Phrase-based machine translation based on text mining and statistical language modeling techniques

Chiraz Latiri[1], Kamel Smaïli[2], Caroline Lavecchia[2], Cyrine Nasri[1], and David Langlois[2]

[1] URPAH Team, Computer Sciences Department, Faculty of Sciences of Tunis
El Manar University, Tunisia
[2] LORIA, Speech Group, Vandoeuvre Nancy, France
chiraz.latiri@gnet.tn
kamel.smaili,caroline.lavecchia,david.langlois@loria.fr

Abstract. In this paper, we introduce two new methods dedicated to phrase-based machine translation. Both are based on mining a parallel corpus in order to find out the couples of linguistic units which are translation of each other. The presented methods do not rely on any alignment in contrast to what is done usually by the statistical machine translation community. Each of them proposes a complete translation table containing translations of single words and phrases. The first method is inspired from the well-known trigger language model while the second one is inspired from the association rules mining technique. All experiments are conducted on a large part of EUROPARL corpus and highlight the utility of both proposed approaches.

Keywords: Statistical machine translation, Sequence mining, Inter-lingual triggers, Inter-lingual association rules, Bilingual corpora.

1 Introduction

Since the apparition of the pioneering work of IBM researchers [2], almost all the proposed papers in Statistical Machine Translation (SMT) are based on their formalism. This is due to the strength of the approach and the availability of tools, such as GIZA++ for producing the translation table [12], CMU or SRILM for developing language model [15] and PHARAO [4] or MOSES [6] for decoding. These tools make developing SMT a very easy process.

In this paper, we would like to show that it is possible to investigate other issues which could constitute an alternative to IBM methods and their generalization to support phrase-based models [7]. The proposed methods do not rely on any alignment in contrast to what is done usually by the SMT community. The first method is inspired from the well-known trigger language model which we adapted to automatically learn word and phrase equivalents from bilingual corpora. The second one is inspired from the association rules mining technique, well-known in data mining. We adapt this latter to make it supporting two different languages.

The remainder of the paper is organized as follows: Section 2 recalls the basic foundations of statistical machine translation. We devote Section 3 to present the machine translation approach based on inter-lingual triggers. Section 4 introduces the second one based on inter-lingual association rules. Then, in Section 5, we present results of the mixture of the two above methods. The conclusion and future works are presented in Section 6.

2 Principle of Statistical Machine Translation

In SMT framework, the translation process comes back essentially to the search for the most probable sentence f in the target language given a sentence e in the source language. Let $e = e_1, .., e_j$ be the source sentence (*i.e.,* to be translated) and $f = f_1, .., f_i$ be the sentence generated by the translation system, namely:

$$\hat{f} = \arg\max_f P(f|e) \tag{1}$$

By using Bayes, we obtain:

$$\hat{f} = \arg\max_f P(f)P(e|f) \tag{2}$$

In Equation (2), $P(f)$ is estimated by a *language model*. Its role is to propose a sentence supposed to be correct in the target language. $P(e|f)$ is computed from a *translation model* and is supposed to reflect the truthfulness of the translation. Then, the decoder such PHARAO [4] or MOSES [6] generates the best hypothesis by making a compromise between, at least, these probability distributions.

3 Statistical Machine translation using inter-lingual triggers

The concept of *triggers* has been largely used in statistical language modeling [14, 16]. Roughly speaking, a statistical language model yields a probability to each potential sequence of words belonging to a vocabulary. A trigger model enhances the probability of a list of words which are correlated to a word w_i. To develop such a model, all the correlated words are retrieved. Triggers are determined by computing *mutual information* between two linguistic units x and y, each of them takes its values in the list of words belonging to the vocabulary V. Given two words x, y, the correlation $MI(x, y)$ is given by:

$$MI(x, y) = P(x, y)log_2 \frac{P(x, y)}{P(x)P(y)} \tag{3}$$

$P(x, y)$ is the joint probability between x and y, while $P(x)$ and $P(y)$ are the marginal probabilities of x and y, respectively.

3.1 Inter-lingual triggers

In [9], the concept of triggers is adapted to handle relationships between words for any two different languages. This approach is called *inter-lingual triggers*. An inter-lingual trigger is henceforth a set composed of a word (or a phrase) f in a source language, and its corresponding best correlated words (or phrases) in a target language $e_1, e_2, \dots, e_n$. This will be written as:

$$Trig(f) \longrightarrow e_1, e_2, \dots, e_n \tag{4}$$

In inter-lingual triggers f takes its values from the source vocabulary (French) and e_i from the target one (English). The translation table is obtained by assigning to each inter-lingual trigger a probability calculated as follows:

$$\forall\, f, e_i \in Trig(f), P(e_i|f) = \frac{MI(e_i, f)}{\sum\limits_{e_j \in Trig(f)} MI(e_j, f)} \tag{5}$$

where $Trig(f)$ is the set of k English linguistic units triggered by the French unit f.

In [8, 9], a word-based machine translation using inter-lingual triggers is detailed and results are presented. This approach is extended in the following to achieve a phrase-based machine translation approach.

3.2 A New Approach to achieve phrase-based MT

Since more than ten years, researches showed that the use of phrases in translation instead of words leads to better SMT system quality. In order to retrieve phrases, several approaches have been proposed in the literature. Most of them require word-based alignments. For instance, Och *et al.* in [11] collect all phrase pairs that are consistent with the word alignment provided by Brown's models.

In this paper, we will show how to take advantage from inter-lingual triggers and how to make machine translation supporting phrases generated by triggers without any alignment. A sequence of n French words can trigger a sequence of m English words with $n, m \in \mathbb{N}$. This kind of correlation is denoted by $n.m$-Trigger. In the remainder, we will detail the different steps of our approach.

Phrase extraction To retrieve from a corpus pertinent phrases, we use a method developed in [17], to rewrite the source training corpus in terms of phrases. To achieve that, an iterative process selects phrases by grouping words which have a high Mutual Information value. Only phrases improving the perplexity are kept for the forthcoming steps. At the end of the process, we get a list of phrases and a source corpus rewritten in terms of these discovered phrases. With this source corpus expressed with pertinent phrases, we aim to find their potential phrase translations in the target corpus by using inter-lingual triggers.

Learning phrases translation Since our method does not require any alignment, we assume that each source phrase of l words could be translated by several sequences of $l \pm \mathit{\Delta} l$ words. This means, to each source phrase, we associate $(2\mathit{\Delta} l + 1)$ sets of its k best inter-lingual triggers. Each set S_i is composed of the potential translations of i words with $i \in [l - \mathit{\Delta} l, .., l + \mathit{\Delta} l]$ with $l - \mathit{\Delta} l \geq 1$. Thus, we allow a source phrase to

	$n.m$-**Triggers**		
Source phrase	2.1	2.2	2.3
porter plainte	press	press charges	can press charges
	charges	can press	not press charges
	easy	not press	you can press

Table 1. Potential translations of the source phrase *"porter plainte"*.

Algorithm 1: Simulated Annealing (SA) algorithm.

```
1  Begin
2      Start with a high temperature $T$;
3      repeat
4          From the current temperature $T$, state $i$ and a BLEU $B_i$, randomly add a
           subset of $n.m$-Triggers into the translation table which makes the system
           moving from state $i$ to $j$ leading to a new BLEU $B_j$
5          If $B_j - B_i >= 0$ then
6              state $j$ is kept as the new current state
7          Else
8              $j$ is accepted as the new current state with a probability
               $random(P) < e^{\frac{B_i - B_j}{T}}$ with $P \in [0 \dots 1]$
9      until BLEU equilibrium with temperature T is reached;
10     Decrease the temperature and go to line 3 until the given low temperature is
       reached or until the BLEU stops increasing;
11 End
```

be translated by different target sequences of variable sizes. Table 1 shows the potential translations of the source phrase *Porter plainte*.

For the cited example, we guess that Δl is set to 1. Consequently, *"porter plainte"* could be translated by a sequence of at least one word and at most by a sequence of 3 words. In this example, we have selected 9 potential translations. Obviously, only *"press charges"* is correct. In the general case, each phrase could be translated by k potential units. That is why we propose to select those which are pertinent and discard the noisy ones.

All source phrases and their sets of inter-lingual triggers constitute the set of $n.m$-Triggers. The main challenge is how to select the best $n.m$-Triggers. In other words, what are the pertinent phrases and their translations. The choice of the best sub-set phrases is a combinatorial problem. Simulated annealing (SA) algorithm is one of the algorithms which can give a good solution to this kind of problem. To achieve that, we start with a word-based MT system based on 1.1-Triggers presented in [8]. Then, we randomly add phrases ($n.m$-Triggers) into the translation table until an optimal BLEU score is reached on a development corpus. In other words, we only keep phrases which improve the quality of translation on a development corpus. This method is summarized in Algorithm 1.

For each French unit (a word or a sequence) of l words, we select from the target training corpus its $10 \times (l \pm \Delta l)$ best inter-lingual triggers (translations). This means that if a sequence of l words is allowed to be translated by $l-2$, $l-1$, l, $l+1$, $l+2$ words, then 50 potential candidates are kept. All this inter-lingual triggers make the set of candidate phrase translations (called $n.m$-Triggers) required by the SA algorithm.

In the next section, we present another original method which uses association rules between terms in SMT.

4 Machine translation with inter-lingual association rules

The association rules mining problem has been introduced by Agrawal *et al.* [1]. The motivation for searching associations from texts is to discover correlations between terms that are occurring together as well as to look for regularities in corpora. Before presenting our method, let us give a brief review of the basic definitions related to association rule mining [1].

Definition 1. *An extraction context (or* corpus*, in our case) is a triplet* $\mathcal{K} = (\mathcal{P}, \mathcal{T}, \mathcal{R})$ *where:*

- $\mathcal{P} = \{s_1, s_2, \ldots, s_n\}$ *is a finite set of* n *distinct sentences of a corpus.*
- $\mathcal{T} = \{t_1, t_2, \cdots, t_m\}$ *is a finite set of* m *distinct terms of a corpus.*
- *Both sets* $\mathcal{T}$ *and* $\mathcal{P}$ *are linked through a binary relation* $\mathcal{R}$ *such that* $\mathcal{R} \subseteq \mathcal{P} \times \mathcal{T}$*. That is, each sentence* $s \in \mathcal{P}$ *is represented a set of terms* m *terms* $T \in \mathcal{T}$ *named termset, that occur together in the sentence.*

The support of $X \subseteq \mathcal{T}$ in $\mathcal{K}$, denoted by $Supp(X)$, is the absolute number of a randomly chosen sentences from $\mathcal{P}$ containing the termset X. A k-termset $T \in \mathcal{T}$, *i.e.,* a termset of length k, is called *frequent* if the k terms of T occur simultaneously in the corpus more than a user-defined frequency threshold denoted *minsupp*.

Definition 2. *An association rule* R *over* $\mathcal{K}$ *is an implication of the form* $R : X \Rightarrow Y$*, where* X *and* Y *are subsets of* $\mathcal{T}$*, and* $X \cap Y = \emptyset$*. The termsets* X *and* Y *are, respectively, called the premise and the conclusion parts of* R*.*

The *support* of a rule R and its *confidence* are defined as:

$$Supp(R) = Supp(X \cup Y) \qquad\qquad Conf(R) = \frac{Supp(X \cup Y)}{Supp(X)} \tag{6}$$

An association rule R is *valid* if its confidence value is greater than or equal to a user-defined threshold, denoted *minconf*.

For word-based machine translation, we introduced in [8] the concept of *Inter-lingual association rules*, named ILAR. The potential translations of a French term f are obtained by selecting all the English terms $e_1, e_2, \ldots, e_n$ which are present in the conclusion of a inter-lingual association rule for which f is its premise.

Since we are interested in phrase-based machine translation, we investigate in the following the problem of mining frequent closed sequences from highly sized bilingual corpora. Our aim is to extend the concept of inter-lingual association rules to the context of phrase-based machine translation.

4.1 Mining frequent closed sequences for phrase-based machine translation

Our approach is inspired from an efficient sequential pattern mining algorithm, called BFSM [3]. Our choice of BFSM is argued by the fact that this latter is well adapted for handling very large corpora and, especially for low values of the support threshold. A set of frequent closed sequences is retrieved and then inter-lingual association rules are obtained from this latter as explained further.

We consider the context $\mathcal{K} = (\mathcal{P}, \mathcal{T}, \mathcal{R})$ (*cf.* Definition 1).

Definition 3. *A sequence $S = \langle t_1, \ldots, t_j, \ldots, t_n \rangle$, such that $t_k \in \mathcal{T}$ and n is its length, is a n-termset for which the position of each term in the sentence is maintained. S is called a n-sequence.*

Definition 4. *A sequence $S_\alpha = \langle a_1, a_2, \ldots, a_n \rangle$ is a sub-sequence of $S_\beta = \langle b_1, b_2, \ldots, b_m \rangle$, denoted by $S_\alpha \subseteq S_\beta$, if there is a set of indices $1 \leq i_1 \leq i_2 \leq \ldots \leq i_n \leq m$, such that $a_1 = b_{i_1}, a_2 = b_{i_2}, \ldots, a_n = b_{i_n}$. S_β is called a* super-sequence *of S_α.*

Definition 5. *Given S a sequence discovered from $\mathcal{K}$. The support of S is the number of sentences in $\mathcal{P}$ that contain S,* i.e., *$Supp(S) =\parallel p \in \mathcal{P}$ s.t. $S \subseteq p \parallel$. S is said to be* frequent *if and only if its support is greater than or equal to the minimum support threshold $minsupp$.*

Definition 6. *A frequent closed sequence (FCS) S is a frequent sequence that has no frequent super-sequence $S^{'}$ with the same support.*

Definition 7. *Given a sequence S, its information position, denoted by $\mathcal{POS}_S$, is a set of pairs (id_s, pos_seq), where id_s represents the rank of the sentence in the corpus and pos_seq the position of the last term of the sequence in the sentence.*

Sentence	Sequence
s_1	$S_1 : \langle text, mining, tools, for, text \rangle$
s_2	$S_2 : \langle text, mining, for, analysis \rangle$
s_3	$S_3 : \langle text, for, analysis \rangle$
s_4	$S_4 : \langle text, mining, for, analysis \rangle$

Table 2. Example of a sequences dataset.

To illustrate this concept, let us take an example. Given a dataset of sequences depicted in Table 2 and a value of the $minsupp$ threshold equals to 2. We can get the information positions of the frequent 1-sequences as shown in Table 3. We notice that the term $tools$ is pruned since its support is lower than the $minsupp$ value.

Proposition 1. *Given two sequences S_k and $S_{(k+n)}$. If $S_{(k+n)}$ is a super-sequence of S_k and if they have the same k first terms and the same support, then $S_{(k+n)}$ is called* ***a backward super-sequence*** *of S_k* [3].

1-Sequence	Information position	Support
$\langle text \rangle$	(1 , 1) (1 , 5) (2 , 1) (3 , 1) (4 , 1)	4
$\langle mining \rangle$	(1 , 2) (2 , 2) (4 , 2)	3
$\langle for \rangle$	(1 , 4) (2 , 3) (3 , 2) (4 , 3)	4
$\langle analysis \rangle$	(2 , 4) (3 , 3) (4 , 4)	3

Table 3. The frequent 1-sequences.

For instance, the sequence $\langle$*Statistical machine translation evaluation*$\rangle$ is a backward super-sequence of the sequence $\langle$*Statistical machine translation*$\rangle$ since they share the three first terms, while assuming that they have the same support. The second sequence is then considered redundant.

Thus, the set of the frequent closed sequences, denoted by $\mathcal{FCS}$ only contains non-redundant patterns, *i.e.,* those not covered by other ones of the same support (or equivalently, do not have a backward super-sequence, *cf.* Proposition 1).

An algorithm for discovering frequent closed sequences The main idea of frequent closed sequence mining is based on the principle of *sequence-extension* [3]. The sequence-extension of a k-sequence S_k adds a new term to S_k as a new last element. The frequent $(k+1)$-sequences can be produced by extending the current found frequent k-sequences. Indeed, for each frequent k-sequence S_k, we pick up each frequent 2-sequence S_α whose first term is the same as the last term of S_k and matches the information position of S_k. The result is a new frequent $(k+1)$-sequence. Algorithm 2 details how extending a frequent k-sequence by a frequent 2-sequence according to the explained process.

Algorithm 2: Sequence-Extension.

```
Input: a $k$-sequence $S_k$ and a 2-sequence $S_\alpha$.
Output: a $(k+1)$-sequence
Begin
    If $last\text{-}term(S_k) = first\text{-}term(S_\alpha) \ \wedge \ id\_s(S_k) = id\_s(S_\alpha)$
    $\wedge \ pos\_seq(S_\alpha) = pos\_seq(S_k) + 1$ then
        $S_{k+1} = \langle S_k \oplus (S_\alpha \backslash first\text{-}term(S_\alpha)) \rangle$;
    Return $S_{k+1}$;
End
```

$\oplus$ denotes the concatenation operator.

Our approach proceeds in four steps, namely:

Step 1: Extraction of frequent 1-sequences and their information positions. We scan the extracted context $\mathcal{K}$ once to record the information position (*cf.* Definition 7) of each distinct term in the corpus.

Step 2: Generation of the frequent 2-sequences. The frequent 2-sequences are produced by applying join operation on the 1-sequences as in the APRIORI-like methods [1]. Note that, we do not need to scan the corpus to count their supports. Indeed, the information positions of the frequent 1-sequences are used to get those of frequent 2-sequences.

Let us consider the two 1-sequences $\langle text \rangle$ and $\langle mining \rangle$ given in Table 3. The 2-sequence $\langle text\ mining \rangle$ is generated as follows: the pair $(1, 2)$ of $\langle mining \rangle$ is matched with the pair $(1, 1)$ of $\langle text \rangle$. By the same way, the pairs $(2, 2)$ and $(2, 1)$ are matched to form the pair $(2, 2)$, and the pairs $(4, 2)$ and $(4, 1)$ to form the pair $(4, 2)$. So, the information position of the 2-sequence $\langle text\ mining \rangle$ are the pairs $(1, 2)$, $(2, 2)$ and $(4, 2)$ (*cf.* Table 4).

Frequent 2-sequence	Information position
$\langle text\ mining \rangle$	(1 , 2) (2 , 2) (4 , 2)
$\langle text\ for \rangle$	(1 , 4) (2 , 3) (3 , 2) (4 , 3)
$\langle text\ analysis \rangle$	(2 , 4) (3 , 3) (4 , 4)
$\langle mining\ for \rangle$	(1 , 4) (2 , 3) (4 , 3)
$\langle mining\ analysis \rangle$	(2 , 4) (4 , 4)
$\langle for\ analysis \rangle$	(2 , 4) (3 , 3) (4 , 4)

Table 4. The frequent 2-sequences.

Step 3: Generation of the frequent sequences of length greater than 2 We use the frequent 2-sequences to generate frequent sequences of length greater than two. The matching is based on the sequence-extension principle as shown in Algorithm 2.

For instance, the frequent 3-sequence $\langle text\ mining\ for \rangle$ can be generated by extending $\langle text\ mining \rangle$ with the 2-sequence $\langle mining\ for \rangle$ (*cf.* Table 4).

Thus, longer frequent sequences are iteratively derived starting from the frequent 2-sequences. For our running example, only one 4-sequence is found, namely :
$\langle text\ mining\ for\ analysis \rangle$.

Step 4: Pruning step We only retain *frequent closed sequences* since we look for compact set of term sequences by pruning redundant ones. Our pruning procedure is based on the backward super-sequence condition (*cf.* Proposition 1) which is tested on each candidate k-frequent sequence. Hence, if a sequence $S_{(k+n)}$ is a backward super-sequence of another sequence S_k, *i.e.,* they have the same support, then this latter is pruned from the set $\mathcal{FCS}$. For example, the frequent 3-sequence $\langle mining\ for\ analysis \rangle$ is a backward super-sequence of the 2-sequence $\langle for\ analysis \rangle$. Therefore, the sequence $\langle for\ analysis \rangle$ is discarded from the set $\mathcal{FCS}$.

4.2 Inter-lingual association rules based on frequent closed sequences

In what follows, we describe the way to derive from $\mathcal{FCS}$ the inter-lingual association rules, named ILAR-n-to-m, for phrase-based machine translation.

While considering frequent closed sequences sets $\mathcal{S}^{fr}$ and $\mathcal{S}^{en}$, an ILAR is an implication of the form: $R : S_{fr} \Rightarrow S_{en}$ such that S_{fr} and S_{en} are two frequent closed sequences of terms of lengths n and m, respectively. In machine translation, this means that the English sequence S_{en} is a potential translation of the French sequence S_{fr}. Note that, we keep the same definitions for the support and the confidence of an ILAR as given in Equation (6).

In order to use inter-lingual association rules in statistical machine translation, we need to assign to each rule $R : S_{fr} \Rightarrow S_{en}$ a probability computed as follows:

$$\forall\, S_f \in \mathcal{S}^{fr},\ S_{e_j} \in \mathcal{S}^{en}, P(S_{e_j}|S_f) = \frac{Conf(S_f \Rightarrow S_{e_j})}{\sum\limits_{i\in[1\ldots n]} Conf(S_f \Longrightarrow S_{e_i})} \qquad (7)$$

We present in the next section the experimental evaluation of the two approaches described above in the context of phrase-based machine translation.

5 Experimental study

All experiments are carried out on a part of the proceedings of the European Parliament EUROPARL [5]. The proposed models have been tested in a whole translation decoding system by using PHARAO decoder [4] and then compared to the performance of state-of-the-art both for word and phrase-based machine translation [10]. We use the BLEU (BiLingual Evaluation Understudy) score [13] for evaluation.

5.1 Material

We used a French-English parallel corpus of 596831 sentence pairs. Table 5 gives more details about the used parallel corpus EUROPARL.

		French	**English**
Train	Sentences	$596K$	
	Words	$17.3M$	$15.8M$
	Singletons	$26.6K$	$22.2K$
	Vocabulary	$77.5K$	$60.3K$
Development	Sentences	1444	
	Words	$15.0K$	$14.0K$
Test	Sentences	500	
	Words	$5.2K$	$4.9K$

Table 5. Quantitative description of the used corpus.

Experimental results The direction of translation is from English to French. Tests have been achieved on a corpus of 500 sentences. The phrase translation table of the state-of-the-art system, denoted in the remainder by Koehn-Och, is acquired from a word-aligned parallel corpus by extracting all phrase-pairs that are consistent with the word alignment [10].

Method based on inter-lingual triggers retrieves from the source language a set of 11 212 pertinent phrases which are composed of two or three words, only 8.31% of these phrases occur in the test corpus. This percentage is very low in order to hope to noticeably improve the results.

Table 6 illustrates performances of different systems on both development and test corpora. On the development corpus, the use of pertinent $n.m$-Triggers improved the results achieved by 1.1-Triggers by 13.27%. For the state-of-the-art methods, the use of phrases increases the performance by 19.90% compared to the word-based method. On the test corpus, both methods improve the results by respectively 5% and 25%. This difference may be explained by the fact that the state-of-art method uses a translation table of more than 21 millions of entries (of one or more words) whereas ours uses only 5.2 millions (where the number of phrases do not exceed 20 000 phrases). Consequently, our translation table has a weak coverage of the training corpus. We should thus increase the size of the translation table in order to get closer values to those of the state-of-art one. By adding 1.71 millions of phrases, we achieve a BLEU result of 34.41. This improvement reduces the gap between our results and the state-of-art one to 2.74, knowing that our translation table is very small in comparison to the one used by Koehn-Och.

	Inter-lingual triggers		ILAR		State of the art	
	1.1	n.m	1-to-1	n-to-m	IBM model 3	Koehn-Och
Development	31.02	35.27	19.71	32.66	29.23	35.07
Test	30.97	32.75	22.06	34.18	29.57	37.15

Table 6. Experimental evaluation in terms of the BLEU score on the development and the test corpora.

As illustrated in Table 6, the phrase-based MT system based on ILARs fulfilled a BLEU score of 34.18. So, by adding pertinent frequent closed sequences, we achieved an improvement of more than 12 points in terms of BLEU compared to the initial word-based MT system which only consider ILAR between single terms (22.06). Note that, we do not restrict the length of the generated frequent closed sequences, although derived sequences from EUROPARL may reach a size of 25 terms. We also experimentally observed that beyond a certain length (sequences of 14 terms), the BLEU score does not increase. This could be explained by the fact that these sequences are very rare in the training corpus, *i.e.,* they have a very low support, and their frequency in the test corpus is even lower. Therefore, they can not improve the BLEU score.

6 Table translation mixture

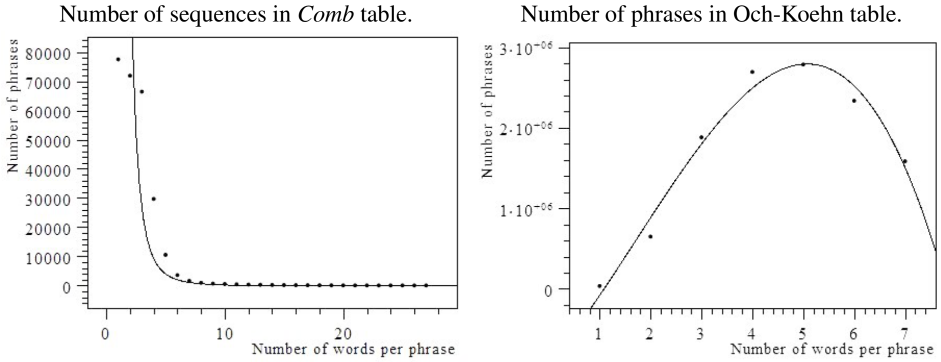
In order to take advantage of the two original methods presented above, we decided to combine their two translation tables. Because the $n.m$-Triggers table does not contain a great number of phrases and since 1.1-Triggers achieved better results than IBM 3 model, we decided to put in a new translation table the word translation got from 1.1-Triggers and phrases obtained by association rules, *i.e.,* ILAR-n-to-m. The result is presented in Table 7. The combination (the line referenced by *Comb*) outperforms both proposed methods. This result shows that we come closed to the result of Koehn-Och. We are just 1.63 below of the standard method. This illustrates that it is possible to retrieve pertinent phrases and their corresponding translations without any need of alignment which constitutes the advantage of the two original methods presented in this paper.

Moreover, Figure 1 illustrates the evolution of the number of phrases in accordance to the number of words per phrase. We can see that the Koehn-Och curve has a cubic form [3] whereas

[3] Koehn-Och method: $y = -28953X^3 + 148491X^2 + 720014X - 90643$

Model	**BLEU**
Koehn-Och	37.15
ILAR-n-to-m	34.18
$n.m$Triggers	34.41
$Comb$	35.52

Table 7. Evaluation in terms of BLEU score on test corpus.


Number of sequences in *Comb* table.

Number of phrases in Och-Koehn table.

Fig. 1. Evolution of the number of phrases in *Comb* table and Och-Koehn one.

ours has a decreasing power aspect [4]. The Koehn-Och curve grows until 5 and decreases for longer phrases. Whereas in our method the curve sharply decreases with the length of phrases.

Consequently, the gap between our result and the Koehn-Och one could be explained by the fact that the influence of phrases of four and five words would have a real impact. While the number of sequences of 5 words in Koehn-Och method is around 2.8 millions, in our table we got just 10500. Globally, our method produces pertinent phrases, however this number is not sufficient to reach the state-of-art result.

7 Conclusion

In this paper, we proposed two new phrase-based machine translation methods. Each of them proposes a complete translation table containing translations of single words and phrases. The advantage of these methods is their easiness to develop statistical machine translation and more important than that, the fact that our methods, in contrast to what is done by the community, do not need any alignment.

We experimented our approaches on a large part of EUROPARL English-French language pair with a vocabulary of more than 60 000 linguistic units, and we evaluated them by using BLEU measure. Obviously, our methods have been compared to the pioneer ones. The results presented here are very encouraging. They show that it is possible to consider the issue of statistical machine translation differently with the aim to improve the literature results. The advantage of our methods

[4] Our method: $y = 1.36E6X^{-3.6}$

is that the selected phrases are pertinent but their number is not huge. In the future, we plan to improve our methods by making them selecting more sequences without losing the quality of translations.

References

1. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference, Washington, DC, USA*, pages 207–216, May 1993.
2. P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
3. K. Y. Chang. Efficient sequential pattern mining by breadth-first approach. Master thesis, Available at http://web.management.ntu.edu.tw/chinese/im/theses/r92/r91725010.pdf, National Taiwan University, Accessed on september 21, 2010., 2004.
4. P. Koehn. PHARAOH: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of Meeting of the American Association for Machine Translation (AMTA)*, pages 115–124, 2004.
5. P. Koehn. EUROPARL: A multilingual corpus for evaluation of machine translation. In *Proceeding on the MT Summit, Thailand*, 2005.
6. P. Koehn and al. MOSES: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*, 2007.
7. P. Kohen, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedingd of the Human Language Technology and North American Association for Computaional Linguistics Conference*, pages 48–54, Edmonton, May-June 2003.
8. C. Latiri, K. Smaili, C. Lavecchia, and D. Langlois. Mining monolingual and bilingual corpora. *Intelligent Data Analysis (IDA)*, 14(6), 2010.
9. C. Lavecchia, K. Smaili, D. Langlois, and J. P. Haton. Using inter-lingual triggers for machine translation. In *Proceedings of the eighth Conference Interspeech 2007, Antwerp, Belgium*, August 2007.
10. F. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In *'Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, PA, USA, July 2002.
11. F. J. Och. An efficient method for determining bilingual word classes. In *Proceedings of EACL, Bergen*, pages 71–76, 1999.
12. F. J. Och and H. Ney. Improved statistical alignment models. In *Association of Computational Linguistics*, pages 440–447, Hongkong, China, October 2000.
13. K. Papineni, S. Roukos, T. Ward, and W-J. 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, 2001.
14. R. Rosenfeld. *Adaptive statistical language modeling: a maximum entropy approach.* PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1994.
15. R. Rosenfeld. The CMU statistical language modeling toolkit and its use in the 1994 arpa csr evaluation. In *Proceeding of the Spoken Language Systems Technology Workshop*, pages 47–50, Austin, 1995.
16. C. Tillmann and H. Ney. *Selection criteria for word trigger pairs in language modeling*, volume 1147, pages 98–106. LNAI, Springer Verlag, 1996.
17. I. Zitouni, K. Smaili, and J. P. Haton. Statistical language modeling based on variable length sequences. *Computer Speech and Language*, 17(4-5):27–41, 2003.