Training Statistical Machine Translation with Multivariate Mutual Information

Cyrine Nasri, Kamel Smaïli, Chiraz Latiri

Training Statistical Machine Translation with Multivariate Mutual Information

C. Nasri[1], K. Smaïli[2] and C. Latiri[1]

[1]URPAH, Faculty of Sciences of Tunis, Tunisia
[2]Loria BP 23 54506 Vandoeuvre Lès-Nancy, France
cyrine.nasri@gmail.com, smaili@loria.fr, chiraz.latiri@gnet.tn

Abstract

In this paper, we describe a new model for phrase-based statistical machine translation. Roughly speaking, statistical approach uses a language and a translation model. This latter could be viewed as a lexical and an alignment model. The approach we propose does not need any alignment, it is based on inter-lingual triggers determined by multivariate mutual information (MMI). This measure depends on conditional mutual information, this means that a source phrase is directly linked to a target one. The conditional mutual information is used in both directions (source-target and target-source languages). We present an experimental evaluation conducted on EUROPARL corpora (French and English) and using the decoder MOSES. We compare then our results to those of a previous work in which we used inter-lingual triggers determined by a simple mutual information (MI) as well as to those given by baseline model (Koehn et al., 2003).

Keywords: Phrases, Machine translation, inter-lingual triggers, Conditionnal Mutual Information, BLEU

1. Introduction

The machine translation issue could be handled by several ways, among them are syntax-based ones (Wu, 1997; Yamada and Knight, 2001; Gildea, 2003; Chiang, 2005), others are based on statistical models and some of them combine statistical and syntax models. 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^* which maximizes the probability of b given the source sentence a. Bayes rule allows to formulate the probability $P(b|a)$ as follows:

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

The translation process needs a language model $P(b)$, a translation model $P(a|b)$ and a decoder. 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 into another. Then, the decoder provides the best target sentence by taking into account several parameters provided among other by the previous models.

In literature, first statistical machine translation (SMT) systems were word-based (Brown and al., 1993). Nowadays, all SMT are phrase-based. A phrase is a sequence of words determined automatically by using a complex algorithm based on the alignment of several words from the source language with one or several sequence of words in the target language. In order to retrieve phrases, several approaches have been proposed in the last decade, we can cite for example those which are based on statistical approach (Wang and Waibel, 1998; Och, 1999). Most of them require word-based alignments. For instance, in (Och, 1999), Och collected all phrase pairs that were consistent with the word alignment provided by Brown's models. Thus, any contiguous source words must be a translation of any contiguous target words if and only if words are aligned with each other. Besides, the retrieved phrases are not based on linguistic knowledge, consequently they could lead to noisy sequence of words.

In this respect, we proposed in a previous work (Lavecchia et al., 2008) a method which retrieves valid linguistic phrases without using any alignment. This method identifies first the best part-of-speech phrases and then from these class phrases we extract 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 phrases are then used to rewrite the source training corpus in terms of phrases. Let us give an example, NOUN [1] DET [2] NOUN is one of the retrieved part-of-speech phrases and from this pattern and the source corpus

[1]A noun class
[2]A determinant class

a phrase as *Table de Salon*[3] is extracted. The words of this phrase are gathered and used to rewrite the source training corpus.

With the inter-lingual triggers method (Lavecchia et al., 2007), we can extract the corresponding target phrases. The proposed algorithm consists in finding out all the triggered sequence of words of length 1, 2, 3 and so on. However, this algorithm does not out-perform the baseline method in termes of BLEU score.
In this article, we propose an original method which retrieves automatically the 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 inter-lingual triggers. In Sections 3 and 4 we present our method for learning phrase translations. Section 5 describes how we integrate and test our approach into a entire translation process. Conclusion in Section 6 points out the strength of our method and gives some tracks about future work.

2. Inter-Lingual Triggers

Inter-lingual triggers are inspired from triggers concept used in statistical language modeling (Tillmann and Ney, 1997). A trigger is a set composed of a word and its best correlated triggered words in terms of mutual information (MI). We proposed in (Lavecchia et al., 2007) 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 units. Inter-lingual triggers are determined on a parallel corpus according to mutual information, namely:

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

where a and b are respectively a source and a target units. Notice tnat $P(a,b)$ is the joint probabilities and $P(a)$ and $P(b)$ are 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 (Lavecchia et al., 2008) and (Latiri et al., 2011). 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, sequence of two words are built, the following iteration produces phrase of three words and so on until the stop-criteria is checked. Then, once all the source phrases built, their corresponding phrases in the target language are retrieved by using *n-to-m* inter-lingual trigger approach. This methods leads to an improvement in comparison to single word approach of $5,7\%$ as shown in table 1. This method has been combined with another based on data mining approach (DM method) (Latiri et al., 2011), the obtained result outperformed the one-to-one method by $14,6\%$. Despite of these significative improvements, we have not succeed to outperform Och's method. In order to improve our result and to avoid the propagation of errors due to the cascade of steps in the previous method, we propose a new approach based on conditional mutual information which allows to retrieve source phrases given target once.

Method	**BLEU**
one-to-one	30,97
n-to-m triggers	34,41
n-to-m triggers+DM method	35,52

Table 1: Performances of phrase-based machine translation based on triggers and on a combination with datamining-based method (under PHARAO decoder)

3. Description of the method

The new approach is based on multivariate mutual information which itself is founded on conditional mutual information (CMI). Before presenting our approach, we introduce some necessary formalizations related to CMI.

3.1. Principle of conditional mutual information

Given 3 discrete random variables X,Y,Z, the conditional mutual information of X, Y given Z is expressed as follows:

$$I(X,Y|Z) = \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} P(x,y,z)log\frac{P(x,y,z)P(z)}{P(x,z)P(y,z)} \qquad (3)$$

The multivariate mutual information of three variables is expressed as the difference between the mutual information of X, Y given Z, *i.e.,* conditional

[3]coffee table

mutual information, and the mutual information between $X$and Y. Thus, formally, we have:

$$I(X,Y,Z) = I(X,Y) - I(X,Y|Z) \quad (4)$$

We can prove that the mutual information verifies the following formula:

$$\begin{aligned} I(X,Y,Z) \quad &= I(X,Y) - I(X,Y|Z) \\ &= I(X,Z) - I(X,Z|Y) \\ &= I(Y,Z) - I(Y,Z|X) \end{aligned} \quad (5)$$

The projection of X, Y, Z on respectively the words x, y and z leads to equation 6:

$$I(x,y,z) = I(x,y) - I(x,y/z) \quad (6)$$

This formula will be used in the remaining of this paper. Indeed, Formula 5 could be generalized to several variables as follows:

$$\begin{aligned} I(X_1, X_2, \ldots, X_n) \quad &= I(X_1, X_2, \ldots, X_{n-1}) \\ &\quad -I(X1, X_2, \ldots X_{n-1}|X_n) \end{aligned} \quad (7)$$

This formula will be used in a forthcoming work.

3.2. How to take advantage from conditional mutual information in order to build phrases?

Multivariate mutual information calculates recursively the correlation relationship between n variables by splitting the sequence of variables respectively into two segments composed of $(n-1)$ words and a single one. Then the same splitting operation is done recursively with the segment of $(n-1)$ variables. This concept is very interesting since we propose to take advantage from this principle by associating the first variables to the words of the source language and the last one to a word taken from the target language. The objective as in (Lavecchia et al., 2007) is to use the principle of inter-lingual triggers except that we use a multivariate mutual information. As illustrative example, guess that we are interested by phrases of length 2 which are translated by one word. For instance, in French "*petit déjeuner*" is translated by "*breakfast*" in English. We can then calculate directly the correlation degree between these two linguistic units as follows:

$$\begin{aligned} \text{I(x,y,z)=} \quad & P(x,y) \times \log \frac{P(x,y)}{P(x)P(y)} \\ & \text{-}P(x,y,z) \times \log \frac{P(z)P(x,y,z)}{P(x,z)P(y,z)} \end{aligned} \quad (8)$$

With x="petit", y="déjeuner" and z="breakfast". This formula shows that the relationship between the words of the source target phrase and the word of the target language is stronger. In fact, the equation takes into account the relationship between each composant of the the source target phrase and the word of the target language. We believe that this will lead to more realistic phrases with more relevant translations.

4. A new algorithm for training phrases

One of the famous algorithm allowing to develop a phrase-based model (Koehn et al., 2003) is based on splitting the source language on several segments and each segment is then translated. The segments correspond to what we call phrases, they are those which are consistent with the word alignment. Words are aligned bidirectionally and the phrases are those with a high intersection precision alignment and 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. 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.

The algorithm we propose is based on retrieving phrases and their translations by using multivariate mutual information without any alignment and without using the previous parameters as detailed in algorithm 1. Firstly, this algorithm provides a one-to-one translation table, and MMI permits to find phrases like $xy \rightarrow z$. Then the source corpus is rewritten with the best phrases.

Calculating again (step 5) triggers $1 \rightarrow 1$ provides pairs of phrases respecting the pattern $n_{words} \rightarrow n_{words}$ and $(n+1)_{words} \rightarrow n_{words}$. By inverting S and T, step 2 provides, this time phrases in the form of $n_{words} \rightarrow n_{words}$ and $(n+1)_{words} \rightarrow n_{words}$. This means that we produce a translation longer or equal than the French one. By iterating the different steps of Algorithm 1, we get a list of phrases and their translations.

5. Experiments

In the following, we present several results of experiments conducted on french-english EUROPARL corpus (Koehn, 2005). Our phrase based model is compared to the reference one (Koehn et al., 2003). A trigram language model is used for all the experiments and the parameters are tuned for each model by using MERT provided with MOSES's decoder (Koehn et al., 2007).

5.1. Material

Table 2 shows the parallel corpus statistics used in our experiments.

Algorithm 1: A phrase model based on multivariate mutual information

1. S is a source corpus and T a target corpus.
2. Train a trigger model $1 \rightarrow 1$ where the left sequence is taken from S and the right one from T. For each source sequence keep the k best ones.
3. Train $MIM(x, y, z)$ which correspond to triggers $2 \rightarrow 1$ where a couple of words from S triggers a word from T quoted $Trig_{2\rightarrow 1}^{S\rightarrow T}$. Include the retrieved phrases xy and its translation z into the dictionary by grouping x and y.
4. Rewrite S with the right member of the best triggers. To achieve that, sort the phrases (the right member of triggers) in a decreasing order of the probability. We do not start rewriting from left to right but by replacing in the corpus the best phrases and by the best second phrases, etc.
5. Calculate triggers $1 \rightarrow 1$ where the left sequence is taken from S and the right one from T.
6. Clean the translation table by selecting the best phrases
7. Inverse S and T and Goto step 2 until the desired length of phrases is reached.

Corpus		English	French
Training	Sentences	596831	596831
	Words	15138093	16613485
	Vocabulary	59838	76946
Development	Sentences	1444	1444
	Words	14077	13770
	Vocabulary	2274	2701
Test	Sentences	500	500
	Words	4945	5249
	Vocabulary	1153	1352

Table 2: An overview of the experiment material

We conduct several tests in order to determine the best phrases which improve the BLEU score. This is done by cleaning the translation table: removing useless phrases and their translations and by keeping only some combination of phrases and their translations as illustrated in Table 3.

Set	Selected triggers	BLEU
S1	$1_{en} \rightarrow 1_f$	31.02
S2	$S1 + 1_{en} \rightarrow 2_f$	31,18
S3	$S2 + 1_{en} \rightarrow 3_f$	31,2
S4	$S3 + 2_{en} \rightarrow 2_f$	32,28
S5	$S4 + 2_{en} \rightarrow 3_f + 2_{en} \rightarrow 4_f$ $+2_{en} \rightarrow 5_f$	38,4
S6	$S5 + 3_{en} \rightarrow 2_f + 3_{en} \rightarrow 3_f$ $+3_{en} \rightarrow 4_f$	38,52
S7	$S6 + 4_{en} \rightarrow 4_f$	38,52
S8	$S7 + 4_{en} \rightarrow 5_f + 4_{en} \rightarrow 6_f$	38,93
S9	$S8 + 5_{en} \rightarrow 5_f + 5_{en} \rightarrow 6_f + 5_{en} \rightarrow 7_f$ $+ 6_{en} \rightarrow 6_f + 6_{en} \rightarrow 7_f + 6_{en} \rightarrow 8_f$ $+ 7_{en} \rightarrow 7_f + 7_{en} \rightarrow 8_f + 7_{en} \rightarrow 9_f$ $+ 8_{en} \rightarrow 8_f + 8_{en} \rightarrow 9_f + 8_{en} \rightarrow 10_f$	39,43
	Koehn	42,78

Table 3: Evolution of BLEU in accordance of phrases's types

Table 3 illustrates the importance of long phrases, each set of this table includes phrases of the previous set and new longer ones. For each new set of phrases, we improve the performance of the precedent one. This highlights the importance of using long phrases. The last set is composed of at most eight English words, this is motivated by the fact that the state of the art's translation table contains phrases of at most 8 words.
The parameters used in Koehn's method as well as in our MMI based approach are given in table 5.1..

Method	LM	TM	WP	WD
Koehn	0,11	0,09 0,17 0,35	-0,12	0,14
MMI	0,09	0,14 0,05 0,5	-0,07	0,04

Table 4: MERT optimized configuration for both methods (LM: Language Model weight, TM: Translation Model weight, WP: word penalty and WD: Weight distorsion).

6. Conclusion

Out contribution in this paper performs phrase-based statistical machine translation. The proposed approach is based on the concept of multivariate mutual information. In fact, 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, this could be proved by the high improvement obtained by the introduction of phrases in the translation

table which enhances the result by $27,1\%$ whereas the progress we get by the method presented in previous work (Lavecchia et al., 2008) and (Latiri et al., 2011) are respectively $5,6\%$ and $14,7\%$. The second advantage is that our method does not need any alignment. In fact, we succeed to identify common phrases in the source and target languages by using multivariate mutual information, technically we do not need to include an 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. The best segment is found out by using multivariate mutual information which is depending on conditional mutual information. Consequently, the calculation of the translation table become faster and our method produces less noise.

The investigation we did to explain the discrepancy between our method and Koehn one 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 limite.

7. References

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

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Gildea. 2003. 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. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. 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. Association for Computational Linguistics.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit*, september.

Chiraz Latiri, Kamel Smaïli, Caroline Lavecchia, Cyrine Nasri, and David Langlois. 2011. Phrase-based machine translation based on text mining and statistical language modeling techniques. In *12th International Conference on Intelligent Text Processing and Computational Linguistics - CICLing2011*, Tokyo, Japon.

Caroline Lavecchia, Kamel Smaili, David Langlois, and J.P. Haton. 2007. Using inter-lingual triggers for machine translation. In *Proc. Interspeech*, pages 2829–2832, Antwerp, Belgium.

C. Lavecchia, K. Smaili, and D. Langlois. 2008. Discovering phrases in machine translation by simulated annealing. In *proceedings of the Eleventh Interspeech, Brisbane, Australia, September*.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *EACL*, pages 71–76.

C. Tillmann and H. Ney. 1997. Word trigger and the EM algorithm. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 117–124, Madrid, Spain.

Y. Wang and A. Waibel. 1998. Modeling with structures in statistical machine translation. In *ACL Proceedings*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23:377–403, September.

Kenji Yamada and Kevin Knight. 2001. 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. Association for Computational Linguistics.