A new method for learning Phrase Based Machine Translation with Multivariate Mutual Information

Cyrine Nasri, Kamel Smaïli, Chiraz Latiri, Yahya Slimani

International Journal of Advanced Intelligence
Volume 0, Number 0, pp.XXX-YYY, November, 20XX.

AIA International Advanced Information Institute

A new method for learning Phrase Based Machine Translation with Multivariate Mutual Information

Cyrine Nasri
URPAH, Faculty of Sciences of Tunis, Tunisia
PAROLE, LORIA, Campus scientifique, BP 139, 54500
Vandoeuvre Lès Nancy Cedex, France
cnasri@loria.fr

Kamel Smaili
PAROLE, LORIA, Campus scientifique, BP 139, 54500
Vandoeuvre Lès Nancy Cedex, France
smaili@loria.fr

Chiraz Latiri
URPAH, Faculty of Sciences of Tunis, Tunisia
chiraz.latiri@gnet.tn

Yahya Slimani
URPAH, Faculty of Sciences of Tunis, Tunisia
Yahya.slimani@fst.rnu.tn

Current statistical machine translation systems usually build an initial word-to-word alignments before learning phrase translation pairs. This operation needs so many matching between different single words of both considered languages. We propose a new approach for phrase-based machine translation which does not need any word alignments, it is based on inter-lingual triggers determined by Multivariate Mutual Information. This algorithm segments sentences into phrases and finds their alignments simultaneously. The main objective is to build directly valid alignments between source and target phrases. Inspite of the youth of this method, experiments showed that the results are competitive but needs some more efforts in order to overcome the one of state-of-the-art methods.

Keywords: statistical machine translation; inter-lingual triggers; multivariate mutual information.

1. Introduction

Machine translation issue could be handled by several ways, some of them are syntax-based (see Ref. 1, 2, 3 and 4), others are based on statistical models. Nowadays, a more attractive approach is to see how to combine a purely statistical technique with some linguistic rules or model (see Ref. 6)
The work presented in this paper is based on statistical method. The principle consists in finding the best translation of a source sentence among several ones. Thus,

translating a sentence from language A into B involves finding the best target sentence b^* wich maximizes the probability of b given the source sentence a. Bayes rule allows to formulate the probability $\mathrm{P}(b|a)$ as follows:

$$b^* = \underset{b}{\mathrm{argmax}} \mathrm{P}(b|a) = \underset{b}{\mathrm{argmax}} \mathrm{P}(a|b)\mathrm{P}(b) \quad (1)$$

The translation process needs a language model $\mathrm{P}(b)$, a translation model $\mathrm{P}(a|b)$ and a decoder wich calculates b^*. Language model parameters are trained on a target corpus and its task is to build up a correct sentence from partial translations, whereas parameters of the translation model are determined from a parallel corpus and provides the probability that a linguistic unit is translated to another. Then, the decoder provides the best target sentence by taking into account several parameters provided among other the previous models. In this work, we develop a new algorithm for extracting phrase pairs from parallel corpus. This algorithm does not require an initial segmentation on the monolingual text. It uses inter-lingual triggers based on Multivariate Mutual Information between the source and target phrases. In fact, we propose an original method which retrieves automatically phrases and their corresponding translations in one step. It means that a phrase translation is not constructed by agglutining connected words in the target language.
The remainder of the paper is organized as follows : section 2 gives an overview of statistical phrase-based machine translation and interlingual triggers. In section 3 and 4 we present our method for learning phrase translations. Section 5 describes how we estimate probabilities for phrase pairs to fit into the SMT decoder. Section 6 shows how we integrate and test our new approach into an entire translation process. Conclusion in section 7, points out the strength of our method and gives some tracks about future work.

2. Background and related works

2.1. *Statistical phrase-based machine translation*

At present, best performing statistical machine translation systems are based on phrase-based models: models that translate small word sequences at a time.
First statistical methods (see Ref. 5) were word-based models, but words, as shown, in later works gige worse results than those based on longer units. But words may not be the best candidates for the smallest units for translation. Sometimes one word in foreign language is translated into two English words, or vice versa. For instance, in French "*petit déjeuner*" is translated in English by "*breakfast*". It is important to note that current phrase-based models are not rooted in deeply linguistic in approach of phrases. Koehn *and al.* defined a phrase as a contiguous multiword sequence, without any linguistic motivation (see Ref. 6). Phrases are mapped one-to-one based on a phrase translation table, and may be reordered. All phrase pairs that are consistent with the word alignment are added to the phrase table.
Modern statistical phrase-based models are based on alignment template models (see Ref. 7 and 8).

These models defined phrases over word classes that were then instantiated with words. Several methods to extract phrases from a parallel corpus have been proposed. Most make use of word alignments (see Ref. 9, 10 and 11). Phrase alignment may be done directly from senetce-aligned corpora using a probabilistic model (see Ref. 12), pattern mining methods (see Ref. 13), or matrix factorization (see Ref. 14).
In this respect, Lavecchia proposed in Ref. 15 a method which retreives 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. 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. Let us give an example, NOUN DET 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 words of this phrase are gathered and used to rewrite the source training corpus.

2.2. *Inter-lingual Triggers*

Inter-lingual triggers are inspired from triggers concept used in statistical language modeling (see Ref. 16). A trigger is a set composed of words and its bests correlated triggered words in terms of mutual information (MI). In Ref. 17, authors proposed to determine correlations between words coming from two different languages. Each inter-lingual trigger is composed by a triggering source linguistic unit and its best correlated triggered target linguistic units. Based on this idea, we 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:

$$MI(a,b) = \mathrm{P}(a,b) \log \frac{\mathrm{P}(a,b)}{\mathrm{P}(a)\mathrm{P}(b)} \tag{2}$$

where a et b are respectively a source and a target words. Notice that $\mathrm{P}(a,b)$ is the joint probability and $\mathrm{P}(a)$ and $\mathrm{P}(b)$ are the marginal probabilities.
For each source unit a, we kept its k best target triggered units. Interestingly enough, this approach has been extended to take into account triggers of phrases (see Ref. 15). The drawback of this method is that phrases are built in an iterative process starting from single words and joining others to them until expected size of phrases is reached. In others words, at the end of the first iteration, sequences of two words are built, the following iteration produces phrases 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-n* inter-lingual trigger approach (see Ref. 15). 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 retreive source phrases given target once.

3. Description of the method

The new approach is based on multivariate mutual information. Before presenting our new approach, we introduce some necessary formalizations related to the multivariate mutual information (MMI).

3.1. *Principle of multivariate mutual information*

Typically, mutual information is defined and studied between just two variables. Though the approach to evaluate bivariate mutual information is well established, several problems in multi-user information theory require the knowledge of interaction between more than two variables. Since there exists dependency between the variables, we cannot decipher their relationship without considering all of them at once. The seminal work on the information-theoretic analysis of the interaction between more than two variables was first studied in Ref. 22.
The definition of mutual information has been extended to e general case (over more than three variables) by Fano (see Ref. 23) and re-formulated in a lattice-theoretic framework by Han (see Ref. 24). Though each have taken different approaches and the expressions are in terms of different entities (mutual informations on one case and entropies in the other), they can be simplified to be the same.
In this sense, we define the multivariate mutual information as follows:

$$MMI(X,Y) = \sum_{\substack{x_1..x_n \in X \\ y_1..y_m \in Y}} P(x_1, x_2, ..., x_n, y_1, y_2, ..., y_m) \log \frac{P(x_1, x_2, ..., x_n, y_1, y_2, ..., y_m)}{P(x_1, x_2, ..., x_n)P(y_1, y_2, ..., y_m)} \tag{3}$$

3.2. *How to take advantage from multivariate mutual information in order to learn phrase translation?*

Multivariate mutual information calculates the correlation between respectively n and m variables. This concept is very interesting since we propose to take advantage from this principle by associating n words in a source language to m words in a target language. The objective as in Ref. 17 is to use the principle of interlingual triggers except that we use a multivariate mutual information. As illustrative example, guess that we are interested by phrase of length 3 which are translated by two words. For instance, in French "*le petit déjeuner*" is translated by "*the breakfast*" in English. We can calculate directly the correlation degree between two linguistics units as follows:

$$MMI(v,w,x,y,z) = P(v,w,x,y,z) \log \frac{P(v,w,x,y,z)}{P(v,w,x)P(y,z)} \tag{4}$$

With v = "le", w = "petit", x = "déjeuner", y = "the" and z = "breakfast".

4. A new algorithm for training phrases

One of the famous algorithm allowing to develop a phrase based model (see Ref. 6) is based on splitting the source language on several segments and each segment is then translated. Segments correspond to what are called phrases, they are those which are consistent with the word alignment. Words are aligned bidirectionally and the phrases are those with a high union recall alignment. A reordering model is trained using a joint probability which has the role to put in order the phrases of the target language. A reordering model is trained using a joint probability which has the role to put in order phrases of the target language. Consequently, at least the following parameters are necessary to develop a phrase based model: a bidirectional phrase translation probability and a bidirectional lexical translation probability.
In this work, we start by identifying the longest phrases with thier translations and then the less longest and we finish with 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 propose in the following is based on retrieving phrases and their translations by using multivariate mutual information without any alignment. Firstly, this algorithm provides a many-to-many translation table, and MMI permits to find phrases like x_1, x_2, ..., $x_m \longrightarrow y_1$, y_2, ..., y_n, such as m and n are respectively the maximal length of source and target phrases. suppose that an English phrase is in general longer than the French one. This is not true in all the cases but in most cases this hypothesis is true. By iterating the different steps of Algorithm 1, we get a list of phrases and their translations.

Algorithm 1 A phrase model based on multivariate mutual information

1: **for** $i = lenMax$ to *2* **do**
2: Extract the n_i best phrases (most frequent) from the french corpus F
3: Extract the n_i best phrases (most frequent) from the english corpus E
4: **end for**
5: **for** $i = lenMax$ to *2* **do**
6: $\mathrm{MMI}(f_1, ..., f_i, a_1, ..., a_i) = \mathrm{P}(f_1, ..., f_i, a_1, ..., a_i) \log \frac{P(f_1,...,f_i,a_1,...,a_i)}{P(f_1,...,f_i)P(a_1,...,a_i)}$
7: **end for**

5. Estimate probabilities for phrase pairs

Translation probabilities reflect the probability that a sequence of words in a source language translates into sequence of words in a target language. We use the principle proposed in Ref. 15 to compute phrase translation probabilities.

Table 1. Examples of retrieved phrases and their translations

French	English	MMI	Prob
autour de	around	0.0054	0.276
	around the	0.0022	0.122
a été prise	was taken	0.001	0.212
	been taken	0.00037	0.074
	taken	0.00034	0.068
semaine dernière	last week	0.016	0.194
	week	0.0095	0.16
	last	0.009	0.158

Table 2. An overview of the experimental material

Corpus	Sentences words	English words	French
Train	596831	15138093	16613485
Dev	1444	14077	13776
Test	500	1153	1352

In our algorithm phrases and their translations are obtained by selecting those which have the best values of multivariate mutual information. In order to build, a translation table which can be processed by Moses, we have to transform each MMI into a probability. For that, we proceed as in Ref. 16

$$\forall f, e_i \in Trig(f) P(e_i|f) = \frac{MMI(e_i, f)}{\sum_{e_i \in Trig(f)} MMI(e_i, f)} \tag{5}$$

Where *Trig(f)* is the set of k English segments triggered by the French phrases f. Table I illustrates retrieved phrases by our algorithm. The first column presents French phrases. For each French phrase, best corresponding translations are presented in column 2. The third column indicates the value of MMI assigned to the translation proposed in the second column. In the same way, the fourth column indicates the probability. A qualitative analysis showed that our method leads to pertinent inter-lingual triggers. Thus, triggered sequences could often be considerated as potential translation of the triggering French phrase. And finally the fourth column shows the probability of the french phrase. It is calculated as we have shown in equation 5.

6. Experiments

In this section we evaluate our phrase-based system based on Multivariate Mutual Information. The experiments presented below have been conducted on the proceeding of the European Parliament (see Ref. 18). We used French-English parallel corpus.

Table 2 gives details about the used the parallel corpus. We use a train corpus to extract French phrases and to compute inter-lingual triggers (few examples are given in table 1). A development corpus is used to select the best phrase translations among all those determined by the set of inter-lingual triggers. Finally, we use a test corpus to validate our approach.

Table 3. Evaluation results on the Europarl corpora : French to English MT task. The test corpus contains 500 aligned sentences

System	BLEU
Baseline	44.3
MMI	43.79

Table 4. Evolution of BLEU in accordance of phrases's types

Set	Selected Triggers	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

We compare the output of MOSES (see Ref. 19) using its default phrase table (refined alignments from Giza++ (see Ref. 20)), against those produced by our method. For the language model, we use a trigram for all the experiments. Results are presented in table 3 ans 4 in terms of BLEU (see Ref. 21).
Table 3 shows the results obtained with our method are close to the baseline system. Table 4 illustrates the evolution of BLEU in accordance to the size of phrases introduced in translation table. The first conclusion is the one that proved several years ago by the community that the introduction of phrases improve the results. Results S1 corresponds to word-to-word translation and S8 corresponds to a translation using all the possible phrases. The improvement eceeds 9 points which is a considerable achievement. Then what we can remark is that, the introduction of phrases of 8,7 and 6 words improve the results with more than 2 points. This mean that influence of long phrases is 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.

7. Conclusion

The proposed approach in this work is based on the concept of multivariate mutual information. This measure is used to determine directly many-to-many phrases. The first positive result is that this approach allowed to find out valid linguistic phrases and their corresponding translations.
The second advantage is that our method does not need word alignments. Technically we do not need to include any alignment variable in the calculation of the translation probability. So, the translation probability is calculated directly through the correlation between the source and the target corpora. The matching between the source and the target segments is handled by associating the best target segment to a source segment. Consequently, the calculation of the translation table is

faster than the baseline method and provides results less noisy.
The investigation we did to explain the difference between our method and the baseline system is likely due to the non discriminative probabilities of our translations. Indeed, the translation probability assigned to a pair of phrases is calculated by a standard normalization of the multivariate mutual information, the consequence is that the probabilities are close to each other and this does not allow a high discrimination between partial translations in the decoding step. Work is under progress to overcome this limit.

References

1. Dekai Wu. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23:377–403, 1997.
2. Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 523–530, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.
3. Daniel Gildea. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 80–87, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
4. David Chiang. Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228, June 2007.
5. Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311, June 1993.
6. Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
7. Franz Josef Och and Hans Weber. Improving statistical natural language translation with categories and rules. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, COLING '98, pages 985–989, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.
8. Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 295–302, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
9. Christoph Tillmann and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, 2003.
10. Bing Zhao and Stephan Vogel. A Generalized Alignment-Free Phrase Extraction. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 141–144, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.
11. Hendra Setiawan, Haizhou Li, Min Zhang, and Beng Chin Ooi. Phrase-based statistical machine translation: A level of detail approach. In *IJCNLP*, pages 576–587, 2005.
12. Jung H. Shin, Young S. Han, and Key sun Choi. Bilingual knowledge acquisition from korean-english parallel corpus using alignment method (korean-english alignment at word and phrase level). In *COLING-96: The 16th International Conference on Computational Linguistics*, pages 230–235, 1996.
13. Kaoru Yamamoto, Taku Kudo, Yuta Tsuboi, and Yuji Matsumoto. Learning sequence-to-sequence correspondences from parallel corpora via sequential pattern mining. In Rada Mihalcea and Ted Pedersen, editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 73–80, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.

14. Cyril Goutte, Kenji Yamada, and Eric Gaussier. Aligning words using matrix factorisation. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
15. Caroline Lavecchia, David Langlois, and Kamel Smaïli. Phrase-based machine translation based on simulated annealing. In *LREC*, 2008.
16. Christoph Tillmann, , Christoph Tillmann, Hermann Ney, and Lehrstuhl Fur Informatik Vi. Word triggers and the em algorithm. In *In Proceedings of the Workshop Computational Natural Language Learning (CoNLL 97*, pages 117–124, 1997.
17. Caroline Lavecchia, Kamel Smaïli, David Langlois, and Jean-Paul Haton. Using inter-lingual triggers for machine translation. In *Annual Conference of the International Speech Communication Association*, pages 2829–2832, 2007.
18. Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. Draft, 2002.
19. Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondřej Bojar. Moses: Open source toolkit for statistical machine translation. pages 177–180, 2007.
20. Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March 2003.
21. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
22. William McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954.
23. R. M. Fano. Transmission of Information: A Statistical Theory of Communication. 1961.
24. Te Sun Han. Multiple Mutual Informations and Multiple Interactions in Frequency Data. *Information and Computation/information and Control*, 46:26–45, 1980.