139, 54500 Vandoeuvre lès Nancy Cedex, France
cyrine.nasri@loria.fr
smaili@loria.fr
chiraz.latiri@gnet.tn

Abstract. The machine translation systems usually build an initial word-to-word alignment, before training the phrase translation pairs. This approach requires a lot of matching between different single words of both considered languages. In this paper, we propose a new approach for phrase-based machine translation which does not require any word alignment. This method is based on inter-lingual triggers retrieved by Multivariate Mutual Information. This algorithm segments sentences into phrases and finds their alignments simultaneously. The main objective of this work is to build directly valid alignments between source and target phrases. The achieved results, in terms of performance are satisfactory and the obtained translation table is smaller than the reference one; this approach could be considered as an alternative to the classical methods.

Index Terms: Statistical Machine Translation, Inter-lingual triggers, Multivariate Mutual Information.

1 Introduction

The current best performing statistical machine translation systems are based on phrase-based models: the basic idea of phrase-based translation is to segment the given source sentence into phrases, then translate each phrase and finally compose the target sentence from these phrase translations.
It is important to point out that the current phrase-based models are not based on any deep linguistic concept.
Interestingly enough, the power of phrase-based translation is due to the quality of the phrase table. State-of-the-art statistical machine translation uses phrases as translation units to incorporate context into translation models, as described in [4], [14] and [15]. There are many ways to acquire such a table.
The mostly applied phrase pairs extraction method is the so-called Viterbi Extract [15]. In this approach, a source and a target phrase are considered to be

translations of each other, if their words are only aligned within this phrase pair and not to the words outside. Collecting phrases and their corresponding translations extracted from all the sentences in the bilingual training corpus, achieves a phrase table with a set of phrase pairs with scores indicating their translation accuracy.
The decoder based on log-linear model produces target sentences from left to right by covering the source phrases in a certain order [8]. The log-linear model uses several features such as relative frequencies of the phrase pairs, a word-based lexicon model, a target language model, a source phrase reordering model, as well as a word and phrase penalty model.
Currently, this is the most widely method used for producing phrases and decoding.
Other approaches have been investigated to obtain phrase pairs in less heuristic ways. Zhang in [16] presented an integrated phrase segmentation/alignment algorithm (ISA) for statistical machine translation, which segments and aligns phrases simultaneously. Without training a word alignment model, phrases are identified based on the similarities of mutual information values among word alignment points. Venugopal in [13] presented a technique that begins with an improved IBM models to create knowledge, that represents effectively local and global phrase contexts.
Another method proposed by Lavecchia in [8] retrieves valid linguistic phrases without using any alignments. This method identifies first the best part-of-speech phrases and then from these class phrases, they extracted the corresponding phrases which improve the perplexity of the source language. For instance, NOUN PRE NOUN is one of the retrieved part-of-speech phrases and from this pattern and the source corpus a phrase as *Table de Salon* is extracted. The obtained phrases are linguistically pertinent and consequently the derived phrases are also relevant. These obtained phrases are then used to rewrite the source training corpus in terms of phrases. The words of this phrase are gathered and used to rewrite the source training corpus.
In the following, we detail our method which is based on the inter-lingual triggers.

2 Inter-lingual Triggers

Inter-lingual triggers are inspired from triggers concept used in statistical language modeling [12]. A trigger is a set composed of words and its best correlated triggered words in terms of mutual information (MI). In [7], the authors proposed to determine correlations between words coming from two different languages. Each inter-lingual trigger is composed of a triggering source linguistic unit and its best correlated triggered target linguistic units. Based on this idea, they found among the set of triggered target units, potential translations of the triggering source words. Inter-lingual triggers are determined on a parallel corpus according to mutual information measure namely:

$$MI(a,b) = P(a,b)\log\frac{P(a,b)}{P(a)P(b)} \quad (1)$$

where a and b are respectively a source and a target words. $P(a,b)$ is the joint probabilities and $P(a)$ and $P(b)$ are marginal probabilities.

For each source unit a, the authors kept its k best target triggered units. This approach has been extended to take into account triggers of phrases [8]. The drawback of this method is that phrases are built in an iterative process starting from single words and joining others to them until the expected size of phrases is reached. In other words, at the end of the first iteration, sequences of two words are built, the following iteration produces phrase of three words and so on until the stop-criteria is reached. Then, once all the source phrases are built, their corresponding phrases in the target language are retrieved by using *n-to-m* inter-lingual trigger approach which means that a phrase of n words triggers a phrase of m words [8]. In order to avoid the propagation of errors due to the cascade of steps in the previous method, we propose a new approach based on multivariate mutual information which allows to retrieve source phrases given target ones.

3 Training phrase with Multivariate Mutual Information

Multivariate Mutual Information (MMI) calculates the degree of correlation between n random variables. This concept is very interesting since we propose to take advantage of this principle by associating k words in the source language and r words in the target language with $n = k + r$.

$$MMI(A_1, A_2, ..., A_n) = P(A_1, A_2, ..., A_n)\log\frac{P(A_1, A_2, ..., A_n)}{P(A_1)P(A_2)...P(A_n)} \quad (2)$$

Our method allows creating inter-lingual triggers, their estimation is based on MMI. For instance for the trigger *petit déjeuner* $\rightarrow$ *breakfast*, we proceed as follows: A_1 = *petit*, A_2 = *déjeuner*, and A_3= *breakfast*.
$P(A_1, A_2, A_3)$ is the probability that the words*petit, déjeuner* and *breakfast* occur simultaneously. $P(A_1)$, $P(A_2)$, $P(A_3)$ are respectively the probabilities of *petit*, *déjeuner* and *breakfast*.

3.1 Selecting phrases in terms of their size

In this work, we started by identifying the longest phrase with their translations and then, the less longest and finally arriving to phrases of two words. This is motivated by the fact that we would like to appreciate the real contribution of each segment without the influence of its sub-segments. In fact, a long segment is linguistically more informative than a shorter one included into it. The algorithm, we proposed is based on retrieving phrases and their translations by using MMI as described in the following algorithm. For a fixed length of a trigger, the

concatenation of its words constitute the source phrase. And all the words of its triggered sequence constitute the target phrase. It should be noted, that in this algorithm we suppose that the translation table is from English to French. Now we know that, in most cases for a French sentence, its equivalent in English is shorter. That is why, in the proposed algorithm, for each fixed length of phrase, we look for a triggered phrase of the same length or longer by one word. More details are given in [9].

Algorithm 1: Oriented-Size Phrases Discovering (OSPD)

m : maximum size of a source phrase
n : maximum size of a target phrase
for i = m to 1 do
for j = n to 1 do
if i = j or i = j+1 then

1. Train triggers model $X_i \longrightarrow Y_j$ (X_i: source phrase composed of i words, Y_j : target phrase composed of j words)
2. Calculate $MMI(x_1, x_2, \ldots, x_i, y_1, y_2, ..., y_j)$ as shown in formula 2.
3. Include the retrieved phrases $Xi = x_1, x_2, \ldots x_i$ and their best translations $Y_j = y_1, y_2, \ldots y_j$ into the translation table

endif

Fig. 1. Algorithm 1: Oriented-Size Phrases Discovering (OSPD).

This method leads to remarkable triggers where the triggered words could be considered as potential translations of the trigger or very close in terms of meaning. For each source phrase, the 20 best triggers are kept. Table 1 illustrates some examples of obtained French-English triggers.

Table 1. Examples of retrieved phrases and their translations.

French	**English**	**MMI**
autour de la table	around the table	0.0054
	around the	0.0022
	the table	0.0011
a été prise	was taken	0.001
	been taken	0.00037
	has been taken	0.00034
semaine denière	last week	0.016
	week	0.0095
	last	0.009

The experiments presented below have been conducted on the proceeding of the

European Parliament[5]. We used French-English parallel corpus.
Table 2 shows the parallel corpus statistics used in our experiments. So far, we

Table 2. An overview of the experimental material.

Corpus	Sentences	English words	French words
Training	0,5M	15M	16,6M
Dev	1,4 K	14K	13,7K
Test	500	1153	1352

have only discussed how to collect a set of phrase pairs. More is needed to turn this set into a probabilistic phrase translation table. For that, we use the principle proposed in [8] to compute phrase translation probabilities. The translation table is obtained by assigning for each trigger a conditional probability calculated as follows:

$$\forall f, e_i \in Trig(f) \quad P(e_i|f) = \frac{MMI(e_i, f)}{\sum_{e_i \in Trig(f)} MMI(e_i, f)} \tag{3}$$

where *Trig(f)* is the set of k English events triggered by the French event *f*.
In table 3, we present the results of OSPD method. S_1 corresponds to word-to-word translation and S8 corresponds to a translation using all the discovered phrases. The introduction of phrases of 8, 7 and 6 words improve the results by more than 2 points. This means that long phrases are suitable but not as relevant as the introduction of phrases of 5 words. These phrases bring more than 2.5 in terms of BLEU. But the best improvement is brought by sequence of words of 4,3, and 2 words: more than 4.5! Consequently, all these sequences of different sizes are necessary to improve the results. Phrases beyond of 8 words are not relevant. This method has a performance which is 1.1% less than the baseline method. This shows the feasibility to develop machine translation without any word alignment with acceptable results.

4 How to improve this method?

We would like to go further and to achieve results closer to those obtained by the baseline method which needs word alignment. For that, we analyze our method to identify its drawbacks.

4.1 Improving by selecting the best phrases

One of the drawback of this method is that we keep for each phrase its 20 best translations. This could include bad translations which would corrupt the results. In fact, the examples given in table 4 are considered as bad translations

Table 3. Evolution of BLEU in accordance to the length of phrases introduced in the translation table.

Set	Selected Triggers	Score BLEU
S_1	1FR $\longrightarrow$ 1EN	34.16
S_2	S_1 + 8FR	36.32
S_3	S_2 + 7FR	36.36
S_4	S_3 + 6FR	36.8
S_5	S_4 + 5FR	39.58
S_6	S_5 + 4FR	41.12
S_7	S_6 + 3FR	43
S_8	S_7 + 2FR	**43.79**
	Baseline (Och method)	**44.3**

Table 4. Examples of bad English-French triggers.

French phrase	English phrase	MMI $\times 10^{-5}$
les droits de l'homme	human rights situation	3.2
madame le président	president ladies and gentlemen	6
la semaine dernière	president last week	4.6

but unfortunately yet they are in the translation table. In [11] the authors argue that extracting only minimal phrases, i.e the smallest phrases pairs that map each entire sentence pairs, does not degrade performances. By using significance test, they remove unlikely phrase pairs which reduces the phrase table drastically and may even yield increases in performance [3].
We propose in the following pruning phrase table by incorporating best translation pairs in terms of MMI without the need of integrating sequences step by step in terms of their length. We will call this method **Best-Phrases Discovering (BP)**. To do that, we extract all the inter-lingual n-to-m-triggers as in [8] except, that instead of using classical mutual information, we use MMI. Then we sort all the discovered phrases in descending order according to their score. Only triggers that have a MMI greater than a fixed threshold are kept. The impact of this selection is positive in terms of BLEU as presented in the evaluation section.

4.2 Lexical weights

Lexical weights were proposed in [4] to validate the quality of alignments. Given a bilingual phrase to score, the objective consists in checking how well each source word translates into the target words it links to. When a source word links to multiple target words, the average of their translation probabilities is

calculated. A source-to-target lexical weight is then the product of all scores. The same calculation is done from target to source, and the result is a pair of lexical weights between 0 and 1. Because in our method, we do not proceed to any alignment and because lexical weights is important in Moses, we decided to adapt the classical technique with one major change.
As, our method proposes phrases without any initial word-to-word alignments, we estimate a simple lexical translation probability distribution D based on word-to-word triggers [7]:

$$D(w_j|w_i) = \frac{MI(w_j|w_i)}{\sum MI(w_j|w_i)} \tag{4}$$

Where w_i is a word in a language i and w_j is a word in a language j.

5 Evaluation

In the following, we present several experiments to evaluate the impact of the improvements we proposed in order to boost our initial method. The method is evaluated by comparing it to the baseline one with a refined alignments from Giza++ [10]. Default set of options is used: 2 translation probabilities, 2 lexical weights and length penalty.
In figure 2, we evaluate the impact of the method based on selecting the best

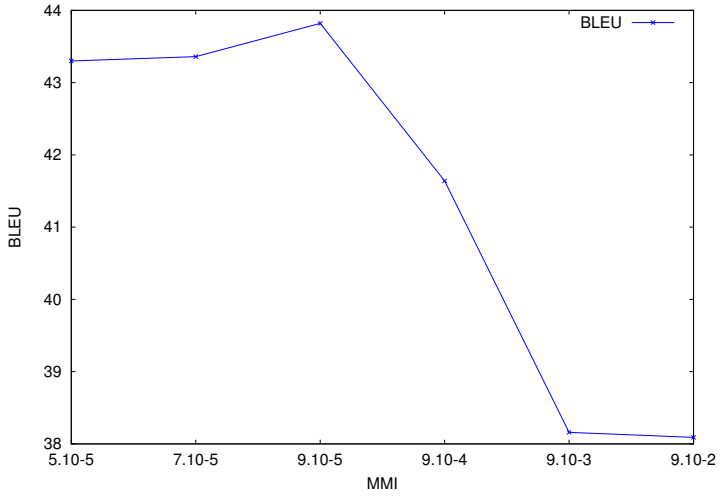

Fig. 2. Evolution of Bleu score in terms of MMI threshold on the DEV corpus

phrases (BP). We can notice that the best performance is achieved for a MMI threshold equal to 9.10^{-5}. This curve shows also, if phrases are not selected

judiciously, then the degradation of the performance is serious, more than 5 BLEU points are lost.
On the Test corpus, the best performance is achieved for a MMI threshold equal to 5.10^{-5} (figure 3). Obviously, the only threshold used is the one get on the development corpus which is close to the one get from the test corpus.

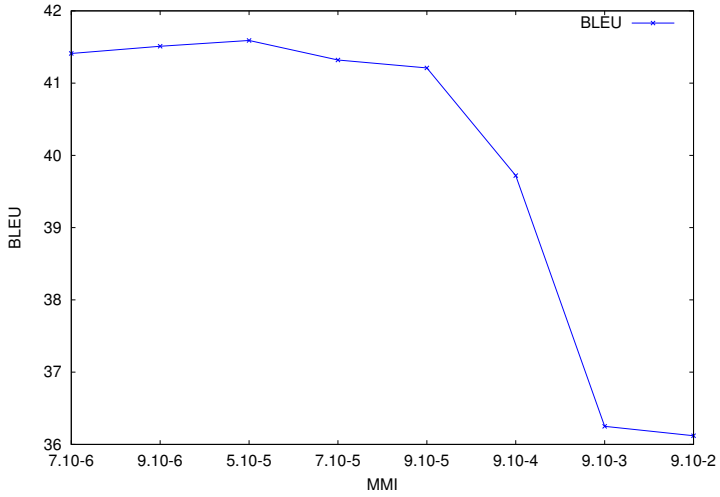

Fig. 3. Evolution of Bleu score in terms of MMI threshold on the Test corpus

In this test, we improve slightly the results, the performance reaches a BLEU of 43.82. The results are still bellow the standard one (44.3). To improve our results we combined several translation tables OSPD, BP and the baseline one. This table shows that by combining our two translation tables we are only 0.6%

Table 5. Experiments with different models

System	**BLEU**
Baseline	44.3
OSPD	43.79
BP	43.82
OSPD+BP	44.02
BP+Baseline	44.48
OSPD+BP+Baseline	44.57

from the baseline method. This result confirms that it is possible to do almost as

good as Och method without any word process alignment. By combining BP with the baseline method, we outperform the baseline result by 0.4%. This improvement reachs 0.6% when both translation tables (OSPD and BP) are combined with the baseline one. These last results illustrates that it is possible to improve the baseline translation table by using other phrases and for some of them, they get better scores than those proposed by the baseline one. Furthermore, our best method (BP) uses a smaller (12%) translation table than the baseline.

Table 6. Size of the different translation tables

Baseline	**OSPD**	**BP**
33,3 M	51 M	22,9 M

6 Conclusion

In conclusion, we proposed two methods which do not require any word alignment. These two methods achieve results closer to the baseline method, around 1%, when they are used alone. When they are combined, we approach the baseline system and the difference is only about 0.6%. The size of the BP translation table is 12% smaller than the baseline one. Then one question may arise, could our community accept a new method of retrieving phrases which does not require any word alignment, getting results closer to the classical one and with a smaller and more cleaned translation table? Besides that, the obtained phrases have an added value since they can enrich those achieved by the baseline method since we have shown that they can improve the reference results

References

[1] Abramson, N. "Information theory and coding", McGraw-Hill electronic sciences series, McGraw-Hill, 1963.

[2] Hoang, H., Birch, A., Callison-burch, C., Zens, R., Aachen, R., Constantin, A., Federico, M., Bertoldi, N., Dyer, C., Cowan, B., Shen, W., Moran, C. and Bojar, O. "Moses: Open source toolkit for statistical machine translation", 177–180, 2007.

[3] Johnson, H., Martin, J., Foster, G. and Kuhn, R., "Improving Translation Quality by Discarding Most of the Phrase table", Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)", 967–975, 2007.

[4] Koehn, P., Franz.J. and Marcu, D. "Statistical phrase-based translation", Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, - NAACL '03, Edmonton, Canada, 48–54 2003.

[5] Koehn, P. “Europarl: A Parallel Corpus for Statistical Machine Translation”, Conference Proceedings: the tenth Machine Translation Summit, 79–86, 2005.
[6] Lavecchia, C., Smaïli, K., Langlois, D. and Haton, J-P. “Using inter-lingual triggers for machine translation”, INTERSPEECH, 2829-2832, 2007.
[7] Lavecchia, C., Smaïli, K., Langlois, D. and Haton, J-P. ”Using inter-lingual triggers for machine translation“, INTERSPEECH, 2829-2832 2007
[8] Lavecchia, C., Langlois, D. and Smaïli, K. ”Discovering phrases in machine translation by simulated annealing“ Interspeech, 2354-2357, 2008.
[9] Nasri, C. Smaïli, K., Latiri, C. and Slimani, Y., “A new method for learning Phrase Based Machine Translation with Multivariate Mutual Information”, NLP-KE'12, HuangShan, China, 2012.
[10] Och, F. and Hermann, N., “A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, 29, 2003.
[11] Quirk, C. and Menezes, A. “Do we need phrases? Challenging the conventional wisdom in Statistical Machine Translation”, HLT-NAACL, 2006.
[12] Tillmann, C., Ney, H. and Lehrstuhl Fur Informatik Vi “Word Triggers and the EM Algorithm”, In Proceedings of the Workshop Computational Natural Language Learning (CoNLL 97), 117–124.
[13] Venugopal, A., Vogel, S., and Waibel, A. “Effective Phrase Translation Extraction from Alignment Models”, ACL, 319-326, 2003.
[14] Zens, R., Och, F.J. and Hermann, N., “Phrase-Based Statistical Machine Translation”., Springer Verlag, 18–32, 2002.
[15] Zens, R. and Ney. H. “Improvements in Phrase-Based Statistical Machine Translation”, The Human Language Technology Conf, HLT-NAACL, 257–264, 2004.
[16] Zhang, Y., Vogel, S. and Waibel, A., “Integrated Phrase Segmentation and Alignment Algorithm for Statistical Machine Translation”, Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'03), Beijing, China, 2003