This paper presents an unsupervised method in application of extracting parallel sentence pairs from a comparable corpus. A translation system is used to mine the comparable corpus and to withdraw the parallel sentence pairs. An iteration process is implemented not only to increase the number of extracted parallel sentence pairs but also to improve the quality of translation system. A comparison between this unsupervised method and a semi-supervised method is also presented. The unsupervised extracting method was tested in a hard condition: the parallel corpus did not exist and the comparable corpus contained up to 50% of non parallel sentence pairs. However, the result shows that the unsupervised method can be really applied in the case of lacking parallel data.

Index Terms—unsupervised method, extract parallel sentence pairs, comparable corpus.

1. INTRODUCTION

Over the past fifty years of development [1], machine translation (MT) has obtained good results when applied to several pairs of languages such as English-French, English-Italian, 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 SMT for low-resourced languages always faces the challenge of getting enough data to support any particular approach.

Statistical machine translation tries to generate translations using statistical methods based on large parallel bilingual corpora for source and target languages. These corpora are used 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 [2], [3]. Thus, a large parallel bilingual text corpus is a prerequisite. Such a corpus is not always available, especially for low-resourced languages.

The most common methods to build parallel corpora consist in automatic methods which collect parallel sentence pairs from the Web [4], [5], or alignment methods which extract parallel documents/sentences from two monolingual corpora [6], [7], [8]. Beside these “traditional” methods, there is also the method of extracting parallel sentence pairs from a comparable corpus. For instance, Sadaf and Schwenk present a semi-supervised extracting method [9]. This kind of method requires an initial parallel corpus (see more in section 2.1). We assume that in the case of a low-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 noisy comparable corpus, is able to solve the problem of lacking parallel data?

This paper presents a fully unsupervised extracting method, in comparison with 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 two methods of extracting parallel sentence pairs from a comparable corpus: semi-supervised method versus fully unsupervised method. Section 3 gives our experiments and our results on testing the unsupervised method. The next section presents an application of this method for a real low-resourced language pair: Vietnamese-French. The last section concludes and gives some perspectives.

2. SEMI-SUPERVISED V/S UNSUPERVISED LEARNING

2.1 Semi-supervised learning method

Using a comparable corpus to extract parallel data has been presented in some previous works. D.S. Munteanu and D. Marcu present a method for extracting parallel sub-sentential fragments from comparable bilingual corpora [10].

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. Parallel sentence pairs are then filtered and parallel sub-sentential fragments are extracted from this document pair (see more in [10]).

S. Abdul-Rauf and H. Schwenk also present a method for extracting parallel data from a comparable corpus. To mine a comparable French-English corpus, for example, a statistical machine translation system is used to translate the French side to English. These translated texts are then compared with the English side, using the evaluation metric TER, and the parallel sentence pairs are filtered out. A post-processing is then applied to smooth the results. This technique is similar to that of [10], but a proper statistical machine translation system is used instead of the bilingual dictionary, and an evaluation metric is used to decide the degree of parallelism between two sentences.

All these methods are presented as effective methods to extracting parallel fragments/sentences from a comparable corpus.

2.2. Unsupervised learning method

The two above mentioned methods can be considered as semi-supervised methods, which need an initial parallel corpus to build the extracting system. We assume that in the case of low-resourced languages, this parallel corpus, even small, may be not available. So, we try to propose a fully unsupervised method, here, where the starting point is a simple noisy comparable corpus containing a significant amount of non parallel sentences. One of the challenges of this work is to see if such a different starting point (noisy comparable corpus, versus truly parallel corpus) can lead to the design of an acceptable SMT system.

Firstly, a baseline statistical machine translation system $S_0$ is built based on a comparable corpus ($C_2$) (in semi-supervised method, the system $S_0$ is built from a parallel corpus ($C_1$)). Of course the quality of $S_0$ is not high. We propose to use this system to mine another comparable corpus ($D$), and also to improve the quality of the translation system.

Secondly, the source side of the corpus $D$ (in our case the source language is French) is translated by the system $S_0$. The translated output is then compared with the target side (English in our case) of the corpus $D$. The evaluation metric is calculated for each sentence pairs. The pairs are considered as parallel sentence pairs if the evaluation metric is larger than a threshold.

In our research, several evaluation metrics are used to determine which one is the most suitable. The scores are estimated at sentence level. Four common evaluation metrics are used: BLEU [11], NIST [12], TER [13] and a modified PER* (see in section 3.3).

The extracted sentence pairs are then combined with the baseline system $S_0$ in several ways to create a new translation system. An iteration process is performed which re-translates the French 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.

![Diagram](image)

Figure 1: Semi-supervised v/s unsupervised methods

Again, to reuse the extracted parallel data in translation system, different combinations can be proposed:
- **W1:** The translation system at step $i$ is retrained on a training corpus consisting of $C_2$ and $E_{i-1}$ (the extracted data from the last iteration); $E_0$ being the data extracted when translation system is trained on $C_2$ only ($S_0$).
- **W2:** The translation system at step $i$ is retrained on training corpus consisting of $C_2$ and $E_0+E_1+...+E_{i-1}$ (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 $S_0$ and the new one are weighted, e.g. 1:2.

The section 3 presents our experiments on this unsupervised method.

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. Data was chosen from the Europarl corpus.
[6], version 3. The correct parallel sentence pairs were extracted directly from the Europarl corpus and a comparable corpus was simulated by introducing a significant amount of wrong sentence pairs in the data (about 50%).

To make it comparable with the real case treated in section 4 (low-resourced language pair), the size of the experimental data was chosen small. 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 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 [14]. 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.

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

3.3. Starting with parallel or comparable corpus?

One question that we want to answer first is whether the translation system based on a comparable 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 scores were used in this comparison: BLEU, NIST, TER and PER*. Our modified position-independent word error rate (PER*) is calculated based on the similarity, while the PER [15] measures the difference, of words occurring in hypotheses and reference.

\[
\text{PER}^* = \frac{2 \times \text{number of identical words}}{\text{length of hypothesis} + \text{length of reference}}
\]

Then 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 comparable corpus can replace a parallel corpus for constructing an initial translation system. Remember that the initial comparable corpus here contains up to 50% non-parallel sentence pairs. Therefore, this kind of unsupervised method can be really applied in the case of lacking parallel data. 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 and filter out the wrong ones. Table 1 presents the precision and recall of filtering parallel sentence pairs from two systems.
3.4. The iterations of the unsupervised method

Section 3.3 has shown that an unsupervised method can be also used to filter the parallel sentence pairs from a comparable 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 (table 1)). 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.

3.4.1. The number of correct extracted sentence pairs

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

The combination W2 brought the largest number of correct extracted sentence pairs.

3.4.2. The precision and the recall of 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 Sys1 (97.85) (PER*=0.3).

3.4.3. 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 can increase quickly during some first iterations, then increase slowly and then it can be decreased after several iterations. It can be explained that for the first iterations, the new parallel sentence pairs are included into the translation model, so it increase the translation quality. However, for the next iterations, 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 translation system was reduced.

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

4. APPLICATION FOR FRENCH-VIETNAMESE 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 [1]. Until now, in Vietnam, there are only four research groups working on MT [16]. However the results are still modest.

We focus on building a French-Vietnamese statistical machine translation (SMT) system. The training corpus was created by mining a bilingual news corpus from the Web. The mining process was presented in [17]. In [17], the parameters of mining process were adjusted to obtain parallel sentence pairs. But, in this research, to test the unsupervised method, we adjust the parameters to obtain comparable sentence pairs corresponding to a comparable corpus similar to that of previous section (including wrong parallel sentence pairs).

The initial translation system was built from a comparable training corpus C2 of 30,000 French-Vietnamese sentence pairs. The corpus D contains 21,000 French-Vietnamese sentence pairs. In these corpora, we do not know how many correct parallel sentences are included. The unsupervised method was applied. There is no tuning process for the statistical models. The number of extracted sentence pairs after several iterations was reported in figure 6.

The quality of the translation systems was also evaluated on a test set of 400 manually extracted French-Vietnamese parallel sentence pairs [17]. Each French sentence has only one Vietnamese reference. The evaluation scores were reported in figure 7.

The unsupervised method was applied in a real low-resourced language pair: French-Vietnamese. The result shows that this method can be really applied in the case of lacking parallel data. The quality of the translation system increased during several iterations. We intend to apply this method on a large scale of mining the real comparable data stream extracted from the web.

5. RELATED WORKS

Beside several researches on the semi-supervised method mentioned in previous sections ([9], [10]), there are also researches involving our work. In [18], Zhao and Vogel propose a maximum likelihood criterion which combines sentence length models 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 by using the extracted data. Sarikaya et al. present a semi-supervised method with iterations and the initial translation system is based on parallel corpus [19]. The method is also presented as an efficient method in filtering the parallel sentence pairs from a comparable corpus. In this research, authors use a
different evaluation metric (Bleu), and use the type of combination like our W2 type. However, their research does not provide a full explanation about how they choose evaluation metric, or combination method, and further more, the problem of decreasing the quality of translation system after several iterations is not mentioned.

6. CONCLUSION AND PERSPECTIVES

This paper presents an unsupervised method for extracting parallel sentence pairs from a comparable corpus. An initial translation system was built based on a comparable corpus, instead of a parallel corpus. The initial translation system was then used to translate another comparable corpus, to withdraw the parallel sentence pairs. An iteration process was implemented to increase the number of extracted parallel sentence pairs and to improve the quality of translation system. The method was tested in a hard condition: the parallel corpus does not exist and the comparable 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 using this method was also presented, with different evaluation metrics 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 the first iterations, but it becomes worse later because of adding the noisy data into the statistical models.

The next work of this research focuses on how to decrease this undesired problem. After some first iterations, the filtering may be altered to respect the precision, instead of the recall. Additionally the new way of reuse extracted parallel sentence pairs will be researched.

11. REFERENCES