# Discovering Phrases in Machine Translation by Simulated Annealing

Caroline Lavecchia, David Langlois, Kamel Smaïli

# Discovering Phrases in Machine Translation by Simulated Annealing

*Caroline Lavecchia[1,2], David Langlois[1,3], Kamel Smaïli [1,2]*

[1] Loria, Speech Group, Vandoeuvre, France
[2] University Nancy2
[3] IUFM de Lorraine
{lavecchi, langlois, smaili}@loria.fr

## Abstract

In this paper, we propose a new phrase-based translation model based on inter-lingual triggers. The originality of our method is double. First we identify common source. Then we use inter-lingual triggers in order to retrieve their translations. Furthermore, we consider the way of extracting phrase translations as an optimization issue. For that we use simulated annealing algorithm to find out the best phrase translations among all those determined by inter-lingual triggers. The best phrases are those which improve the translation quality in terms of Bleu score. Tests are achieved on the proceedings of the European Parliament corpora. The training is made on a corpus containing 596K parallel sentences (French-English) and tests on a corpus of 1444 sentences. With only $8.1\%$ of the identified source phrases occurring in the test corpus, our system overcomes the baseline model by almost 3 points.

**Index Terms**: statistical machine translation, inter-lingual triggers, simulated annealing

## 1. Introduction

The goal of Machine Translation (MT) is to look for a sentence in a *target* language given a sentence in a *source* language. Different approaches exist to deal with this difficult challenge. Some approaches require *a priori* human knowledge in order to model each of the source and target languages, and how to switch from one to another. The Systran MT system [1] is based on this approach and proposes a translation model depending on transfer rules. The statistical approach follows a completely different direction.

A statistical MT does not require any external knowledge, it uses only parallel corpora to model the translation process. Such corpora are aligned at word or sentence level in order to link both source and target languages. The translation issue is treated as an optimization problem. Translating a sentence from English into French involves finding the best French target sentence $f^*$ which maximizes the probability of $f$ given the English source sentence $e$. The Bayes rule allows to formulate the probability $P(f|e)$ as follows:

$$f^* = \underset{f}{argmax} P(f|e) = \underset{f}{argmax} P(e|f)P(f) \qquad (1)$$

Thus, the translation process consists in a language model $P(f)$ and a translation model $P(e|f)$. Language model parameters are trained on a target corpus, whereas parameters of the translation model are determined from the parallel corpus. Then, a decoder provides the best target sentence given the source sentence and the table translation parameters.

First statistical MT systems were word-based [2]. Obviously, the human translation is a very complex process which is not only word based. Following this fact, recent researches showed that the use of phrase translation instead of word translation leads to better MT system quality. Dealing with phrases allows an explicit modeling of lexical units and captures easily local reordering. For example, without use of phrases, the translation of *Tour Eiffel* gives *Tower Eiffel*, then reordering process may produce the correct English translation. By using phrases, we reduce the imprecision of translation and avoid some reordering problems.

In order to retrieve phrases, several approaches have been proposed in literature. Most of them require word-based alignments. For example, Och in [3] 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 on the condition that words are aligned with each other. Retrieved phrases are not based on linguistic knowledge, consequently they could lead to noisy sequence of words.

In this paper, we propose an original idea based on inter-lingual triggers to build phrase translation tables without requiring word-based alignments.

First we give an overview of inter-lingual triggers. Then we present our method for learning phrase translations. Finally, we integrate and test our approach into a entire translation process. We end with a conclusion which 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 [4]. A trigger is a set composed of a word and its best correlated triggered words in terms of mutual information (MI).

Since classical triggers allow to establish a triggering-triggered relationship between two events from the same language, we propose to determine correlations between words coming from two different languages. These triggers will be qualified by inter-lingual triggers. Each of them is composed by a triggering source event and its best correlated triggered target events. With this, we hope to find among the set of triggered target events, potential translations of the triggering source event. Inter-lingual triggers are determined on a parallel corpus according to mutual information formula:

$$MI(f, e) = P(f, e) log \frac{P(f, e)}{P(f)P(e)} \qquad (2)$$

where $f$ and $e$ are respectively French and English events. Where $P(f, e)$ is the joint probabilities and $P(e)$ and $P(f)$ are

marginal probabilities.

For each French event $f$, we kept its $k$ best English triggered events. In the following, an event is a word or a sequence of words. Thus a sequence of $n$ French words can trigger a sequence of $m$ English words with $n, m \in \mathbb{N}$. This kind of triggers are called *n-To-m* triggers.

Inter-lingual triggers have been used in [5] to enrich resource deficient languages from those which are considered as potentially important. Our purpose is to use them in order to carry out statistical phrase-based machine translation. In the following, we propose an original method based on *n-To-m* inter-lingual triggers to build phrase translation tables.

# 3. A new method for learning phrase translations

Most of methods which use phrases on MT require word-based alignments. For example, Och in [3] collected all phrase pairs that were consistent with the word alignment. In his method, any contiguous source words must be a potential translation of any contiguous target words provided that words are aligned with each other.

We are convinced that if we succeed in identifying common phrases in the source part of the training corpus, inter-lingual triggers will allow to retrieve its translations in the target part. This would generate less noise. Since source phrases are selected beforehand, our method does not require any word alignment.

## 3.1. Phrase extraction

In the few last years we developed a statistical method to extract pertinent phrases [6] from corpus. We use this method 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 value of Mutual Information. Only the phrases which improve 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 phrases. With this source corpus expressed with pertinent phrases, we hope to find their potential phrase translation in the target corpus by using inter-lingual triggers.

## 3.2. Learning phrase translation

In the following, we assume that each source phrase of $l$ words could be translated by several sequences of $l \pm \Delta l$ words. At this step, no word alignment is performed. For this reason, we associate to each source phrase $(2\Delta l + 1)$ sets of its $k$ best inter-lingual triggers. In each set, inter-lingual triggers are sequences of $j$[1] target words. Thus, we allow a source phrase to be translated by different target sequences of variable sizes. Table 1 shows the potential translations of the source phrase *Aide financière*. As this source phrase is a sequence of 2 words, we assume that it could be translated by a sequence of one, two or three target words. For this reason, we associate it with its best *2-To-1*, *2-To-2* and *2-To-3* inter-lingual triggers. Among all the potential translations, we have then to choose the most pertinent and to remove those which could upset the translation system.

All source phrases and their sets of inter-lingual triggers constitute the set of *n-To-m* inter-lingual triggers. Now, the issue is how to select the best *n-To-m* inter-lingual triggers? In other words, what are the pertinent phrases and their transla-

---

[1] $j$ taking its value in $l \pm \Delta l$

| Source phrase | n-To-m triggers | | |
| | *2-To-1* | *2-To-2* | *2-To-3* |
| --- | --- | --- | --- |
| Aide financière | Financial | Financial aid | Financial aid to |
| | Aid | Financial assistance | Financial assistance to |
| | Assistance | Financial support | Financial aid for |

Table 1: Potential translations of the source phrase *Aide financière*

tions. To answer this question, we first evaluate in terms of Bleu score the quality of the automatic translations using a decoder and a word translation table based on *1-To-1* triggers. The translation table is obtained by assigning to each inter-lingual trigger a probability calculated as follows:

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

where $Trig(f)$ is the set of $k$ English events triggered by the French event $f$. The model based on *1-To-1* inter-lingual trigger will constitute our baseline model. Then we add phrases (*n-To-m* inter-lingual triggers) to the initial translation table. This will be done in accordance to what follows:

---

**Algorithm 1** Method for learning and selecting the best phrase translations

1: Extract phrases from the source corpus.
2: Determine the *n-To-m* inter-lingual triggers which allow to associate each source phrase with its best target phrases of variable size.
3: Compute the baseline Bleu score by using *1-To-1* inter-lingual triggers.
4: Select an optimal subset of *n-To-m* inter-lingual triggers on an iterative process handled by Simulated Annealing algorithm.

---

## 3.3. Simulated Annealing tuning

Simulated Annealing (SA) algorithm is a technique applied to find a good solution to an optimization problem that becomes unmanageable using combinatorial methods. The set *n-To-m* triggers constitute a list of candidate phrase translations. We have to integrate a subset of this list in the translation table in order to improve the quality of translation. Naturally, it is unreasonable to try all possible combinations of translation. That is why we decided to use simulated annealing.

The entire algorithm is given by Algorithm 2.

At the end of the SA algorithm, only the *n-To-m* inter-lingual triggers which allow to improve the performance of the initial word-based MT system are selected. Since the SA algorithm makes the system moving from one state to another with the objective to improve the energy, consequently this method achieves the phrases which necessary increase BLEU.

## 3.4. Corpora

In this section we evaluate our phrase-based system based on inter-lingual triggers. The experiments presented below have

**Algorithm 2** Simulated Annealing algorithm
1: Start with a high temperature $T$.
2: With a temperature $T$ and until the equilibrium is reached do

From the current temperature $T$ of the system and from the current state $i$ which has an Energy $E_i$, perturb the system which makes it moving from state $i$ to $j$. The energy of state $j$ is $E_j$.

If $E_j - E_i >= 0$ then state $j$ is accepted as the current state; Otherwise, state $j$ is accepted with a probability $random(P) < e(Ei - Ej)/T$ with $P \in [0 - 1]$
3: Decrease the temperature and go to step 3 until the given low temperature is reached or until the energy stops increasing

---

been conducted on the proceedings of the European Parliament [7]. We used the French-English parallel corpus. Table 2 shows the parallel corpus statistics of our experiments. We use a train corpus to extract French phrases and to compute inter-lingual triggers. The best translation of phrases are selected by using inter-lingual triggers on the test corpus.

|       |            | French | English |
|-------|------------|--------|---------|
| Train | Sentences  | 596K   |         |
|       | Words      | 17, 3M | 15.8M   |
|       | Singletons | 26.6K  | 22.2K   |
|       | Vocabulary | 60.3K  | 77.5K   |
| Test  | Sentences  | 1444   |         |
|       | Words      | 15.0K  | 14.0K   |

Table 2: An overview of the experiment material

Our approach leads to significant inter-lingual triggers as shown in Table 3.

| French      | English     | $MI \times 10^{-4}$ |
|-------------|-------------|---------------------|
| croix rouge | red cross   | 5.08                |
|             | cross       | 4.53                |
|             | red         | 4.26                |
| abaisser    | lower       | 1.51                |
|             | reduce      | 0.75                |
|             | bring down  | 0.43                |
| bonjour     | good morning| 0.92                |
|             | morning     | 0.44                |
|             | good        | 0.30                |
| date butoir | deadline    | 1.77                |
|             | date        | 1.07                |
|             | the deadline| 0.59                |
| excusez moi | sorry       | 2.68                |
|             | excuse me   | 2.27                |
|             | forgive me  | 1.32                |

Table 3: Examples of English phrases triggered by French phrases

The first column presents French sequences of one or several words. Sequences that have more than one word are automatically picked up by the iterative process explained in section (3.1). For each French sequence, the second column refers to the best correlated English sequences of one, two or three

words in terms of MI. And finally the third column shows the MI value associated with each inter-lingual trigger. A qualitative analysis showed that our method leads to pertinent inter-lingual triggers. Thus, triggered sequences could often be considered as potential translation of the triggering French sequence. Furthermore, inter-lingual triggers allow to retrieve synonyms as it is shown for the French word *abaisser* which can be translated by *lower* or *bring down*. Note also that they take into account the fact that $n$ French words are not necessarily translated into $n$ English words but rather $m$ words. Thus, *bonjour* is associated with *good morning* or *excusez moi* with *sorry*. Finally, when a French sequence is translated into several English words, inter-lingual triggers favor the whole English sequence rather than subparts of it. This is illustrated by the example of *croix rouge* which is associated with *red cross* with a higher MI value than *cross* or *red*.

In the following section, we design and evaluate our PBMT system based on inter-lingual triggers. To achieve that, we use *n-To-m* inter-lingual triggers and simulated annealing algorithm to build a phrase translation table required by the decoder Pharaoh [8] in order to translate an English source corpus into French. Bleu score allows us to evaluate the quality of the obtained French translation.

### 3.5. Experiments

#### 3.5.1. French extracted phrases

We extracted from the French part of the training corpus, a set of 11212 phrases which are composed of two or three words. Only $8.31\%$ of the phrases extracted from the training corpus occur in the test corpus.

#### 3.5.2. Selecting the best phrase translations

For each French linguistic unit (word or phrase) of size $l$, we kept for further steps the 100 best potential translations of size $l \pm \Delta l$. These translations are actually triggers which constitute the set of candidates (called *n-To-m* triggers) required by SA algorithm.

#### 3.5.3. Initial configuration for SA algorithm

As an initial configuration of our system, we employ *1-To-1* triggers for which each French word is associated with a list of $k$ English words. Different experiments showed that 100 is the optimal value for the parameter $k$. From this we calculate the initial translation table required by Pharaoh decoder. For an optimal use of the decoder, the weights of the models involved in the decoding process are tuned [2]. The target language model is a trigram model (smoothed by Good-Turing method). This leads to a baseline Bleu score of 31.02 as shown in Table 4. The third line of this table reminds the results obtained by the third IBM model on the same corpus. Note, that the value of the parameters of the two models are different. This is due to the fact that parameters are optimized on the same corpus but with two different translation tables.

#### 3.5.4. Results with Simulated Annealing

Different experiments have been conducted to optimize the parameters of the SA algorithm. In this section, we present the

---

[2]$tm$ (respectively $lm$, $d$) indicate the weight of the translation (respectively target language, distortion) model, the parameter $w$ is the word penalty

| Initial conf. | tm | lm | d | w | Bleu |
|---|---|---|---|---|---|
| *1-To-1 Triggers* | 0.9 | 0.8 | 0.4 | −3 | **31.02** |
| IBM | 0.6 | 0.7 | 0.4 | −1 | **29.23** |

Table 4: Baseline Bleu score associated with the initial configuration of our PBMT system

performance with the optimal set of parameters. The suitable value for the initial temperature is $T = 10^{-4}$. The initial configuration was the MT system based on *1-To-1* triggers, this leads to an initial energy of 31.02 (Bleu score). Then, at each step of the algorithm, we agitate the configuration by adding randomly phrase translations from the retrieved set of *n-To-m* triggers. Each perturbation is simulated by inserting 10 phrases and by keeping, for each of them, the 5 best translations. Equilibrium at a fixed temperature is reached when the energy stops increasing.

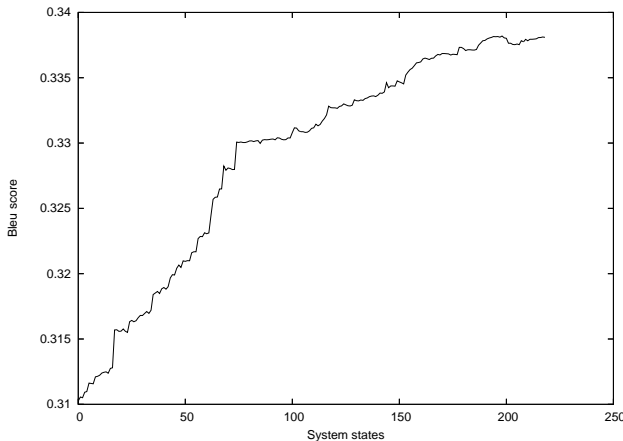The evolution of the Bleu score is shown in Figure 1. At the



Figure 1: Evolution of the Bleu score

end of the SA process, our phrase-based MT system fulfilled a Bleu score of 33.93. In other words, by adding pertinent phrase translations, we achieved an improvement of more than 2 point in terms of Bleu compared to the initial word-based MT system.

## 4. Conclusion

In this paper, we presented a phrase-based Machine Translation system based on inter-lingual triggers. This approach allows to associate a triggering source phrase with its best triggered target phrases. Thus, we used inter-lingual triggers in order to construct our PBMT (Phrase-Based Machine Translation) system.

In order to extract relevant phrase translations and therefore improve the translation quality, we proposed an original method that does not need any word alignment. First we identified common source phrases by an iterative process. Then, we retrieved their potential translations by using inter-lingual triggers. Finally, we used simulated annealing algorithm to select the best phrase translations among all those determined by inter-lingual triggers.

We trained and tested our PBMT system on the Parliament European corpus. Once phrase translation table were induced by inter-lingual triggers and simulated annealing algorithm, we used Pharaoh to translate text from English into French. Knowing that the test corpus contains only 8.1% of phrases, results showed that our approach overcomes the baseline model (*1-To-1*) by almost 3 points. With a corpus containing more phrases we will probably achieve better results. Preliminary comparison with a state of the art PBMT showed that our approach achieves better results on a sub-title corpus. Tests are underwork to optimize the state of the art method parameters in order to compare it with our approach on the European Parliament corpus.

Our results are very encouraging and efforts are done to improve our model. The idea of using inter-lingual triggers seems to be very attractive. Up till now, we focus on word surface forms, in future work we will consider inter-lingual triggers on syntactic features. We hope then to integrate more linguistic knowledge in the translation process.

## 5. Acknowledgments

## 6. References

[1] T. V. Jean Senellart, Pter Dienes, "New generation systran translation system," in *MT Summit VIII*, Santiago de Compostela, Spain, September 2001.

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

[3] F. J. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in *the joint conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, University of Maryland, College Park, MD, 1999, pp. 20–28.

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

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

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

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

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