Mining a comparable text corpus for a Vietnamese - French statistical machine translation system

Thi-Ngoc-Diep Do *,, Viet-Bac Le *, Brigitte Bigi*,**
Laurent Besacier*, Eric Castelli**
*LIG Laboratory, CNRS/UMR-5217, Grenoble, France
** MICA Center, CNRS/UMI-2954, Hanoi, Vietnam
thi-ngoc-diep.do@imag.fr

Abstract

This paper presents our first attempt at constructing a Vietnamese-French statistical machine translation system. Since Vietnamese is an under-resourced language, we concentrate on building a large Vietnamese-French parallel corpus. A document alignment method based on publication date, special words and sentence alignment result is proposed. The paper also presents an application of the obtained parallel corpus to the construction of a Vietnamese-French statistical machine translation system, where the use of different units for Vietnamese (syllables, words, or their combinations) is discussed.

1 Introduction

Over the past fifty years of development, machine translation (MT) has obtained good results when applied to several pairs of languages such as English, French, German, Japanese, etc. However, for under-resourced languages, it still remains a big gap. For instance, although Vietnamese is the 14th widely-used language in the world, research on MT for Vietnamese is very 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 for Vietnamese-English (Ho, 2005). However the results are still modest.

MT research on Vietnamese-French occurs even more rarely. Doan (2001) proposed a translation module for Vietnamese within ITS3, a multilingual MT system based on the classical analysis-transfer-generation approach. Nguyen (2006) worked on Vietnamese language and Vietnamese-French text alignment. But no complete MT system for this pair of languages has been published so far.

There are many approaches for MT: rule-based (direct translation, interlingua-based, transfer-based), corpus-based (statistical, example-based) as well as hybrid approaches. We focus on building a Vietnamese-French statistical machine translation (SMT) system. Such an approach requires a parallel bilingual corpus for source and target languages. Using this corpus, we build a statistical translation model for source/target languages and a statistical language model for target language. Then the two models and a search module are used to decode the best translation (Brown et al., 1993; Koehn et al., 2003).

Thus, the first task is to build a large parallel bilingual text corpus. This corpus can be described as a set of bilingual sentence pairs. At the moment, such a large parallel corpus for Vietnamese-French is unavailable. (Nguyen, 2006) presents a Vietnamese-French parallel corpus of law and economics documents. Our SMT system was trained using Vietnamese-French news corpus created by mining a comparable bilingual text corpus from the Web.

Section 2 presents the general methodology of mining a comparable text corpus. We present an overview of document alignment methods and sentence alignment methods, and discuss the document alignment method we utilized, which is based on publishing date, special words, and sentence alignment results. Section 3 describes our experiments in automatically mining a multilingual news website to create a Vietnamese-French parallel text corpus. Section 4 presents

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 165–172,
Athens, Greece, 30 March – 31 March 2009.

our application to rapidly build Vietnamese-French SMT systems using the obtained parallel corpus, where the use of different units for Vietnamese (syllables, words, or their combination) is discussed. Section 5 concludes and discusses future work.

2 Mining a comparable text corpus

In (Munteanu and Daniel Marcu, 2006), the authors present a method for extracting parallel sub-sentential fragments from comparable bilingual corpora. However this method is in need of an initial parallel bilingual corpus, which is not available for the pair of language Vietnamese-French (in the news domain).

The overall process of mining a bilingual text corpus which is used in a SMT system typically takes five following steps (Koehn, 2005): raw data collection, document alignment, sentence splitting, tokenization and sentence alignment. This section presents the two main steps: document alignment and sentence alignment. We also discuss the proposed document alignment method.

2.1 Document alignment

Let *S1* be set of documents in language *L1*; let *S2* be set of documents in language *L2*. Extracting parallel documents or aligning documents from the two sets *S1, S2* can be seen as finding the translation document *D2* (in the set *S2*) of a document *D1* (in the set *S1*). We call this pair of documents *D1-D2* a *parallel document pair (PDP).*

For collecting bilingual text data for the two sets *S1, S2*, the Web is an ideal source as it is large, free and available (Kilgarriff and Grefenstette, 2003). For this kind of data, various methods to align documents have been proposed. Documents can be simply aligned based on the anchor link, the clue in URL (Kraaij et al., 2003) or the web page structure (Resnik and Smith, 2003). However, this information is not always available or trustworthy. The titles of documents *D1, D2* can also be used (Yang and Li, 2002), but sometimes they are completely different.

Another useful source of information is invariant words, such as named entities, dates, and numbers, which are often common in news data. We call these words *special words*. (Patry and Langlais, 2005) used numbers, punctuation, and entity names to measure the parallelism between two documents. The order of this information in document is used as an important criterion. However, this order is not always respected in a PDP (see an example in Table 1).

French document	Vietnamese document
*Selon l'Administration nationale du tourisme, les voyageurs en provenance de l'Asie du Nord-Est (Japon, République de Corée,...) représentent **33**%, de l'Europe, **16**%, de l'Amérique du Nord, **13**%, d'Australie et de Nouvelle-Zélande, **6**%.* *En outre, depuis le début de cette année, environ **2,8** millions de touristes étrangers ont fait le tour du Vietnam, **78**% d'eux sont venus par avion.* *Cela témoigne d'un afflux des touristes riches au Vietnam.…*	*Trong số gần **2,8** triệu lượt khách quốc tế đến Việt Nam từ đầu năm đến nay, lượng khách đến bằng đường hàng không vẫn chiếm chủ đạo với khoảng **78**%.* *Điều này cho thấy, dòng khách du lịch chất lượng cao đến Việt Nam tăng nhanh.* *Theo thống kê thị khách quốc tế vào Việt Nam cho thấy khách Đông Bắc Á (Nhật Bản, Hàn Quốc) chiếm tới **33**%, châu Âu chiếm **16**%, Bắc Mỹ **13**%, Ôxtrâylia và Niu Dilân chiếm **6**%.…*

Table 1. An example of a French-Vietnamese parallel document pair in our corpus.

2.2 Sentence alignment

From a PDP *D1-D2*, the sentence alignment process identifies parallel sentence pairs (PSPs) between two documents *D1* and *D2*. For each *D1-D2*, we have a set SenAlignmentD1-D2 of PSPs.

$SenAlignment_{D1\text{-}D2}$ = {"sen1-sen2"| sen1 is zero/one/many sentence(s) in document D1, sen2 is zero/one/many sentence(s) in document D2, sen1-sen2 is considered as a PSP}.

We call a PSP *sen1-sen2* alignment type *m:n* when *sen1* contains *m* consecutive sentences and *sen2* contains *n* consecutive sentences.

Several automatic sentence alignment approaches have been proposed based on sentence length (Brown et al., 1991) and lexical information (Kay and Roscheisen, 1993). A hybrid approach is presented in (Gale and Church, 1993) whose basic hypothesis is that "longer sentences in one language tend to be translated into longer sentences in the other language, and shorter sentences tend to be translated into shorter sentences". Some toolkits such as Hunalign[1] and Vanilla[2] implement these approaches. However, they tend to work best when documents *D1, D2* contain few sentence deletions and insertions, and mainly contain PSPs of type 1:1.

[1] http://mokk.bme.hu/resources/hunalign
[2] http://nl.ijs.si/telri/Vanilla/

Ma (2006) provides an open source software called Champollion[1] to solve this limitation. Champollion permits alignment type *m:n* (*m, n = 0,1,2,3,4*), so the length of sentence does not play an important role. Champollion uses also lexical information (lexemes, stop words, bilingual dictionary, etc.) to align sentences. Champollion can easily be adapted to new pairs of languages. Available language pairs in Champollion are English-Arabic and English-Chinese (Ma, 2006).

2.3 Our document alignment method

Figure 1 describes our methodology for document alignment. For each document *D1* in the set *S1*, we find the aligned document *D2* in the set *S2*.

We propose to use publishing date, special words, and the results of sentence alignment to discover PDPs. First, the publishing date is used to reduce the number of possible documents *D2*. Then we use a filter based on special words contained in the documents to determine the candidate documents *D2*. Finally, we eliminate candidates in *D2* based on the combination of document length information and lexical information, which are extracted from the results of sentence alignment.

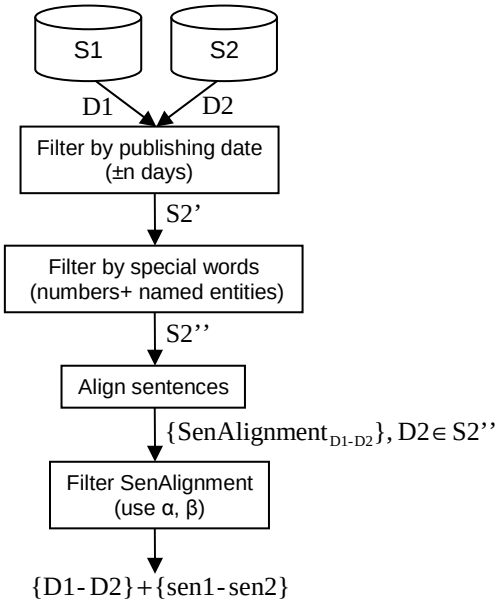

Figure 1. Our document alignment scheme.

2.3.1 The first filter: publishing date

We assume that the document *D2* is translated and published at most *n* days after the publishing date of the original document. We do not know whether *D1* or *D2* is the original document, so we assume that *D2* is published *n* days before or after *D1*. After filtering by publishing date criterion, we obtain a subset *S2'* containing possible documents *D2*.

2.3.2 The second filter: special words

In our case, the special words are *numbers* and *named entities*. Not only numbers (*0-9*) but also attached symbols (*'$', '%', '‰', ',', '.'*…) are extracted from documents, for example: "*12.000$"; "13,45"; "50%";*… Named entities are specified by one or several words in which the first letter of each word is upper case, e.g. "*Paris*", "*Nations Unies*" in French.

While named entities in language *L1* are usually translated into the corresponding names in language *L2*, in some cases the named entities in *L1* (such as personal names or organization names) do not change in *L2*. In particular, many Vietnamese personal names are translated into other languages by removal of diacritical marks (see examples in Table 2).

	French	**Vietnamese**	**Vietnamese -Removed diacritic**
Changed	*Nations Unies*	*Liên Hợp Quốc*	*Lien Hop Quoc*
	France	*Pháp*	*Phap*
Not changed	*ASEAN*	*ASEAN*	*ASEAN*
	Nong Duc Manh	*Nông Đức Mạnh*	*Nong Duc Manh*
	Dien Bien	*Điện Biên*	*Dien Bien*

Table 2. Some examples of named entities in French-Vietnamese.

All special words are extracted from document *D1*. This gives a list of special words $w_1,w_2,\ldots w_n$. For each special word, we search in the set *S2'* documents *D2* which contain this special word. For each word, we obtain a list of documents *D2*. The document *D2* which has the biggest number of appearance in all lists is chosen. It is the document containing the highest number of special words. We can find zero, one or several documents which are satisfactory. We call this set of documents set *S2''* (see in Figure 2).

The way that we use special words is different from the way used in (Patry and Langlais, 2005). We do not use punctuation as special words. We use the attached symbols ('$', '%', '‰', …) with the number. Furthermore, in our method, the order of special words in documents is not important, and if a special word appears several times in a document, it does not affect the result.

[1] http://champollion.sourceforge.net

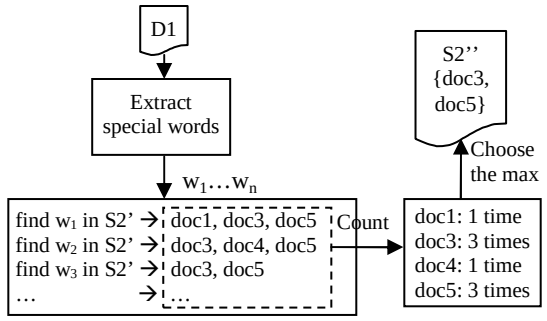

Figure 2. Using special words to filter documents *D2*.

2.3.3 The third filter: sentence alignments

As mentioned in section 2.3.2, for each document *D1*, we discover a set *S2''*, which contains zero, one or several documents *D2*. When we continue to align sentences for each PDP *D1-D2*, we get a lot of low quality PSPs. The results of sentence alignment allow us to further filter the documents *D2*.

After aligning sentences, we have a set of PSPs, $SenAlignment_{D1\text{-}D2}$, for each PDP *D1-D2*. We add two rules to filter documents *D2*.

When *D1-D2* is not a true PDP, it is hard to find out PSPs. So we note the number of PSPs in the set $SenAlignment_{D1\text{-}D2}$ by $card(SenAlignment_{D1\text{-}D2})$. The number of sentence pairs which can not find their alignment partner (when *sen1* or *sen2* is "null") is noted by $nbr\_omitted(SenAlignment_{D1\text{-}D2})$.

When $\frac{\text{nbr\_omitted}(\text{SenAlignment}_{\text{D1-D2}})}{\text{card}(\text{SenAlignment}_{\text{D1-D2}})} > \alpha$, this PDP *D1-D2* will be eliminated.

This first rule also deals with the problem of document length, sentence deletions and sentence insertions.

The second rule makes use of lexical information. For each PSP, we add two scores x_{L1} and x_{L2} for *sen1* and *sen2*.

$$x_{Li} = \frac{number - of - translated - words - in - sen_i}{number - of - words - in - sen_i}$$

Translated words are words having translation equivalents in the other sentence. In this rule, we do not take into account the stop words. Table 3 shows an example for calculating two scores x_{L1} and x_{L2} for a PSP.

In the second rule, when all PSPs in $SenAlignment_{D1\text{-}D2}$ have two scores x_{L1} and x_{L2} that are both smaller than β, this PDP *D1-D2* will be eliminated. This rule removes the low quality PDP which creates a set of low quality PSPs.

sen1 **(in French)** : *ils ont échangé leurs opinions pour parvenir à la signature de documents constituant la base du développement et de l' intensification de la coopération en économie en commerce et en investissement ainsi que celles dans la culture le sport et le tourisme entre les deux pays*	
sen2 **(in Vietnamese)** : *hai bên đã tiến_hành trao_đổi để ký_kết các văn_bản làm cơ_sở cho việc mở_rộng và tăng_cường quan_hệ hợp_tác kinh_tế thương_mại đầu_tư văn_hoá thể_thao và du_lịch giữa hai nước*	
Translated words : *"échanger:trao_đổi" ;"base:cơ_sở","intensification:tăng_cường" ;"coopération:hợp_tác","économie:kinh_tế" ; investissement:đầu_tư","sport:thể_thao" ; "tourisme :du_lịch" ; "pays:nước"*	
Number of non-stop words in *sen1*	19
Number of non-stop words in *sen2*	21
Number of translated words	9
$x_{L1} = 9/19 = 0.47$; $x_{L2} = 9/21 = 0.43$	

Table 3. Example for calculating two scores x_{L1} and x_{L2}.

After using three filters based on information of publishing date, special words, and the results of sentence alignment, we have a corpus of PDPs, and also a corpus of corresponding PSPs. To ensure the quality of output PSPs, we can continue to filter PSPs. For example, we can keep only the PSPs whose scores (x_{L1} and x_{L2}) are higher than a threshold.

3 Experiments

3.1 Characteristics of Vietnamese

The basic unit of the Vietnamese language is syllable. In writing, syllables are separated by a white space. One word corresponds to one or more syllables (Nguyen, 2006). Table 4 presents an example of a Vietnamese sentence segmented into syllables and words.

Vietnamese sentence: *Thành phố hy vọng sẽ đón nhận khoảng 3 triệu khách du lịch nước ngoài trong năm nay* **Segmentation in syllables**: *Thành | phố | hy | vọng | sẽ | đón | nhận | khoảng | 3 | triệu | khách | du | lịch | nước | ngoài | trong | năm | nay* **Segmentation in words**: *Thành_phố | hy_vọng | sẽ | đón_nhận | khoảng | 3 | triệu | khách_du_lịch | nước_ngoài | trong | năm | nay*
Corresponding English sentence: *The city is expected to receive 3 million foreign tourists this year*

Table 4. An example of a Vietnamese sentence segmented into syllables and words.

In Vietnamese, words do not change their form. Instead of conjugation for verb, noun or adjective, Vietnamese language uses additional words, such as "*những*", "*các*" to express the plu-

ral; "*đã*", "*sẽ*" to express the past tense and the future. The syntactic functions are also determined by the order of words in the sentence (Nguyen, 2006).

3.2 Data collecting

In order to build a Vietnamese-French parallel text corpus, we applied our proposed methodology to mine a comparable text corpus from a Vietnamese daily news website, the *Vietnam News Agency*[1] (VNA). This website contains news articles written in four languages (Vietnamese, English, French, and Spanish) and divided in 9 categories including "Politics - Diplomacy", "Society - Education", "Business - Finance", "Culture - Sports", "Science - Technology", "Health", "Environment", "Asian corner" and "World". However, not all of the Vietnamese articles have been translated into the other three languages. The distribution of the amount of data in four languages is shown in figure 3.

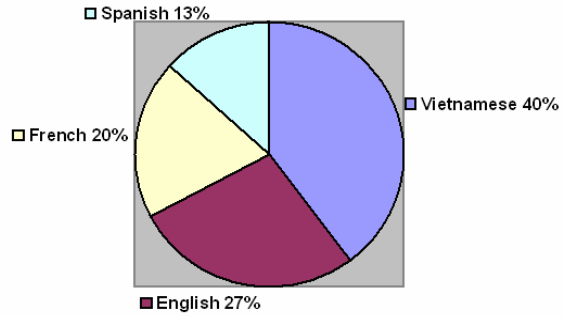

Figure 3. Distribution of the amount of data for each language on VNA website.

Each document (i.e., article) can be obtained via a permanent URL link from VNA. To date, we have obtained about 121,000 documents in four languages, which are gathered from 12 April 2006 to 14 August 2008; each document contains, on average, 10 sentences, with around 30 words per sentence.

3.3 Data pre-processing

We splitted the collected data into 2 sets. The development set, designated S_{DEV}, contained 1000 documents, was used to tune the mining system parameters. The rest of data, designated $S_{TRAIN,}$ was used as a training set, where the estimated parameters were applied to build the entire corpus. We applied the following pre-process to each set S_{DEV} and S_{TRAIN}:

1. Extract contents from documents.
2. Classify documents by language (using TextCat[2], an n-gram based language identification).
3. Process and clean both Vietnamese and French documents by using the CLIPS-Text-Tk toolkit (LE et al., 2003): convert html to text file, convert character code, segment sentence, segment word. The resulting clean corpora are *S1* (for French) and *S2* (for Vietnamese).

3.4 Parameters estimation

Our proposed document alignment method was applied to the sets *S1* and *S2* extracted from the set S_{DEV}. To filter by publishing date, we assumed that *n*=2.

The second filter was implemented on the set *S1* and the new set *S2** which was created by removing diacritical marks from the set *S2* (in the case of Vietnamese).

The sentence alignment process was implemented by using data from sets *S1*, *S2* and the Champollion toolkit. We adapted Champollion to Vietnamese-French by changing some parameters: the ratio of French word to Vietnamese translation word is set to 1.2, penalty for alignment type 1-1 is set to 1, for type 0-1 to 0.8, for type 2-1, 1-2 and 2-2 to 0.75, and we did not use the other types (see more in (Ma, 2006)). After using two filters, the result data is shown in Table 5. The true PDPs were manually extracted.

S_{DEV}	- Number of documents: 1000 - Number of French documents: 173 - Number of Vietnamese documents: 348 - Number of true PDPs: 129
S2''	- Number of found PDPs: 379 - Number of hits PDPs: 129 - Precision = 34.04% , Recall = 100%

Table 5. Result data after using two filters.

The third filter was applied in which α was set to (*0.4, 0.5, 0.6, 0.7*) and β was set to (*0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4*). The precision and recall were calculated according to our true PDPs and the F-measure (F1 score) was estimated.

F-measure

α / β	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.4	0.69	0.71	0.71	0.60	0.48	0.36	0.21
0.5	0.76	0.79	0.77	0.65	0.52	0.39	0.23
0.6	0.77	0.83	0.82	0.70	0.56	0.41	0.26
0.7	0.75	0.84	0.83	0.73	0.59	0.44	0.27

Table 6. Filter result with different values of α and β on the S_{DEV}.

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

[2] http://www.let.rug.nl/~vannoord/TextCat/

From the results mentioned in Table 6, we chose α=*0.7* and β=*0.15*.

3.5 Mining the entire corpus

We applied the same methodology with the parameters estimated in section 3.4 to the set S_{TRAIN}. The obtained corpus is presented in Table 7.

S_{TRAIN}	- Number of documents: 120,218 - Number of French documents: 20,884 - Number of Vietnamese documents: 54,406
Entire corpus	- Number of PDPs: 12,108 - Number of PSPs: 50,322

Table 7. The obtained corpus from S_{TRAIN}.

4 Application: a Vietnamese - French statistical machine translation system

With the obtained parallel corpus, we attempted to rapidly build a SMT system for Vietnamese-French. The system was built using the Moses toolkit[1]. The Moses toolkit contains all of the components needed to train both the translation model and the language model. It also contains tools for tuning these models using minimum error rate training and for evaluating the translation result using the BLEU score (Koehn et al., 2007).

4.1 Preparing data

From the entire corpus, we chose 50 PDPs (351 PSPs) for developing (Dev), 50 PDPs (384 PSPs) for testing (Tst), with the rest PDPs (49,587 PSPs) reserved for training (Trn).

Concerning the developing and testing PSPs, we manually verified and eliminated low quality PSPs, which produced 198 good quality PSPs for developing and 210 good quality PSPs for testing. The data used to create the language model were extracted from 49,587 PSPs of the training set.

4.2 Baseline system

We built translation systems in two translation directions: French to Vietnamese (F→V) and Vietnamese to French (V→F). The Vietnamese data were segmented into either words or syllables. So we first have four translation systems. We removed sentences longer than 100 words/syllables from the training and development sets according to the Moses condition (so the number of PSPs used in the training set differs slightly between systems). All words found are implicitly added to the vocabulary.

System	Direction	Vietnamese is segmented into	Nbr of PSPs
S1FV	F→V	Syllable	Training: 47,081 Developing: 198 Testing: 210
S1VF	V→F		
S2FV	F→V	Word	Training: 48,864 Developing: 198 Testing: 210
S2VF	V→F		

System	Set - Language		Nbr. of vocab (K)	Nbr. of running words/syllables (K)
S1FV S1VF	Trn	Fr	38.6	1783.6
		Vn	21.9	2190.2
	Dev	Fr	1.8	6.3
		Vn	1.2	6.9
	Tst	Fr	1.9	6.4
		Vn	1.3	7.1
S2FV S2VF	Trn	Fr	39.7	1893
		Vn	33.4	1629
	Dev	Fr	1.8	6.3
		Vn	1.5	4.8
	Tst	Fr	1.9	6.3
		Vn	1.6	4.9

Table 8. Our four translation systems.

We obtained the performance results for those systems in Table 9. In the case of the systems where Vietnamese was segmented into words, the Vietnamese sentences were changed back to syllable representation before calculating the BLEU scores, so that all the BLEU scores evaluated can be compared to each other.

	S1FV	S1VF	S2FV	S2VF
BLEU	*0.40*	*0.31*	*0.40*	*0.30*

Table 9. Evaluation of SMTs on the Tst set.

The BLEU scores for *French to Vietnamese* translation direction are around 0.40 and the BLEU scores for *Vietnamese to French* translation direction are around 0.31, which is encouraging as a first result. Moreover, only one reference was used to estimate BLEU scores in our experiments. It is also interesting to note that segmenting Vietnamese sentences into words or syllables does not significantly change the performance for both translation directions. An example of translation from four systems is presented in Table 10.

[1] http://www.statmt.org/moses/

Given a pair of parallel sentences FR: selon le département de gestion des travailleurs à l' étranger le qatar est un marché prometteur et nécessite une grande quantité de travailleurs étrangers VNsyl : theo cục quản lý lao động ngoài nước cata là thị trường đầy tiềm năng và có nhu cầu lớn lao động nước ngoài VNword : theo cục quản_lý lao_động ngoài nước cata là thị_trường đầy tiềm_năng và có nhu_cầu lớn lao_động nước_ngoài	
S1FV	Input: FR Reference: VNsyl Output: theo cục quản lý lao động ở nước ngoài phía cata là một thị trường đầy tiềm năng và cần một lượng lớn lao động nước ngoài
S2FR	Input: FR Reference: VNword Output: theo thống_kê của cục quản_lý lao_động ngoài nước cata là một thị_trường đầy tiềm_năng và cần có sự lớn lượng lao_động nước_ngoài
S1VF	Input: VNsyl Reference: FR Output: selon le département de gestion des travailleurs étrangers cata était un marché plein de potentialités et aux besoins importants travailleurs étrangers
S2VF	Input: VNword Reference: FR Output : selon le département de gestion des travailleurs étrangers cata marché plein de potentialités et la grande travailleurs étrangers

Table 10 : Example of translation from systems.

4.3 Combining word- and syllable-based systems

We performed another experiment on combining syllable and word units on the Vietnamese side. We carried out the experiment on the *Vietnamese to French* translation direction only. In fact, the Moses toolkit supports the combination of phrase-tables. The phrase-tables of the system S1VF (T_{syl}) and system S2VF (T_{word}) were used. Another phrase-table (T_{word*}) was created from the T_{word}, in which all words in the phrase table were changed back into syllable representation (in this latter case, the word segmentation information was used during the alignment process and the phrase table construction, while the unit kept at the end remains the syllable). The combinations of these three phrase-tables were also created (by simple concatenation of the phrase tables). The Vietnamese input for this experiment was either in word or in syllable representation. As usual, the developing set was used for tuning the log-linear weights and the testing set was used to estimate the BLEU score. The obtained results are presented in Table 11. Some performances are marked as X since those combinations of input and phrase table do not make sense (for instance the combination of input in words and syllable-based phrase table).

Phrase-tables used	Input in syllable		Input in word	
	Dev	Tst	Dev	Tst
T_{syl}	**0.35**	**0.31**	X	X
T_{word}	X	X	0.35	0.30
T_{word*}	0.37	0.31	X	X
$T_{syl} + T_{word}$	0.35	0.31	0.36	0.30
$T_{syl} + T_{word*}$	**0.38**	**0.32**	X	X
$T_{word} + T_{word*}$	0.37	0.30	0.36	0.30

Table 11: The BLEU scores obtained from combination of phrase-tables on Dev set and Tst set (Vietnamese to French machine translation).

These results show that the performance can be improved by combining information from word and syllable representations of Vietnamese. (BLEU improvement from 0.35 to 0.38 on the Dev set and from 0.31 to 0.32 on the Tst set). In the future, we will analyze more the combination of syllable and word units for Vietnamese MT and we will investigate the use of confusion networks as an MT input, which have the advantage to keep both segmentations (word, syllable) into a same structure.

4.4 Comparing with Google Translate[1]

Google Translate system has recently supported Vietnamese. In most cases, it uses English as an intermediary language. For the first comparative evaluation, some simple tests were carried out. Two sets of data were used: *in domain data set* (the Tst set in section 4.2) and *out of domain data set*. The latter was obtained from a Vietnamese-French bilingual website[2] which is not a news website. After pre-processing and aligning manually, we obtained 100 PSPs in the out of domain data set. In these tests, the Vietnamese data were segmented into syllables. Both data sets were inputted to our translation systems (S1FV, S1VF) and the Google Translate system. The outputs of Google Translate system were post-processed (lowercased) and then the BLEU scores were estimated. Table 12 presents the results of these tests. While our system is logically better for in domain data set, it is also slightly better than Google for out of domain data set.

[1] http://translate.google.com
[2] http://www.ambafrance-vn.org

	Direction	BLEU score	
		Our system	Google
In domain (210 PSPs)	F→V	0.40	0.25
	V→F	0.31	0.16
Out of domain (100 PSPs)	F→V	0.25	0.24
	V→F	0.20	0.16

Table 12: Comparing with Google Translate.

5 Conclusions and perspectives

In this paper, we have presented our work on mining a comparable Vietnamese-French corpus and our first attempts at Vietnamese-French SMT. The paper has presented our document alignment method, which is based on publication date, special words and sentence alignment result. The proposed method is applied to Vietnamese and French news data collected from VNA. For Vietnamese and French data, we obtained around 12,100 parallel document pairs and 50,300 parallel sentence pairs. This is our first Vietnamese-French parallel bilingual corpus. We have built SMT systems using Moses. The BLEU scores for *French to Vietnamese* translation systems and *Vietnamese to French* translation systems were 0.40 and 0.31 in turn. Moreover, combining information from word and syllable representations of Vietnamese can be useful to improve the performance of Vietnamese MT system.

In the future, we will attempt to increase the corpus size (by using unsupervised SMT for instance) and investigate further the use of different Vietnamese lexical units (syllable, word) in a MT system.

References

Brown, Peter F., Jennifer C. Lai and Robert L. Mercer. 1991. *Aligning sentences in parallel corpora*. Proceedings of 47th Annual Meeting of the Association for Computational Linguistics.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics. Vol. 19, no. 2.

Doan, Nguyen Hai. 2001. *Generation of Vietnamese for French-Vietnamese and English-Vietnamese Machine Translation*. ACL, Proceedings of the 8th European workshop on Natural Language Generation.

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

Ho, Tu Bao. 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.John. 2001. *Machine translation over fifty years*. Histoire, epistemologie, langage: HEL, ISSN 0750-8069, Vol. 23, Nº 1, 2001 , pages. 7-32.

Kay, Martin and Martin Roscheisen. 1993. *Text - translation alignment*. Association for Computational Linguistics.

Kilgarriff, Adam and Gregory Grefenstette. 2003. *Introduction to the Special Issue on the Web as Corpus*. Computational Linguistics, volume 29.

Koehn, Philipp, Franz Josef Och and Daniel Marcu. 2003. *Statistical phrase-based translation*. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1.

Koehn, Philipp. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. Machine Translation Summit.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Richard Zens, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen and Christine Moran. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. Proceedings of the ACL.

Kraaij, Wessel, Jian-Yun Nie and Michel Simard. 2003. *Embedding web-based statistical translation models in cross-language information retrieval*. Computational Linguistics, Volume 29 , Issue 3.

LE, Viet Bac, Brigitte Bigi, Laurent Besacier and Eric Castelli. 2003. *Using the Web for fast language model construction in minority languages*. Eurospeech'03.

Ma, Xiaoyi. 2006. *Champollion: A Robust Parallel Text Sentence Aligner*. LREC: Fifth International Conference on Language Resources and Evaluation.

Munteanu, Dragos Stefan and Daniel Marcu. 2006. *Extracting parallel sub-sentential fragments from non-parallel corpora* . 44th annual meeting of the Association for Computational Linguistics

Nguyen, Thi Minh Huyen. 2006. *Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens*. Thèse présentée pour l'obtention du titre de Docteur de l'Université Henri Poincaré, Nancy 1 en Informatique.

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

Resnik, Philip and Noah A. Smith. 2003. *The Web as a Parallel Corpus*. Computational Linguistics.

Yang, Christopher C. and Kar Wing Li. 2002. *Mining English/Chinese Parallel Documents from the World Wide Web*. Proceedings of the 11th International World Wide Web Conference, Honolulu, USA.