# Improved Vietnamese-French Parallel Corpus Mining Using English Language

*Do Thi Ngoc Diep*[1,2], *Laurent Besacier*[1], *Eric Castelli*[2]

(1) LIG Laboratory, CNRS/UMR-5217, Grenoble, France
(2) MICA Center, CNRS/UMI-2954, Hanoi, Vietnam
thi-ngoc-diep.do@imag.fr

## Abstract

This paper improves our unsupervised method for extracting parallel sentence pairs from a comparable corpus presented in [1]. In this former paper, a translation system was used to mine a comparable corpus and to detect French-Vietnamese parallel sentence pairs. An iterative process was implemented to increase the number of extracted parallel sentence pairs which improved the overall quality of the translation.

This paper validates the unsupervised approach on a new under-resourced language pair (Vietnamese-English) and it also addresses the problem of using triangulation through a third language to improve the parallel data mining process. An extension of the unsupervised method is proposed to make use of triangulation. Two ways to include the additional data from triangulation are carried out. The experiments conducted on Vietnamese – French show that using triangulation through English can improve the quality of the extracted data and slightly improve the quality of the translation system measured with BLEU.

## 1. Introduction

Over the past fifty years of development [2], machine translation (MT) has obtained good results when applied to several pairs of languages including English, French, Chinese, Japanese, etc. Many approaches for MT have been proposed, such as: rule-based (direct translation, interlingua-based, transfer-based), corpus-based (statistical, example-based) as well as hybrid approaches. However, research on statistical MT for low e-resourced languages always faces the challenge of getting enough data to support any particular approach.

Statistical machine translation (SMT) permits to construct rapidly a machine translation system. This approach requires the availability of large parallel bilingual corpora of source and target languages to build a statistical translation model for source/target languages and a statistical language model for the target language. The two models and a search module are then used to decode the best translation ([3]; [4]). Thus, a large parallel bilingual text corpus is a prerequisite. However, such a corpus is not always available, especially for low e-resourced languages.

The most common methods used to build parallel corpora consist either in automatic methods which collect parallel sentence pairs from the Web ([5]; [6]), or alignment methods which extract parallel documents/sentences from two monolingual corpora ([7]; [8], [9]). More recently, there were also increasing contributions where parallel sentence pairs were extracted from a comparable corpus ([10]; [11]; [12]). We assume that in the case of a low e-resourced language pair, even a small parallel corpus might not be available to start developing a SMT system.

In a former work [1], we have proposed a fully unsupervised method, starting with a comparable corpus, which allows us to overcome the problem of lacking parallel data. This method had been applied to mine Vietnamese – French parallel data from a comparable corpus of a daily news website.

The goal of this paper is twofold: firstly, we validate the approach presented in [1] to another language pair (Vietnamese – English); secondly, we investigate the use of triangulation through English to improve the amount (and the quality) of Vietnamese – French parallel data that can be extracted from a dedicated news web site.

The rest of the paper is organized as follows. Section 2 recalls our unsupervised method presented in [1] and its application on mining Vietnamese – French (results already published in [1]) as well as Vietnamese – English (new results) parallel data. Section 3 presents an extension of this unsupervised method using triangulation through English to improve the data mining process, as well as the experiments and results associated. The final section concludes and gives some perspectives.

## 2. A Fully Unsupervised Method to Mine Parallel Data from Noisy Parallel Corpora

### 2.1. Review of the unsupervised method

A comparable corpus contains data which are not parallel but "still closely related by conveying the same information" [10]. It may contain "non-aligned sentences that are nevertheless mostly bilingual translations of the same document" [11] or contain "various levels of parallelism, such as words, phrases, clauses, sentences, and discourses, depending on the corpora characteristics" [13].

Extracting parallel data from comparable corpora has been presented in some previous works. Zhao and Vogel [10] propose a maximum likelihood criterion which combines sentence length model and a statistical translation lexicon model extracted from an already existing aligned parallel corpus. An iterative process is applied to retrain the translation lexicon model with the extracted data. Munteanu and Marcu [12] present a method for extracting parallel sub-sentential fragments from a very non-parallel corpus. Each source language document is translated into target language using a bilingual lexicon/dictionary. The target language document which matches this translation is extracted from a collection of target language documents. A probabilistic translation lexicon based on the log likelihood-ratio is used to detect parallel fragments from this document pair. Abdul-Rauf and Schwenk [14] present a similar technique, but a proper statistical machine translation system is used instead of the bilingual dictionary, and the evaluation metric (TER) is used to decide the degree of parallelism between sentences.

These above methods can be modeled as if containing a translation phase and a filtering phase. To extract parallel data from a comparable corpus A, the source side of A is translated by a translation lexicon model or a proper statistical machine translation system (which is built from an initial parallel corpus or at least a bilingual dictionary). The translated output is then compared with the target side of the corpus A and filtered by a filtering module (using a score or an evaluation metric). These methods require at least a parallel corpus (or a bilingual dictionary) to bootstrap the system. We assume that in the case of low e-resourced languages, even a small parallel corpus, may not be available. In our former work [1], we proposed a fully unsupervised method (called Scheme 1 in this paper), where the starting point is just a comparable corpus, without using additional parallel data.

Let's assume that we have a comparable corpus *A* available. The process contains two steps: initiation step and mining step (Figure 1).

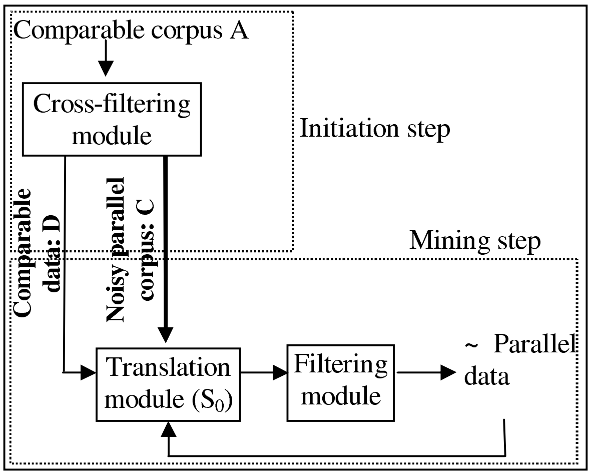

*Figure 1: Our unsupervised method [1] – Scheme 1*

In the initiation step, a cross-filtering process is applied on the corpus *A* to extract a noisy parallel (or still comparable) corpus *C* and a comparable corpus *D* (D=A\C). The process is described as follow (Figure 2):

- Split the comparable corpus A into an even number of sub-corpora (for example 4 sub-corpora $A_1$, $A_2$, $A_3$, $A_4$).
- Build different translation systems from these sub-corpora. ($A_1$ $SMT_{A1}$, $A_2$ $SMT_{A2}$, $A_3$ $SMT_{A3}$, $A_4$ $SMT_{A4}$)
- The source side of a sub-corpus (eg. $A_1$) is translated by using the translation system built from other sub-corpus (eg. $SMT_{A2}$). The translated output is then compared with the target side (eg. of $A_1$) and filtered by a filtering module. We apply the same manner for the rest pairs (eg. ($A_2$, $SMT_{A1}$), ($A_3$, $SMT_{A4}$), ($A_4$, $SMT_{A3}$)). The extracted sentence pairs ($C_1$, $C_2$, $C_3$, $C_4$) form the corpus C. The rest is treated as the corpus D.

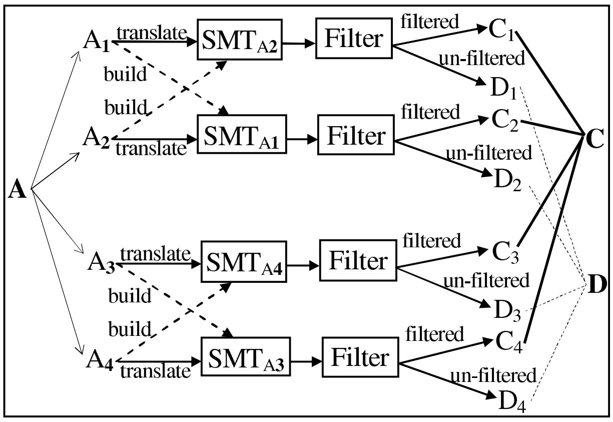

*Figure 2: Cross-filtering process in the initiation step*

Next, in the mining step, the corpus C is considered as reliable enough to build the initial translation system $S_0$. To mine the comparable corpus D, once again, the manner of translating and filtering is applied. The source side of the corpus D is translated by using translation module $S_0$. The translated output is compared with the target side of the corpus D and filtered by a filtering module.

The filtering module in both steps bases on an evaluation metric. A pair of sentences is considered as parallel if its evaluation metric is larger than a threshold. In the former work [1] some metrics like TER, BLEU, NIST and a proposed score PER* were investigated. The modified position-independent word error rate PER* is calculated based on the similarity, while the PER [15] measures an error (the difference of words occurring in hypotheses and reference).

$$\text{PER*} = \frac{\text{2 * number of identical words}}{\text{(length of hypothesis + length of reference)}}$$

According to experiments in [1], PER* threshold =0.3 achieved the best performance on filtering the parallel sentence pairs for pair of languages English – French. So we use PER* in our filtering module.

Since the starting point of our process is not a clean parallel corpus, but a noisy parallel corpus (used to build $S_0$), an iterative scheme is used. The extracted sentence pairs are added to the system $S_0$ to create a new translation system $S_1$ and so on. The iterative process re-translates the source side using this new translation system, re-calculates the evaluation metric and then re-filters the parallel sentence pairs. It was shown in [1] that each iteration not only increases the number of extracted parallel sentence pairs but also improves the quality of the overall translation system.

## 2.2. Application on mining a comparable corpus of a multilingual news website

### *2.2.1. The news website*

Vietnamese is the 14th most widely-used language in the world; however research on MT for Vietnamese is rare. The earliest MT system for Vietnamese is the system from the Logos Corporation, developed as an English-Vietnamese system for translating aircraft manuals during the 1970s [2]. Until now, in Vietnam, there have been only few research groups working on MT [16].

We focus on mining a bilingual news corpus from the Web and building a Vietnamese-French statistical machine translation system. The unsupervised method described in the previous section was applied to mine a text corpus of a multilingual daily news website, the Vietnam News Agency[1] (VNA). This website contains articles written in four languages (Vietnamese, English, French, and Spanish) and divided in 9 categories. This kind of corpus is a truly comparable corpus because it tends to contain parallel sentences or rough translations of sentences on the same topics. To date, we have obtained 20,884 French documents, 54,406 Vietnamese documents and 32,795 English documents. Each document contains, on average, 10 sentences, with around 30 words per sentence.

### *2.2.2. Vietnamese - French sentence pair extraction*

From the comparable corpus VNA, the number of possible Vietnamese – French parallel document pairs was reduced using a publishing date filter. Then each sentence in a Vietnamese document was merged with all sentences in the possible French document. So a pair of one Vietnamese document (containing *m* sentences) and one French document (containing *n* sentences) produced *m x n* pairs of sentences. From the VNA corpus, we obtained a comparable corpus of 1,442,448 Vietnamese – French sentence pairs, which is really noisy parallel. We kept only those pairs where the ratio of the French sentence's length to the Vietnamese sentence's length was between 0.8 and 1.3. This produced a comparable corpus of 345,254 sentence pairs (named $A_{vn\text{-}fr}$). After the cross-filtering process (with PER* threshold=0.45 to ensure the reliability of extracted sentence pairs and an acceptable number of pairs of C to build $S_0$ system), we obtained the corpus $C_{vn\text{-}fr}$ containing 4076 sentence pairs, and the corpus $D_{vn\text{-}fr}$ containing the remaining 341,178 sentence pairs.

The translation modules in this paper were built using the Moses toolkit with the default settings:

- GIZA++ was used for word alignments, the "-*alignment*" option for phrase extraction was "*grow-diag-final-and*"
- 14 features in total were used in the log-linear model: distortion probabilities (6 features), one tri-gram language model probability, bidirectional translation probabilities (2 features) and lexicon weights (2 features), a phrase penalty, a word penalty and a distortion distance penalty.
- A 3-gram target language model was built using the SRILM toolkit.

The unsupervised mining method was applied on the corpus $C_{vn\text{-}fr}$ and $D_{vn\text{-}fr}$. The quality of the translation systems was also evaluated on a test set of 400 manually extracted Vietnamese-French parallel sentence pairs. The Vietnamese sentences were initially segmented into

[1] http://www.vnagency.com.vn/

syllables (no word segmentation pre-processing was applied). Each Vietnamese sentence had only one French reference.

The number of extracted sentence pairs and the SMT system's evaluation score after each iteration are reported in Figure 3. (The results reported here are different from those in [1] because in this version we did not consider the stop words to calculate the PER* used for filtering).

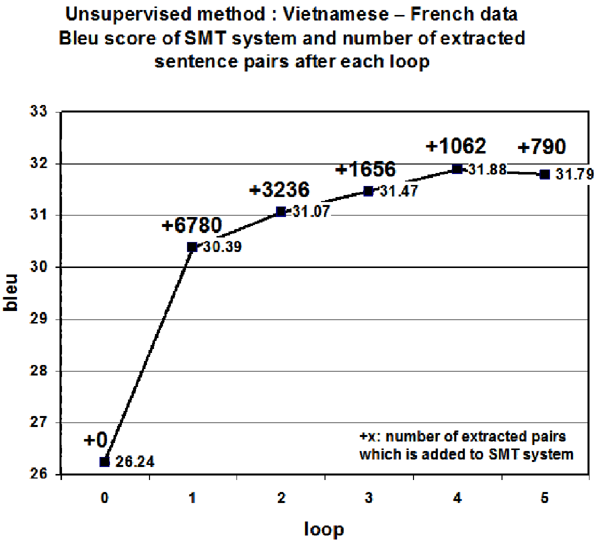

*Figure 3: Mining Vietnamese – French comparable data. BLEU score of SMT system after each loop in Unsupervised method – Scheme 1: D = 341,178 sentence pairs; C = 4076 sentence pairs*

Each iteration brings us a number of extracted sentence pairs. The quality of the translation system increases in the first few iterations and decreases after that. This may be explained by the fact that, in the first iterations, a lot of new parallel sentence pairs are extracted and included to the translation model. However, in subsequent iterations, as the amount of truly parallel sentences decreases, more wrong sentence pairs are added to the system so the quality of the translation system is reduced. However, the quality of the translation system built by extracted data from this unsupervised method is comparable with that of another method which requires better quality data for bootstrapping (bilingual dictionary, etc.) (see more in [1]).

### 2.2.3. Vietnamese - English sentence pair extraction

In the former work, we presented the mining process on Vietnamese – French comparable corpus. In this paper, we first validate this approach for mining a Vietnamese – English comparable corpus. The same mining process was applied on a comparable corpus of 479,865 Vietnamese – English sentence pairs (named $A_{vn\text{-}en}$). After cross-filtering process, we obtained the corpus $C_{vn\text{-}en}$ containing 9407 sentence pairs, and the corpus $D_{vn\text{-}en}$ containing 470,458 sentence pairs.

The quality of the translation systems was evaluated on a test set of 400 manually extracted Vietnamese-English parallel sentence pairs. The number of extracted sentence pairs and the evaluation scores after each iteration are reported in Figure 4. The results in this case are similar to those in the case of Vietnamese – French extraction. The quality of the translation system increases in the first few iterations and decreases after that. However, the usefulness of the iterative process is less clear in that case since the result after iteration 1 is already high.

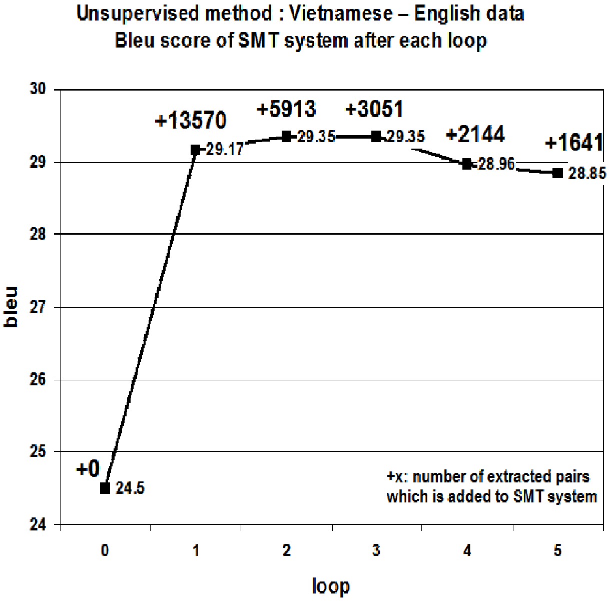

*Figure 4: Mining Vietnamese – English comparable data. BLEU score of SMT system after each loop in Unsupervised method – Scheme 1: D = 470,458 sentence pairs; C = 9407 sentence pairs*

## 3. Using Triangulation through English – an Extension of the Unsupervised Mining Method

### 3.1. Using triangulation through English

Using triangulation through a third language has been proposed in the NLP domain, including machine translation. A language for triangulation, or sometimes also called a bridge language, is an artificial or a natural language used as an intermediary language for translation between multiple different languages. Using triangulation can help when parallel data for a given language pair are lacking. In machine translation, to translate between pair of languages S and T, two independent machine translation systems for pair of languages S – P and P – T are used. These systems can be concatenated or synthesized together [17]. The phrases-tables of the two systems can be merged or interpolated to create a new one for the pair of languages S and T ([18]; [19]). Finally, one or more triangulation languages can be used ([20]; [21]).

In this paper, we want to address the problem of using triangulation to improve the parallel data mining process. Multilingual websites appear more and more and comparable corpora of more than two languages are then also available. One question that we wanted to answer was whether using additional data (for example Vietnamese – English data) can improve the data mining process for another pair of languages (Vietnamese – French).

The unsupervised mining process for a comparable corpus of language pair S – T (Vietnamese-French in our case) was already described in Section 2.1. The same way, a parallel corpus for language pair S – P ($X_{S\text{-}P}$ Vietnamese-English in our case) can be extracted from the multilingual news website. Moreover, for a well-resourced language pair P-T (English-French in our case), it is easy to find or to build a translation system ($SMT_{P\text{-}T}$). So we want to make use of the corpus $X_{S\text{-}P}$ as well as the $SMT_{P\text{-}T}$ system in the mining process to improve the extraction from a comparable corpus $D_{S\text{-}T}$.

The extension mining process is summarized in Figure 5. The data in language P of the parallel corpus $X_{S\text{–}P}$ is translated to language T by using a translation system $SMT_{P\text{-}T}$. Then the new (and probably noisy) $X_{S\text{-}T}$ data obtained is added to the unsupervised mining process (in the mining step). This additional data $X_{S\text{-}T}$ can be added either to the $S_0$ system (Scheme 2), or to the comparable corpus $D_{S\text{-}T}$ (Scheme 2*). Then, unsupervised mining process is applied as usual.

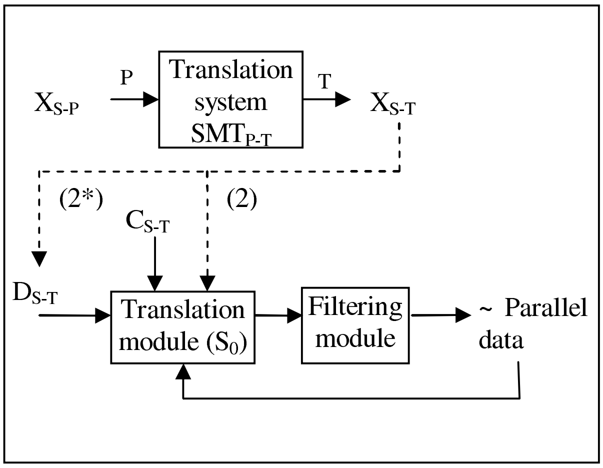

*Figure 5: Extension of the unsupervised method, using a triangulation language*

### 3.2. Experiments using triangulation through English

The two experiments for Scheme 2 and Scheme 2* in Figure 5 were carried out on the VNA web site comparable data. We tried to improve the mining of Vietnamese (S) – French (T) comparable corpus, using parallel Vietnamese – English (P) data.

The statistical machine translation system for English – French $SMT_{P\text{-}T}$ could be any existing commercial system like Systran but we decided to build our own system from the *Europarl* and *News* corpora that are provided for different evaluation campaigns (WMT, IWSLT2010). This translation system was evaluated on the test set provided for the translation task of the WMT 2009. The BLEU score of the system $SMT_{en\text{-}fr}$ obtained was 23.74.

#### *3.2.1. Vietnamese – English data preparation*

The Vietnamese – English parallel data is the extracted data in Section 2.2.3. The English side of this extracted corpus was translated to French using the system $SMT_{en\text{-}fr}$. ($X_{vn\text{-}en}$ was transformed to $X_{vn\text{-}fr}$). Obviously, the corpus $X_{vn\text{-}fr}$ is a very noisy parallel corpus due to the mining and translation errors. So, to ensure the quality of the added data $X_{vn\text{-}fr}$, one has to be more selective. The threshold for PER* score in the unsupervised mining process was changed from 0.3 to 0.5. For each threshold, the $X_{vn\text{-}fr}$ data obtained was added to the initial corpus $C_{vn\text{-}fr}$ of 4076 Vietnamese – French pair of sentences (obtained in section 2.2.2) to build a new translation module $S_0$ (as in the Scheme 1) and the quality of the new translation module $S_0$ was evaluated on the same 400-sentence pair test set.

From Table 1, we can see that the use of triangulation through English can improve the performance of the Vietnamese – French SMT system (baseline corresponds to iteration 0 of our unsupervised process). The 0.4 threshold selected 8218 additional Vietnamese – English sentence pairs and significantly improved the BLEU score of the new $S_0$.

| **PER* Threshold** | **#pairs in $X_{vn\text{-}fr}$** | **#pairs in $S_0$** | **BLEU of $S_0$** |
|---|---:|---:|---:|
| 0.3 | 47.433 | 51.509 | 24.99 |
| 0.35 | 17.159 | 21.235 | 27.66 |
| 0.4 | **8.218** | 12.294 | **28.53** |
| 0.45 | 3.586 | 7.662 | 28.16 |
| 0.5 | 1721 | 5.797 | 26.94 |
| Baseline (No additional $X_{vn\text{-}fr}$ data) | 0 | 4.076 | 26.24 |

*Table 1: Vietnamese – English data preparation*

### 3.2.2. *Experiment for the Scheme 2*

Now (from Section 2.2.2), the corpus $C_{vn\text{-}fr}$ contains 4076 sentence pairs, and the mining corpus $D_{vn\text{-}fr}$ contains 341,178 sentence pairs. From Section 3.2.1, the corpus chosen $X_{vn\text{-}fr}$ contains 8218 sentence pairs. As in the Scheme 2, the corpus $X_{vn\text{-}fr}$ was added to the corpus $C_{vn\text{-}fr}$ to build the initial translation module $S_0$. The iterative mining process was then carried out. The number of extracted data is presented in Table 2, which is compared to that of the unsupervised process – Scheme 1 (section 2.2.2).

| **Loop** | **Scheme 1** | **Scheme 2** | **Loop** | **Scheme 1** | **Scheme 2** |
|---|---|---|---|---|---|
| **0** | 6780 | 6798 | **6** | 460 | 478 |
| **1** | 3236 | 3094 | **7** | 409 | 417 |
| **2** | 1656 | 1596 | **8** | 392 | 335 |
| **3** | 1062 | 1087 | **9** | 324 | 309 |
| **4** | 790 | 765 | **10** | 239 | 282 |
| **5** | 576 | 532 | | | |

*Table 2: The number of extracted data in mining process – Scheme 2 ($S_0$: C=4076+ X=8218) and – Scheme 1 ($S_0$: C=4076)*

The quality of the SMT system after each loop in Scheme 2 is presented as the small dotted line in Figure 6. The quality of the SMT system in Scheme 1 is presented as the large dotted line in the same figure. Recall that the training data for SMT system in Scheme 2 contains $C_{vn\text{-}fr}$, $X_{vn\text{-}fr}$ and the extracted data after each loop, while the training data in the Scheme 1 does not contain $X_{vn\text{-}fr}$.

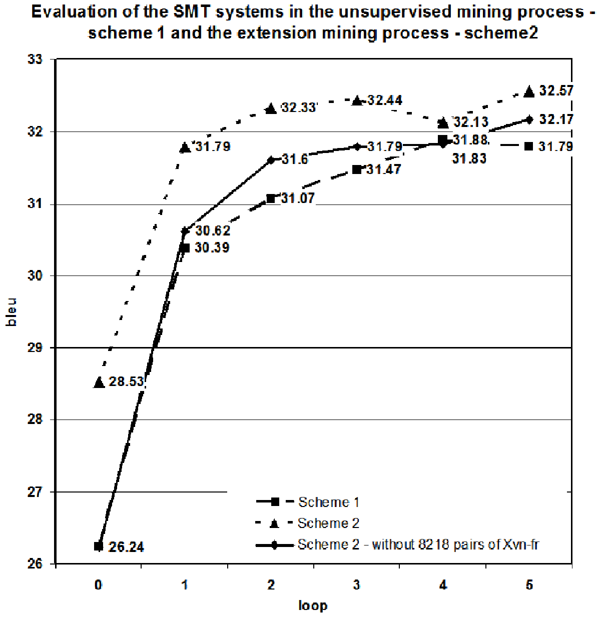

*Figure 6: Evaluation of the SMT systems built in the unsupervised mining process – Scheme 1 and the extension mining process – Scheme 2*

For a deeper analysis, the qualities of only extracted data from both schemes were evaluated. In the Scheme 2, new SMT systems were built based on the $C_{vn\text{-}fr}$ and the extracted data only (that means we removed $X_{vn\text{-}fr}$ from the training data at each loop). The BLEU scores are estimated on the same test set of 400 sentence pairs. They are presented as the continuous line in Figure 6.

Comparing the large dotted line (Scheme 1) and the continuous line (Scheme 2 – without $X_{vn\text{-}fr}$), we can see that, although the number of extracted sentence pairs in two processes is comparable, the quality of the extracted sentence pairs for Scheme 2 is a little higher than that from the baseline Scheme 1, thanks to the additional data of Vietnamese – English.

### 3.2.3. *Experiment for the Scheme 2**

We also carried out the experiment on the Scheme 2* of the extension method (see Figure 5). 8218 sentence pairs of the $X_{vn\text{-}fr}$ were added to the mining corpus $D_{vn\text{-}fr}$ of 341,178 sentence pairs instead of $C_{vn\text{-}fr}$. The translation module was built by the $C_{vn\text{-}fr}$ corpus only. Figure 7 presents the number of extracted sentence pairs from the Scheme 2 and the Scheme 2*.

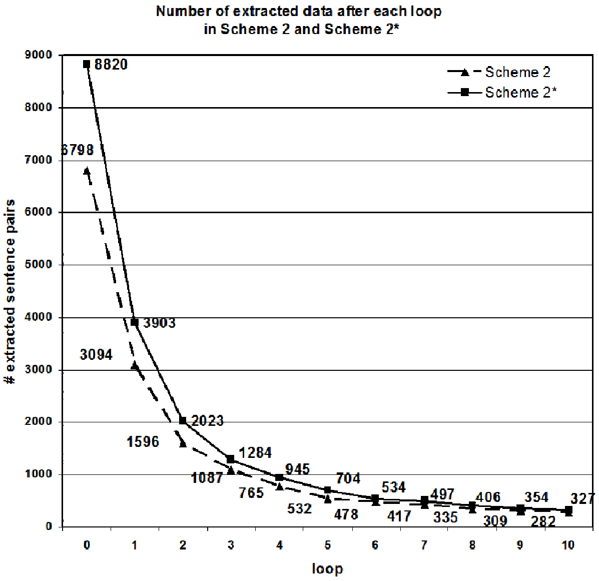

*Figure 7: Extraction result of the Scheme 2 (D: 341,178; $S_0$: C=4076 + X=8218) and the Scheme 2* (D: 341,178 + X=8218; $S_0$: C=4076 )*

Figure 8 presents the BLEU scores estimated on the same test set of 400 sentence pairs of the SMT systems after each loop. The dotted line presents results of the Scheme 2, and the continuous line presents results of the Scheme 2*. The quality of the SMT systems of the Scheme 2 is better than that of the Scheme 2* in the first few iterations because of adding

8218 sentence pairs to the $S_0$. However, in the Scheme 2*, these 8218 sentence pairs were re-filtered through the iterations, so the quality of the SMT systems in Scheme 2* increased in the last iterations. And the max BLEU score was reached at the loop 9 of the Scheme 2*.

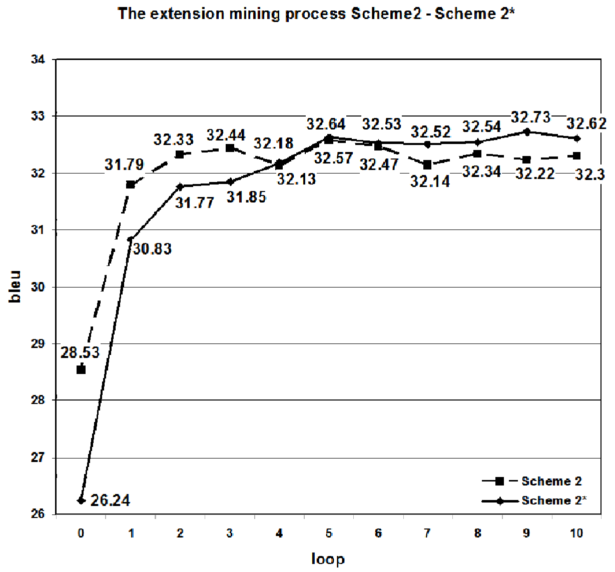

*Figure 8: BLEU scores of the SMT systems after each loop for extension mining process - Scheme 2 and Scheme 2**

## 4. Conclusion and Perspectives

This paper improved our unsupervised method for extracting parallel sentence pairs from a comparable corpus presented in [1]. This paper validated the unsupervised approach on a new under-resourced language pair (Vietnamese-English) and it also addressed the problem of using triangulation through a third language to improve the parallel data mining process. An extension of our unsupervised method was proposed to make use of triangulation.

The unsupervised mining process was applied to the Vietnamese – French and Vietnamese – English language pairs. The results obtained have shown that this method may be applied successfully even in those cases where parallel data are lacking.

As far as triangulation is concerned, the parallel data of a pair of language S – P was used to improve the mining data process for pair of language S – T. The data in language P was translated to language T by using a translation system $SMT_{P\text{-}T}$ (P-T being a well-resourced language pair). Then the translated output and the data in language S of the corpus S – P were added to the unsupervised mining process. Two ways to combine these additional data with the mining process were carried out and both have shown improved results compared to the baseline without triangulation.

Our future works will focus on deeper analysis of the best filtering and data inclusion techniques, on experiments at a larger scale and on human evaluations to confirm improvements obtained with our proposed method.

## 5. References

[1] Do, T.N.D, L. Besacier, E. Castelli, “A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora”, European Association for Machine Translation (EAMT 2010), Saint-Raphael (France), June 2010.

[2] Hutchins, W.J., “Machine translation over fifty years”, *Histoire, epistemologie, langage. ISSN 0750-8069,* 2001.

[3] Brown, P.F., S.A.D. Pietra, V.J.D. Pietra and R.L. Mercer, “The mathematics of statistical machine translation: parameter estimation”, *Computational Linguistics. Vol. 19, no. 2,* 1993.

[4] Koehn, P., F.J. Och and D. Marcu, “Statistical phrase-based translation”, *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Vol. 1,* 2003.

[5] Resnik, P. and N.A. Smith, “The Web as a parallel corpus”, *Computational Linguistics,* 2003.

[6] Kilgarriff, A. and G. Grefenstette, “Introduction to the special issue on the Web as corpus”, *Computational Linguistics, volume 29*, 2003.

[7] Koehn, P., “Europarl: a parallel corpus for statistical machine translation”, *Machine Translation Summit*. 2005.

[8] Gale, W.A. and K.W. Church, “A program for aligning sentences in bilingual corpora”, *Proceedings of the 29th annual meeting on Association for Computational Linguistics*. 1993.

[9] Patry, A. and P. Langlais, « Paradocs: un système d’identification automatique de documents parallèles », 12e Conference sur le Traitement Automatique des Langues Naturelles, 2005.

[10] Zhao B., S. Vogel, “Adaptive parallel sentences mining from Web bilingual news collection”, *International Conference on Data Mining*. 2002.

[11] Fung, P., P Cheung, “Mining very-non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and EM”, *Conference on Empirical Methods on Natural Language Processing*. 2004.

[12] Munteanu, D.S. and D. Marcu, “Extracting parallel sub-sentential fragments from non-parallel corpora” *44th annual meeting of the Association for Computational Linguistics*. 2006.

[13] Kumano, T., H. Tanaka, T. Tokunaga, “Extracting phrasal alignments from comparable corpora by using joint probability SMT model”. *Conference on Theoretical and Methodological Issues in Machine Translation*. 2007.

[14] Abdul-Rauf, S. and H. Schwenk, “On the use of comparable corpora to improve SMT performance”,

*Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. 2009.

[15] Tillmann C., S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf, “Accelerated DP based search for statistical translation”, *In 5th European Conf. on Speech Communication and Technology*. 1997.

[16] Ho, T.B, “Current status of machine translation research in vietnam, towards asian wide multi language machine translation project”, *Vietnamese Language and Speech Processing Workshop*. 2005.

[17] Gispert A., J. B. Marino, “Catalan-English Statistical Machine Translation without parallel Corpus: Bridging through Spanish”, *LREC Fifth International Conference on Language Resources and Evaluation*, 2006.

[18] Utiyama M. and H. Isahara, “A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation”, *NAACL-HLT*, 2007.

[19] Wu H., H. Wang, “Pivot Language Approach for Phrase-Based Statistical Machine Translation”, *ACL*, 2007.

[20] Eisele A., “Parallel Corpora and Phrase-Based Statistical Machine Translation for New Language Pairs via Multiple Intermediaries”, *LREC Fifth International Conference on Language Resources and Evaluation*, 2006

[21] Koehn P., A. Birch, R. Steinberger, “462 Machine Translation Systems for Europe”, *Machine Translation Sumit*, 2009.