A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora

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 presents an unsupervised method for extracting parallel sentence pairs from a comparable corpus. A translation system is used to mine the comparable corpus and to detect parallel sentence pairs. An iterative process is implemented not only to increase the number of extracted parallel sentence pairs but also to improve the overall quality of the translation system. A comparison between this unsupervised method and a semi-supervised method is also presented. The unsupervised method was tested in a hard condition: no available parallel corpus to bootstrap the process and the comparable corpus contained up to 50% of non parallel data. The experiments conducted show that the unsupervised method can be really applied in the case of lacking parallel data. While preliminary experiments are conducted on French-English translation, this unsupervised method is also applied successfully to a low e-resourced language pair (French-Vietnamese).

1 Introduction

Over the past fifty years of development (Hutchins, 2001), machine translation (MT) has obtained good results when applied to several pairs of languages such as English, French, Italia, 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) uses statistical method based on 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 target language. The two models and a search module are then used to decode the best translation (Brown et al, 1993; Koehn et al, 2003). 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 to build parallel corpora consist in automatic methods which collect parallel sentence pairs from the Web (Resnik and Smith, 2003; Kilgarriff and Grefenstette, 2003), or alignment methods which extract parallel documents/sentences from two monolingual corpora (Koehn, 2005; Gale and Church, 1993, Patry and Langlais, 2005). There is also the method of extracting parallel sentence pairs from a comparable corpus (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2006). Abdul-Rauf and Schwenk (2009) present a semi-supervised extracting method requiring an initial parallel corpus in order to build a first SMT system that will be used during the semi-supervised extraction (see more in section 2.1). 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. So, does a fully unsupervised method, starting with a highly noisy parallel corpus, allow to solve the problem of lacking parallel data?

Firstly, it is important to note that we consider that “comparable” and “noisy parallel” have equivalent meanings in the context of our work, since a “noisy parallel” corpus can be extracted from a “comparable” corpus using a minimal information retrieval component (based on basic

© 2010 European Association for Machine Translation.

features like publishing date, sentence length, etc.). Advanced IR approaches for mining comparable corpora are outside of the scope of this paper whose goal is exactly to get rid of complex IR approaches by using an iterative process based on SMT.

This paper presents a fully unsupervised extracting method, which is compared to a semi-supervised extracting method. The first results show that the unsupervised method can be really applied in the case of lacking parallel data. The rest of the paper is organized as follows. Section 2 describes the methods of extracting parallel sentence pairs from a noisy parallel corpus: semi-supervised method and fully unsupervised method. Section 3 presents our experiments and our results on testing the unsupervised method. The next section presents an application of this method for a real low e-resourced language pair: Vietnamese-French. The last section concludes and gives some perspectives.

2 Mining parallel data from comparable corpora

2.1 Extracting methods

A comparable corpus contains data which are not parallel but "still closely related by conveying the same information" (Zhao and Vogel, 2002). It may contain "non-aligned sentences that are nevertheless mostly bilingual translations of the same document" (Fung and Cheung, 2004) or contain "various levels of parallelism, such as words, phrases, clauses, sentences, and discourses, depending on the corpora characteristics" (Kumano et al., 2007).

Extracting parallel data from comparable corpus has been presented in some previous works. Zhao and Vogel (2002) 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 (2006) 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 (2009) 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. Sarikaya et al. (2009) introduce an iterative bootstrapping approach in which the extracted sentence pairs are then added to the initial parallel corpus to rebuild the SMT system. All these methods are presented as effective methods to extracting parallel fragments/sentences from a comparable corpus.

2.2 Semi-supervised v/s Unsupervised learning method

These above methods can be modeled as figure 1a, with a translation phase and a filtering phase (with or without iterations). The source side of a comparable corpus D is translated by using translation module S_0 (a translation lexicon model or a proper statistical machine translation system). The translated output is then compared with the target side of the corpus D and filtered by filtering module (using a score or an evaluation metric). These methods can be considered as semi-supervised methods which require an initial parallel corpus C_1 (or at least a bilingual dictionary) to build the translation module. We assume that in the case of low e-resourced languages, this parallel corpus, even small, may not be available. So, we try to propose a fully unsupervised method, here, where the starting point is just a simple noisy comparable corpus, without using additional parallel data.

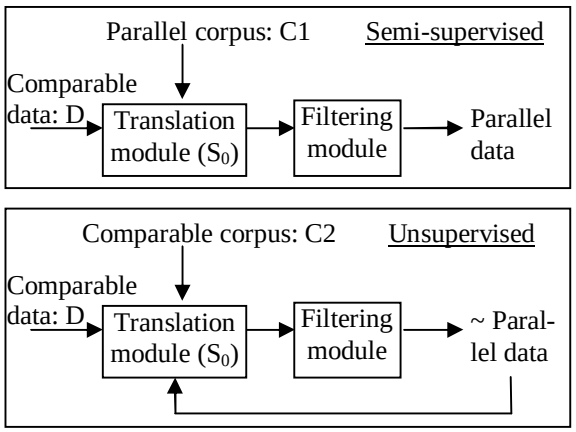

Figure 1. Semi-supervised v/s unsupervised methods.

In the unsupervised learning scheme (figure 1b), the translation module S_0 is built based on another comparable corpus C_2 and the iterative scheme is recommended. One of the challenges of this work is to see if such a different starting point (noisy comparable corpus, versus truly par-

allel corpus) can still lead to the design of an extracting system and also improve the quality of the overall translation system.

In our research, we focus on mining the parallel sentence pairs. The translation module S_0 is a statistical machine translation system, and filtering module bases on evaluation metric estimated for each sentence pair. Several evaluation metrics are used to determine which one is the most suitable: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006) and a modified PER* (see details in section 3.3). A pair is considered as parallel if its evaluation metric is larger (for BLEU, NIST, PER* metrics) or smaller (for TER metric) than a threshold.

The extracted sentence pairs are then combined with the system S_0 in several ways to create a new translation module. An iterative process is performed which re-translates the source side by this new translation system, re-calculates the evaluation metric and then re-filters the parallel sentence pairs. We hope that each iteration not only increases the number of extracted parallel sentence pairs but also improves the quality of the translation system.

The extracted parallel data are re-used in different combinations:

- W1: the translation system at step i is re-trained on a training corpus consisting of C2 and E_{i-1} (the extracted data from the last iteration); E_0 being the data extracted when translation system is trained on C2 only (S_0).
- W2: the translation system at step i is re-trained on training corpus consisting of C2 and $E_0+E_1+\ldots+E_{i-1}$ (the extracted data from the previous iterations).
- W3: at iteration i, a new separate phrase-table is built based on the extracted data E_{i-1}. The translation system decodes using both phrase-table of S_0 and this new one (log-linear model) without weighting them.
- W4: the same combination as W3, but the phrase-table of S_0 and the new one are weighted, e.g. 1:2.

3 Preliminary experiments for French-English SMT

In this section, we present experiments on unsupervised method, in comparison with those on semi-supervised method. Two systems were built, one based on semi-supervised method (Sys1), another based on unsupervised method (Sys2).

3.1 Data preparation

We chose French-English languages for these preliminary experiments. A noisy parallel corpus was "simulated" by gathering parallel and non-parallel sentence pairs in order to control the precision and the recall of the extracting method. The *correct parallel sentence pairs* were taken from the Europarl corpus, version 3 (Koehn, 2005). A significant number of wrong sentence pairs were added in the data (about 50%).

To make it comparable with the real case treated in section 4 (a low e-resourced language pair), the size of data was chosen small for this preliminary setup. The corpus C1 contains only 50K correct parallel sentence pairs. The corpus C2 contains 25K correct parallel sentence pairs (withdrawn from C1) and 25K wrong sentence pairs. The corpus D, the input data for extracting process, was built from 10K correct parallel sentence pairs and 10K wrong sentence pairs, which were different from sentence pairs of C1 and C2. The correct and the wrong sentence pairs of D were marked to calculate the precision and the recall later.

3.2 System construction

Both systems Sys1 and Sys2 were constructed using the Moses toolkit (Koehn et al., 2007). This toolkit contains all of components needed to train the translation model. It also contains tools for tuning these models using minimum error rate training and for evaluating the translation result using the BLEU score. We used the default settings in Moses:

- GIZA++ (Och and Ney 2003) 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 (Stolcke, 2002).

The target (English) language model was built from the English part of the entire Europarl corpus. The baseline translation models were built from corpus C1 and C2 respectively.

3.3 Starting with parallel or comparable corpus?

One question that we want to answer first is whether the translation system based on a noisy parallel corpus can be used to filter the input data like the translation system based on parallel corpus does. To examine this problem, the French side of corpus D was translated by Sys1 and Sys2. Then, the translated outputs were compared with the English side of the corpus D. Four evaluation metrics were used for this comparison: BLEU, NIST, TER and PER*. Our modified position-independent word error rate (PER*) is calculated based on the similarity, while the PER (Tillmann et al., 1997) measures an error (the difference of words occurring in hypotheses and reference).

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

The distributions of evaluation scores for correct parallel sentence pairs and wrong sentence pairs were calculated and presented in figure 2.

From these distributions, we can make the following comments: first, the distributions of scores have the same shape between Sys1 and Sys2. Especially, the distributions of scores for the wrong pairs were nearly identical in both systems. So, a noisy parallel corpus can replace a parallel corpus for constructing an initial translation system. Remember that the initial corpus here contains up to 50% non-parallel sentence pairs. Another important result is that the PER*, a simple and easily calculated score, can be considered as the best score to filter the correct parallel sentence pairs (while TER gave poor result for our experimental setup). Table 1 presents the precision and the recall of filtering parallel sentence pairs from both systems: Sys1 and Sys2.

Sys1 – semi-supervised method					
Filtered by	*Found*	*Correct*	*Precision*	*Recall*	*F1-score*
Bleu=0.1	6908	6892	99.76	68.92	81.52
Nist=0.4	8350	8347	99.96	83.47	90.97
Per=0.3*	10342	9785	94.61	97.85	96.20
Per=0.4*	9390	9333	99.39	93.33	96.27
Sys2 – unsupervised method					
Filtered by	*Found*	*Correct*	*Precision*	*Recall*	*F1-score*
Bleu=0.1	6233	6218	99.75	62.18	76.61
Nist=0.4	7110	7108	99.97	71.08	83.08
Per=0.3*	10110	9468	93.65	94.68	94.16
Per=0.4*	8682	8629	99.38	86.29	92.37

Table 1. Precision and recall of filtering parallel sentence pairs (given 10K correct pairs).

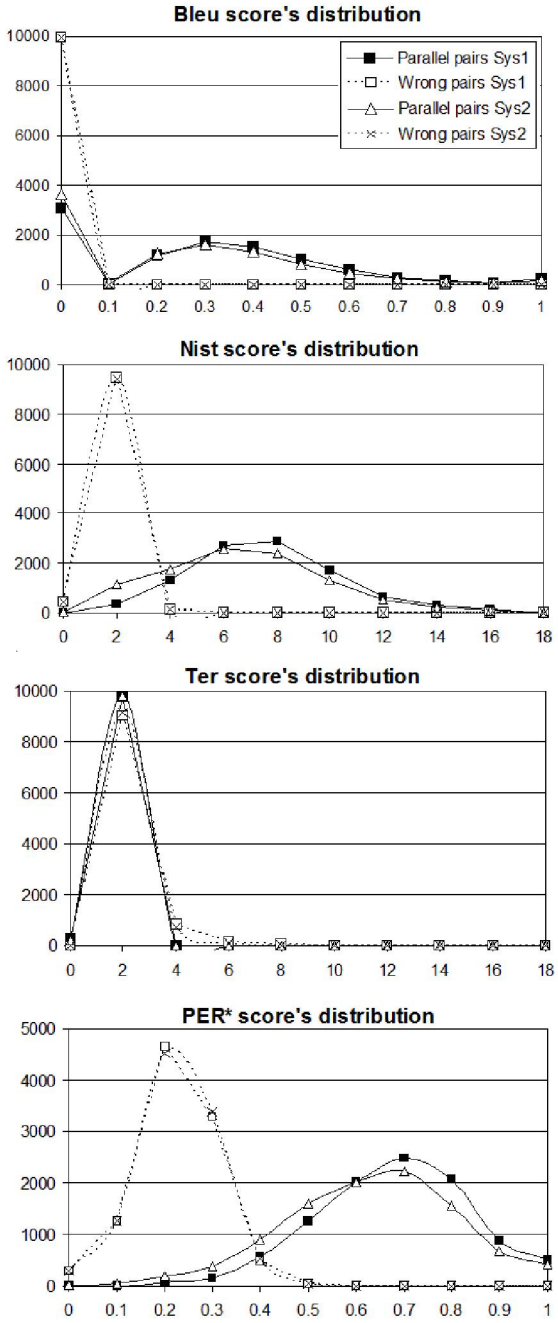

Figure 2. Score distributions for semi-supervised (Sys1) and unsupervised (Sys2) methods.

3.4 The iterations of the unsupervised method

Section 3.3 has shown that translation system based on a noisy parallel corpus can be used to filter parallel data from another corpus. However the result of filtering in Sys2 is lower than that in Sys1 (for example, the number of correct extracted sentence pairs is reduced (table1)). So, we propose, in this section, an iterative process in order to improve the quality of the translation system, and then to increase the number of correctly extracted sentence pairs.

Increasing the number of correct extracted sentence pairs: In Sys2, the extracted sentence pairs were combined with the baseline system S_0 in four ways (as mentioned in section 2.2). In order to receive the maximum number of correct extracted sentence pairs, for all iterations we chose the evaluation score PER* with the threshold=0.3, which gave the maximum recall=94.68% for the baseline system.

Figure 3 presents the number of correctly extracted sentence pairs after 6 iterations for four different combinations: W1, W2, W3 and W4. The number of correct extracted pairs was increased in all cases; however the combination W2 brought the largest number of correct extracted sentence pairs.

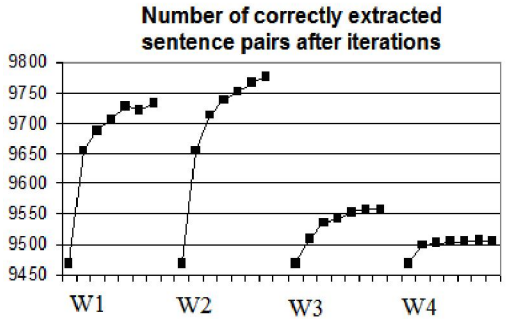

Figure 3. Number of correctly extracted sentence pairs after 6 iterations for four different combinations.

Increasing the precision and the recall of the filtering process: The precision and the recall of these four combinations are presented in figure 4. Because the filtering process focused on extracting the largest number of correct extracted sentence pairs, the precision was decreased. However, using the combination W2, the recall after 6 iterations (97.77) nearly reached the recall of the semi-supervised system Sys1 (97.85).

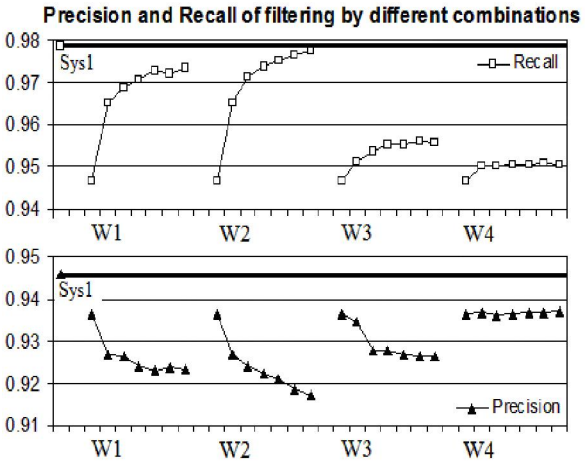

Figure 4. Precision and recall of filtering using different combinations.

Translation system evaluation: The quality of the translation systems was also evaluated. A test set containing 400 French-English parallel sentence pairs was extracted from Europarl corpus. Each French sentence had only one English reference. The quality was reported in BLEU and TER. Figure 5 gives the evaluation scores for the systems after each iteration.

The translation system evaluation revealed an important result. The quality of the translation system was increased quickly during some first iterations, but decreased after that. It can be explained by the fact that, in the first iterations, a lot of new parallel sentence pairs were extracted and included to the translation model. However, in the next iterations, when the precision of the extracting process was decreased, more wrong sentence pairs were added to the system so the translation model got worse and the quality of the translation system was reduced.

In fact, Sarikaya et al. (2009) presents a similar system using a different evaluation metric for filtering (Bleu), and use a combination similar to our W2 type. However, their research does not provide a full explanation about why they choose Bleu and this combination method, and further more, the problem of decreasing the quality of translation system after several iterations is not mentioned.

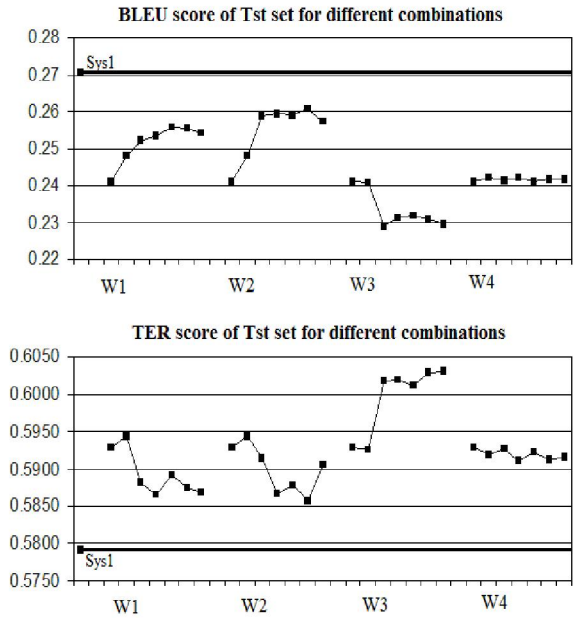

Figure 5. Translation system evaluations.

After about 3 iterations, the Bleu score can increase of about 2 points. Note that there is no tuning for the statistical models (no development data set was used for this experimental setup).

4 Application for Vietnamese - French language pair

Vietnamese is the 14th 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 (Hutchins, 2001). Until now, in Vietnam, there are only four research groups working on MT (Ho, 2005).

We focus on mining a bilingual news corpus from the Web and building a Vietnamese-French statistical machine translation (SMT) system. In a former paper (Do et al., 2009), we have presented a mining method (named *Method1*) based on publication date, special words and sentence alignment result. Firstly, possible parallel document pairs are filtered by using publishing date and special words (numbers, attached symbols, named entities). Secondly, sentences in a possible parallel document pair are aligned using Champollion toolkit (Ma, 2006), which uses lexical information (lexemes, stop words, a bilingual dictionary, etc.). Finally, parallel sentences pairs are extracted based on the sentence alignment information, which combines document length information and lexical information.

This method was applied to mine a text corpus from a Vietnamese daily news website, the Vietnam News Agency[1] (VNA) (containing 20,884 French documents and 54,406 Vietnamese documents). This corpus used is a really comparable corpus because it tends to contain parallel sentences or rough translations of sentences on the same topics. 50,322 parallel sentence pairs were extracted using *Method1*. A SMT system for Vietnamese-French was then built using the Moses toolkit with the same default settings as described in section 3.2.

In this paper, the proposed unsupervised method was applied on the same corpus VNA. Instead of aligning sentences and filtering sentence alignment information, we create a comparable corpus and apply the proposed unsupervised method to extract parallel sentence pairs. Then we compare the unsupervised method with the *Method1*.

4.1 Preparing the data

Firstly, from the comparable corpus VNA, the number of possible parallel document pairs was reduced by using 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 corpus VNA, we obtained a comparable corpus of 1,442,448 pairs of sentences, which is really noisy parallel. We just kept the pairs with the ratio of French sentence's length to Vietnamese sentence's length between 0.8 and 1.3. So we got a comparable corpus of 345,575 pairs of sentences (named C_{all}).

4.2 Building the initial translation system

In order to apply the proposed unsupervised method, we have split the corpus C_{all} into two sets: an initial training corpus C2 and a mining corpus D (C2 and D are referred in figure 1b). To ensure a minimum quality for C2 (and consequently for the initial translation system S_0), we propose the following cross-filtering process to extract C2.

- Split the corpus C_{all} into 4 sub-corpora containing different sentence pairs: SC_1 (85,011 sentence pairs), SC_2 (85,008 sentence pairs), SC_3 (86,529 sentence pairs), SC_4 (89,027 sentence pairs).
- Build 4 different translation systems from 4 sub-corpora: SC_1 à SMT_{sc1}, SC_2 à SMT_{sc2}, SC_3 à SMT_{sc3}, SC_4 à SMT_{sc4}.
- Apply the proposed unsupervised method for each pair of (SC_1, SMT_{sc2}), (SC_2, SMT_{sc1}), (SC_3, SMT_{sc4}), (SC_4, SMT_{sc3}). (with one iteration; PER* threshold=0.45 to ensure the reliability of extracted sentence pairs (according to figure 2) and an acceptable number of pairs to build SMT system). We obtain the extracted sentence pairs $C2_1$, $C2_2$, $C2_3$, $C2_4$, their union is considered as reliable enough for serving as C2 corpus. The rest is treated as corpus D.

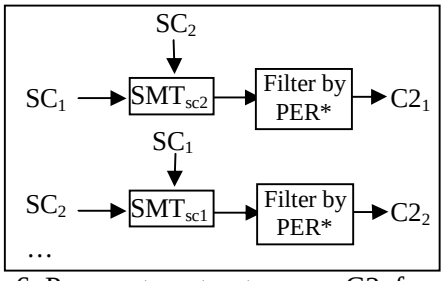

Figure 6. Process to extract corpus C2, for pair (SC_1, SMT_{sc2}), (SC_2, SMT_{sc1}), etc.

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

Sub-corpus	Translated by	Nbr. of extracted pairs (C2)	Nbr. of remaining pairs (D)
SC_1	SMT_{SC2}	$C2_1$: 2916	82095
SC_2	SMT_{SC1}	$C2_2$: 3495	81513
SC_3	SMT_{SC4}	$C2_3$: 3820	82709
SC_4	SMT_{SC3}	$C2_4$: 3892	85135

Table 2. Extracted data for C2 and D.

After this step, we obtained corpus C2 containing 14,123 sentence pairs, and corpus D containing 331,452 sentence pairs. The fully unsupervised method described in section 2.2 was then applied on C2 and D to filter more parallel sentence pairs.

4.3 Applying unsupervised method

The initial translation system S_0 was built from the training corpus C2 of 14,123 French-Vietnamese sentence pairs. The corpus D contains 331,452 French-Vietnamese sentence pairs. The unsupervised method was applied with the type of combination W2 and the evaluation metric PER*. There is no tuning process for the statistical models. The number of extracted sentence pairs after each iteration is reported in figure 7. After 5 iterations, we obtained 39,758 sentence pairs. The quality of the translation systems was also evaluated on a test set of 400 manually extracted Vietnamese-French parallel sentence pairs (same test set as in the implementation of *Method1*). The Vietnamese sentences were initially segmented into syllables (no word segmentation pre-processing was applied). Each Vietnamese sentence has only one French reference. The evaluation scores after each iteration were reported in table 3.

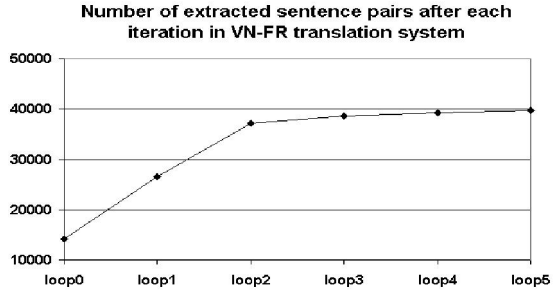

Figure 7. Number of extracted sentence pairs after each iteration

SMT iter.	Training data (nbr. of pairs)	Bleu	Nist	Ter
0	14,123	30.67	6.45	0.59
1	26,517	32.18	6.70	0.57
2	37,210	32.42	6.75	0.56
3	38,530	32.45	6.77	0.55
4	39,254	32.14	6.73	0.56
5	39,758	31.85	6.68	0.56

Table 3. Evaluation scores after each iteration.

The results in this case are similar to those in preliminary test: the number of extracted sentence pairs was increased after iterations; the quality of translation system was increased in some first iterations and decreased after that. Although the number of training sentence pairs increased about two times from iteration 0 to iteration 1, the evaluation score increased only 2 points for Bleu. One reason may be that the initial system (S_0) has already a good performance due to our cross-filtering process described in section 4.2. Moreover, the evaluation is only conducted with automatic metrics using one reference only and a deeper analysis should be conducted with human evaluations.

Furthermore, to compare with the *Method1*, the quality of the translation systems trained by extracted sentence pairs from two methods is given in table 4. Although the number of extracted sentence pairs in our method is lower than that in the *Method1*, the quality of the SMT system is comparable. Note that the *Method1* depends highly on the additional data such as the quality of bilingual dictionary or filtering heuristics.

From these results, we can say that the unsupervised method was applied successfully in a real low e-resourced language pair: Vietnamese - French. The result shows that this method can be really applied in the case of lacking parallel data. Moreover, the quality of the translation system built from extracted data is comparable with the translation system built from other method using lexical information (bilingual dictionary, etc.) and data filtering heuristics. This proposed method requires no more additional data. We intend to apply this method on a larger scale for mining a bigger comparable data stream extracted from the web.

Mining method	Nbr. of training data	Bleu	Nist	Ter
Lexical info. + Heuristics (*Method1*)	50,322	32.74	6.78	0.55
Unsupervised method	38,530	32.45	6.77	0.56

Table 4. Comparison between mining *Method1* and unsupervised method.

5 Conclusion and perspectives

This paper presents an unsupervised method for extracting parallel sentence pairs from a comparable/noisy parallel corpus. An initial translation system was built based on a noisy parallel cor-

pus, instead of a truly parallel corpus. The initial translation system was then used to translate another comparable corpus, to withdraw the parallel sentence pairs. An iterative process was evaluated to increase the number of extracted parallel sentence pairs and to improve the quality of translation system. The method was preliminary tested in a hard condition: the parallel corpus does not exist and the initial corpus contains up to 50% of non parallel sentence pairs. However, the result shows that this method can be really applied, especially in the case of lacking parallel data. Several ways of filtering and use the extracted data were also presented (different evaluation metrics for filtering and different ways of combining the extracted data with the initial translation system). An interesting result is that the quality of the translation system can be improved during some first iterations, but it becomes worse later because of adding noisy data into the statistical models. Moreover, 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.).

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 unsupervised method.

References

Abdul-Rauf, S. and H. Schwenk. 2009. 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*.

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

Doddington G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Human Language Technology Proceedings.*

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

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

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

Hutchins, W.J. 2001. Machine translation over fifty years. *Histoire, epistemologie, langage*. ISSN 0750-8069.

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

Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. *Machine Translation Summit*.

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

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, R. Zens, M. Federico, N. Bertoldi, B. Cowan, W. Shen and C. Moran. 2007. Moses: open source tool-kit for statistical machine translation. *Proceedings of the Association for Computational Linguistics*.

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

Ma, Xiaoyi. 2006. Champollion: A robust parallel text sentence aligner. *LREC: Fifth International Conference on Language Resources and Evaluation.*

Munteanu, D.S. and D. Marcu. 2006. Extracting parallel subsentential fragments from non-parallel corpora. *44th annual meeting of the Association for Computational Linguistics*.

Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29.1*

Papineni K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU:a method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Patry, A. and P. Langlais. 2005. Paradocs: un système d'identification automatique de documents parallèles. *12e Conference sur le Traitement Automatique des Langues Naturelles*.

Resnik, P. and N.A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics.*

Sarikaya R., S. Maskey, R. Zhang, E. Jan, D. Wang, B. Ramabhadran, S. Roukos. 2009. Iterative sentence–pair extraction from quasi–parallel corpora for machine translation. *Interspeech*.

Snover M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*.

Stolcke, Andreas. 2002. SRILM an extensible language modeling toolkit. *Intl. Conf. on Spoken Language Processing.*

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

Do T.N.D., V.B. Le, B. Bigi, L. Besacier, E. Castelli. 2009. Mining a comparable text corpus for a Vietnamese-French statistical machine translation system. *4th Workshop on Statistical Machine Translation.*

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