An Approach of Chunk Alignment for French-Vietnamese Bilingual Corpora

Ngoc Tan LE[1], Ngoc Tien LE[2] and Dien DINH[3]

[1] Faculty of Information Technology, Industrial University of Ho Chi Minh
Ho Chi Minh City, Vietnam

[2] Faculty of Information Technology, Industrial University of Ho Chi Minh
Ho Chi Minh City, Vietnam

[3] Faculty of Information Technology, University of Natural Sciences of Ho Chi Minh
Ho Chi Minh City, Vietnam

Abstract

The machine translation domain has been developed and improved very quickly. But the issue of long sentences is still a problem in this domain. Hence using phrase chunking on the purpose of reducing the length of sentences to improve the translation quality is a promising approach. In this paper, we present the approach of lexical analysis – phrase chunking – applied to French sentences in combination with a French-Vietnamese bilingual dictionary. And we also define the boundaries of the chunks to create a set of French-Vietnamese bilingual segments in order to overcome limitations due to the long sentences. We tested the system model with a French-Vietnamese bilingual corpus composed of 10,000 sentences pairs and evaluated on a sample of 100 sentences pairs in this corpus after the chunking process by our system. And our system has been evaluated with an accuracy more than 90%, and the value of F-measure is 91.61%.

Keywords: *Bilingual corpus, machine translation, extraction of parallel corpus, chunk alignment, French Tree Bank corpus, Conditional Random Fields.*

1. Introduction

The study about corpus was performed by many linguistics, computational linguistics since the 1950s. And it has gone through many ups and downs.With the boom of Internet and information of technology in the world in recent years, the volume of information in many different languages other than English has grown significantly.

To overcome the language barrier between different languages, dictionaries are the most typical tools.

As we know nowadays, the machine translation domain has been developed and improved very fast. But one of many problems in this domain is the issue of long sentences. So using phrase chunking in the purpose of reducing the length of sentences to improve the translation quality is a promising approach.

In this paper, we present the approach of lexical analysis – phrase chunking – applied to French sentences in combination with a French-Vietnamese bilingual dictionary. And we also define the boundaries of the chunks to create a set of French-Vietnamese bilingual segments pairs in order to overcome limitations due to the problem of the long sentences.

Based on ideas from the two articles of [1] and [3], our group presents a model of chunk alignment by solving the issue of long sentences in the French-Vietnamese bilingual language pair.

The rest of the paper is organized as follows: the section 2 presents some related works of this domain. The section 3 describes more about our approach - the method of chunk alignment for the French-Vietnamese language pair. Our experiments are shown in the section 4. And finally we give some conclusions in the section 5.

2. Related works

In Vietnam, with the development of information technology systems, the researches on machine translation for foreign languages (especially English) to Vietnamese began in the late 1980s. However, until now, there are very few research groups working on Vietnamese-English and Vietnamese-French, and the results are still modest. According to the ADD about the Asian languages[1], in Vietnam, there are only four major research teams

[1] Asian applied natural language processing for linguistics Diversity and language resource Development. http://www.tcllab.org/modules.php?name=events&file=ADD

in machine translation for Vietnamese-English language pair.

However, the researches on machine translation for Vietnamese-French is very limited. [Doan N.H., 2001] [8] introduced a translation module for Vietnamese in ITS3, a multilingual machine translation system, developed at the research laboratory of languages (*Laboratoire d'Analyse et de Technologie du langage - LATL*) Geneva. This module is based on transformation rules approaches. But this project is not complete.

In general, the research of French-Vietnamese language pair is performed separately in the phases of grammatical analysis such as the analysis of parsing tree of the author Le Hong Phuong [9] or the analysis of time tenses in Vietnamese by the author Nicolas Boffo [10]. There are also the works on French-Vietnamese bilingual text alignment of the author Nguyen Thi Minh Huyen [11] which was developed in her Ph.D. in 2006.

Indeed, to develop a statistical machine translation system, we need the large parallel corpus for French-Vietnamese or Vietnamese-French language pair. In 2006, [12] mentioned finding medical information through intermediate language Vietnamese-French-English. In which the author Tran Tuan Duc discussed the problem of creating multilingual Vietnamese-French-English corpora. But this method has disadvantages because the author only use monolingual corpus and bilingual dictionaries in order to extract parallel sentence pairs.

Recently, in 2010, for the French-Vietnamese language pair, there is a work about exploiting the comparable corpus for Vietnamese-French statistical machine translation system by the author Do Thi Ngoc Diep and her colleagues [7]. The system [7] uses some features such as the uppercase words for the proper names to find the parallel texts pairs. But this method is difficult to apply in some Asian languages such as Chinese or Japanese.

In many researches applying phrase chunking in machine translation, there are two main approaches. In the first one, they determine the alignment of phrase chunking in sentences pairs and they use this data as the main corpus for statistical translation model (such as the group of Sun Le [3]). Whereas in the second one, they use phrase chunking for building a set of reordering rules, which makes the results more accurate (as the group of Vinh Van Nguyen [4]). Besides, there is also the hybrid approach in combination with statistics and patterns, using keywords to split the sentences into phrase chunks for each language. Then, they use dynamic programming and statistics to find the patterns of phrase chunking alignment (as the group of Francisco Nevado [5]).

For the second approach, the group [4] has experimented with tests on English-Vietnamese language pair. However, the results are still limited, especially the result of the determination of the phrase chunking for Vietnamese language is not high. Whereas for the first approach, there is no article published for Vietnamese. This is why we do the research based on this approach.

There are some authors who used words alignment and the probability of alignment to find the break points splitting the sentences into segments with the best probability. From which, they build the alignment of phrase chunking. In the group [5], they use the marker-words in order to split the sentences into segments for each individual language. Then, they use dynamic programming with the probability of words alignment to determine the alignment of phrase chunking. Inspired by the idea of Sun Le group, we perform the phrase chunking of French language (the language has been deeply studied and have had the high and accurate results of phrase chunking), use also a French-Vietnamese bilingual dictionary and define the boundaries of translated words in order to remove the ambiguity and to determine the boundaries of each phrase chunking in a French sentence as well as in a Vietnamese sentence. From there, we build the phrases chunks alignment of French-Vietnamese language pair.

3. Method of chunk alignment

3.1 Our approach

In this paper, we use a French-Vietnamese bilingual corpus which has been aligned at the sentence level. The French sentences are annotated of part-of-speech (*POS tagging*) and annotated also of chunk tags (*chunk tagging*) by using the SEM[1] tool. This tool has been trained with the French Tree Bank corpus and is introduced by [2]. It is implemented with the Conditional Random Fields (*CRF*) algorithm. And a French-Vietnamese bilingual dictionary is used to make references the results of a word translation from a French sentence to its corresponding word in Vietnamese sentence.

[1] http://www.lattice.cnrs.fr/sites/itellier/SEM.html/

We identify the phrases chunking in the French corpus. Then, we predict, through the results of words translated via French-Vietnamese bilingual dictionary, the boundaries of phrase chunking in the Vietnamese sentence corresponding to each phrase chunking in the French sentence. The disambiguation of the phrase chunking boundaries in the Vietnamese sentences will be resolved on the basis of the position of adjacent translated words in the corresponding segment of the French sentences.

The process consists of five following steps:

Step 1: The French-Vietnamese bilingual corpus will be separated into two monolingual corpus. Next, we continue the phase of phrase chunking for the French sentences. The segments which words number is less than the threshold value θ, will be grouped with the adjacent segments. These groups will be labeled with order number in ascending order.

Step 2: Reference in the French-Vietnamese bilingual dictionary in order to find, in the Vietnamese sentence, words which are the translation of the words in the corressponding French sentence. If there are the correct translations between them then these Vietnamese words will be tagged the same order number label as their corresponding French words.

Step 3: Disambiguate the Vietnamese words which have multiple order number labels based on the position of the translation of the adjacent words in the corresponding French sentence.

Step 4: Establish the phrases chunking in the Vietnamese corpus.

Step 5: Create the French-Vietnamese chunk aligned sentences pairs based on the overlapping order number labels in the French-Vietnamese bilingual sentence pairs.

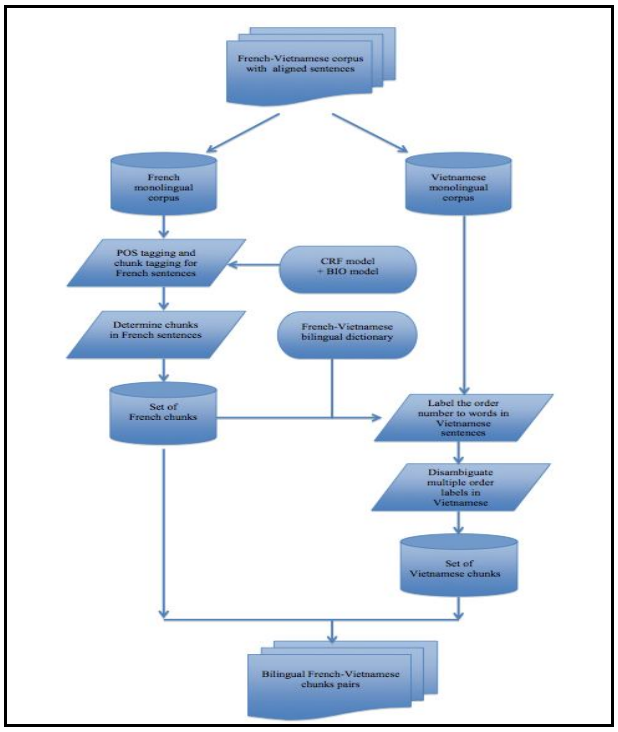

Fig.1 The model architecture of the chunks alignement system.

3.2 Chunk tagging for French sentences

Phrase chunking is a set of fragment of several words or phrases and it creates the syntactic structure of a sentence. To perform the French chunk tagging, we implement the Conditional Random Fields (*CRF*) model introduced by [6]. This model is a probabilistic model that allows to annotate the linear sequence data. Moreover, it allows to combine an observation x with a label y based on a set of examples which has been already tagged (x, y). In this case:

Given that $x = (x_1, x_2, x_3, ..., x_k)$ is a set of input data or in other words, x is a sequence of lexical units corresponding to a part-of-speech (*POS*) tag.

And given that $y = (y_1, y_2, y_3, ... , y_k)$ is a set of states or in other words, y is a sequence of BIO (*Beginning*, *In*, *Out*) labels corresponding to each phrase chunk. The CRF model defines the conditional probability of a state sequence, knowing that a given input string, with the following formula:

$$p(y/x) = \frac{1}{Z(x)} \prod_{c \in \varsigma} \exp(\sum_{k} \lambda_k f_k(y, x, c)) \quad (1)$$

In which :

Z(x) is the normalized coefficient, is defined that the total on y of all the probabilities p(y / x) for a given value of x which is assigned to 1 in this case.

z is a set of child elements in y. These elements include either a single element or a pair of adjacent elements.

f_k is the features function which is defined in each child element c. This is often used to selected to return the binary value 0 or 1. By definition, the value of the features function may depend on the labels of y existing in any child element c as well as the value of the POS tag label x in the input data.

λ_k is the weight in the position k, to be used to adjust the optimal value for each features function f_k.

In the French corpus, each word, after the POS tagging, will be annotated with chunk tags combined with the label followed BIO (*Beginning, In, Out*) model. The BIO model allows to annotate chunk tags. A phrase consists of several words.

The first word will be labeled “B”, followed by “I”. The “O” label is assigned to words of any phrase located individually in a sentence. In Figure 2, we have the following parsing tree:

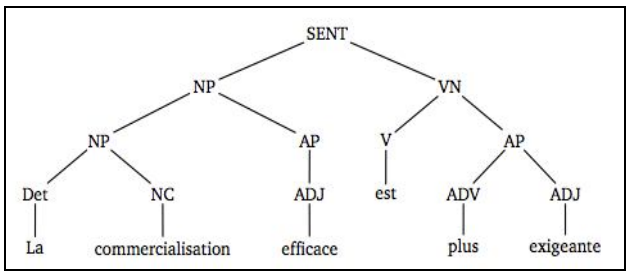

Fig. 2 A parsing tree extracted from the French Tree Bank corpus.

From the parsing tree above, we can have two cases of chunk tagging:

(a) (La commercialisation efficace)NP est plus exigeante.
(b) (La commercialisation efficace)NP (est)VN (plus exigeante)AP.

With the BIO model, we will have these results :
(a’) La/B-NP commercialisation/I-NP efficace/I-NP est/O plus/O exigeante/O.
(b’) La/B-NP commercialisation/I-NP efficace/I-NP est/B-VN plus/B-AP exigeante/I-AP.

The tool we used is SEM. It is a POS tagging tool and also chunk tagging for French corpus. SEM is a set of annotation which is trained on French Tree Bank (Abeillé, 2003) [2] of the University of Paris 7, France.

3.3 Extraction of French phrase chunking

A phrase is considered as true if and only if its boundaries and its category are true. The extraction of phrase chunks is not an easy job. That needs a phase to determine the accuracy of a phrase in the whole sentence.

By combining the BIO model and the Conditional Random Fields algorithm - with the training set of French Tree Bank of the University of Paris 7, France - we have the baseline chunking model for French. We found that if we only use this baseline chunking model, the result is very low. That led to the low efficiency of the disambiguity. Therefore we set an order number as a label for all words in a phrase from left to right. Next, based on the boundaries of phrase chunks, they are grouped together according to the minimum number of words in the phrase (*threshold*), for example with a threshold θ = 3 in the following figure:

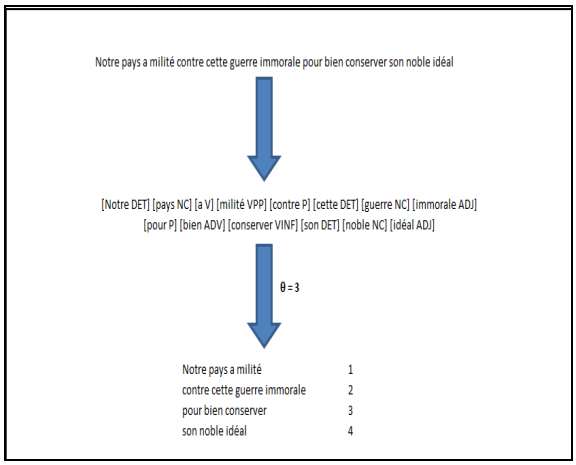

Fig. 3 Example for the extraction of phrase chunking in French corpus with a threshold θ = 3.

3.4 Determine the boundaries of Vietnamese chunks

In the opinion of [3], when translating a English sentence to a Chinese sentence, the words in a English chunk tend to be translated into a group of adjacent Chinese words. And that also exists in the French-Vietnamese language pair. This means that a French chunk will be aligned with a Vietnamese chunk based on words which are considered as the translation of each other by referencing the French-Vietnamese bilingual dictionary.

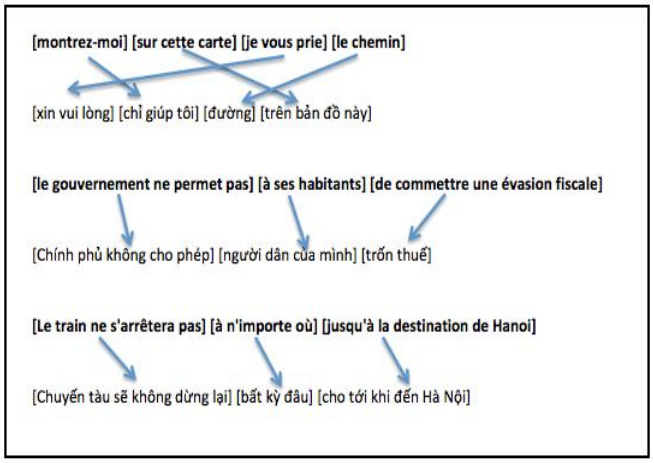

Fig. 4 Some examples about the French-Vietnamese chunks alignment.

Thus, to determine the Vietnamese phrase chunking corresponding with the French phrase chunking, firstly, we label with the order number the French phrase chunking based on the CRF tagging model in combination with the BIO model. In the Vietnamese phrase chunking, we verify if a word is considered a translation of a word in French sentence, it will be annotated with the same order number label as the word in the French sentence. However, in the case that a Vietnamese word has multiple order number labels, we disambiguate by considering the boundaries such as the order number label of the adjacent words. Finally, the Vietnamese words which have the same label in a sentence will be grouped into phrases chunks.

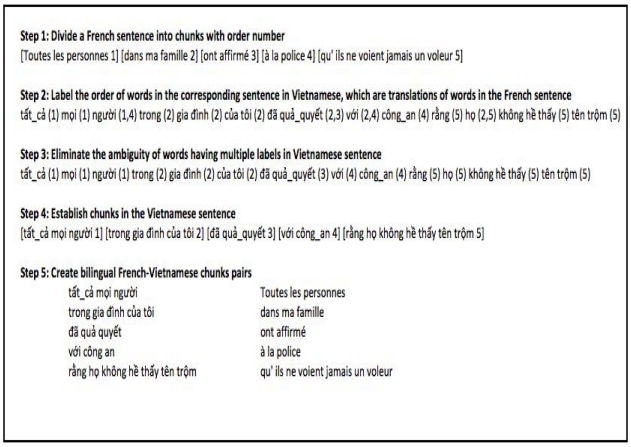

Fig. 5 Our algorithm of extraction for the French-Vietnamese bilingual chunks pairs.

4. Experiments and Evaluation

4.1 Preparation of corpus

For the experiments, we use a French-Vietnamese bilingual corpus composed of 10,000 French-Vietnamese bilingual sentences pairs. In fact, we collected from the conversational French textbook and bilingual French-English dictionary with over 90,000 entries.

The tool used to POS tagging and also chunk tagging for French corpus is SEM. It is a set of annotation which is trained on French Tree Bank (Abeillé, 2003) [2] of the University of Paris 7, France.

This corpus is normalized according to the following criteria :

- Uniformed in terms of content.
- Uniformed in terms of form : each sentence is on a single line and end with a punctuation.
- Spell checked.
- Removed the overlapping sentences in the corpus.

The length of the sentences and the segment is from 1 word to 20 words.

4.2 Results

We evaluated, by hands, the performance of the chunks alignment tool for a French-Vietnamese bilingual corpus by calculating the precision, the recall and by asking help from a linguistic expert.

We have calculated the precision based on a test sample with 100 bilingual sentences pairs that has been aligned over chunks. With this data set for test, we evaluated the accuracy and the recall of the system by adjusting the threshold θ from 1 to 6 in order to determine the optimal threshold values. The table 1 below shows the precision values when adjusting the threshold.
From the precision and the recall, we can calculate the coefficient F-measure with the following formula:

$$Precision = \frac{\sum Correct\ Pairs}{\sum Found\ Pairs} \quad Recall = \frac{\sum Found\ Pairs}{\sum Total\ Pairs} \quad Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (2)$$

Table 1: Results of chunks alignment for French-Vietnamese language pair

θ	Correct	Found	Total	Precision	Recall	$F_{measure}$
1	250	294	302	85,03%	**97,35%**	90,78%
2	209	231	249	**90,48%**	92,77%	**91,61%**
3	172	192	219	89,58%	87,67%	88,62 %
4	147	165	200	89,09%	82,50%	85,67%
5	141	157	196	89,81%	80,10%	84,68%
6	142	158	195	89,87%	81,03%	85,22%

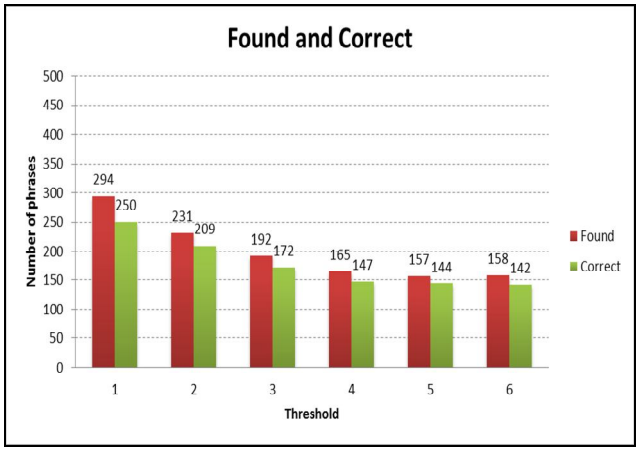

Fig. 6 Results of Found and Correct.

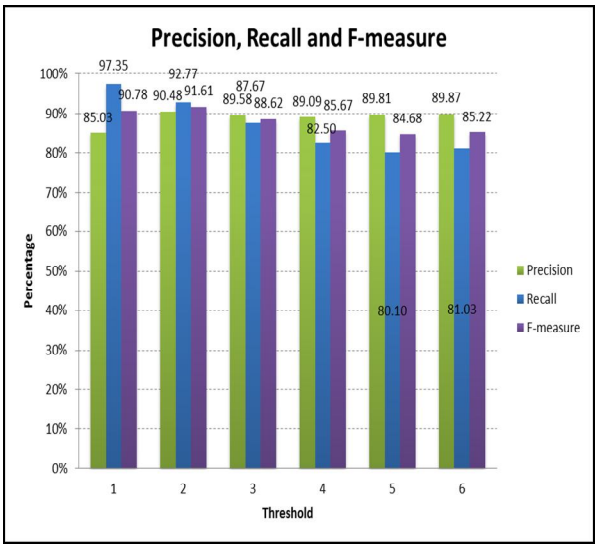

Fig. 7 Results of other parameters – Precision, Recall and Fmeasure.

Remark 1: We found differences in the number of chunks pairs in French text and in Vietnamese text on a sample of 100 bilingual sentences pairs when tuning the value of the threshold θ. The threshold θ is defined such as the smallest number of a phrase during chunks alignment process. By observing the results of chunk tagging in combination with the BIO model, an error on the determination of the boundaries of a chunk involves two chunks segments either deleted (*omission*) or inserted (*insertion*) instead we will have exactly two bilingual chunks segments.

Remark 2: With the results obtained and presented in the table above, we notice that when we align over chunks for the French-Vietnamese language pair, with the threshold θ = 2, the precision and the recall of the process achieve the best results greater than 90%, and the F-measure is 91.61%.

5. Conclusions and perspective

Thus, in this paper, we introduce a new approach about splitting a sentence in several chunks and aligning over chunks for the French-Vietnamese bilingual sentences pairs in order to solve the problem of long sentences. With the experiments of our chunks alignment system, we calculated the precision more than 90%, and the F-measure 91.61%. These are encouraging results

in the research for the French-Vietnamese language pair in Vietnam in particular and in the world in general.

Moreover, one of the advantages of our model is that it will can be easily applied to different language pairs. And the requirements are just lexical information, such as bilingual dictionaries for the language pairs, a list of the function words and we have to change a few parameters for the system configuration.

References

[1] Isabelle Tellier, Denys Duchier, Iris Eshkol, Arnaud Courmet, Mathieu Martinet: Apprentissage automatique d'un chunker pour le français, Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, Grenoble, 4 au 8 Juin 2012 Copyright 2012 ATALA & AFCP, pp. 431-438 (2012).

[2] A.Abeillé, L.Clément et F.Toussenel: Building a treebank for french. In A.ABEILLE, éditeur: Treebanks. Kluwer, Dordrecht (2003).

[3] Sun Le, Jin Youbing, Du Lin, Sun Yufang: Word Alignment of English-Chinese Bilingual Corpus Based on Chunks. Proc. 2000 EMNLP and VLC (2000), pp.110-116 (2000).

[4] Vinh Van Nguyen, Thai Phuong Nguyen, Akira Shimazu and Minh Le Nguyen: Reordering Phrase-based machine translation over chunks. IEEE International Conference on Research, Innovation and Vision for the Future, 2008. RIVF 2008, pp.114-119 (2008).

[5] Francisco Nevado, Francisco Casacuberta, Enrique Vidal: Parallel Corpora Segmentation Using Anchor Words. In Proc. of the EAMT/EACL Workshop on MT and Other Language Technology Tools, Budapest, Hungary, April, pp. 12-17 (2003).

[6] Lafferty, McCallum, and Pereira: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In the proceedings of International Conference on Machine Learning (ICML), pp. 282-289 (2001).

[7] Thi-Ngoc-Diep Do ,Viet-Bac Le, Brigitte Bigi, Laurent Besacier, Eric Castelli: Mining a comparable text corpus for a Vietnamese - French statistical machine translation system. Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 30 March – 31 March 2009, pp. 165–172 (2009).

[8] Doan N.H. Generation of Vietnamese for French-Vietnamese and English-Vietnamese machine translation. In proceedings of European workshop on Natural Language Generation 2001 (2001).

[9] LE Hong Phuong: TAG – Tree Adjoining Grammar. PhD thesis (2009).

[10] Nicolas BOFFO : Formation de la temporalité en Vietnamien pour la traduction automatique. PhD thesis in process, France-Vietnam.

[11] NGUYEN Thi Minh Huyen : Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamien. PhD thesis (2006).

[12] TRAN Tuan Duc, DYALANG, Université de Rouen, France: Système de recherche d'information medical par croisement de langue : Vietnamien–Français–Anglais. GLOTTOPOL, numéro 8 (2006).

Ngoc Tan LE is a lecturer in Department of Computer Science in Industrial University of Ho Chi Minh, Vietnam. He has graduated Master in 2009 from the University of Lyon 1, France. His current research interests include Vietnamese-related NLPs using machine learning of linguistics knowledge from French-Vietnamese bilingual corpora.

Ngoc Tien LE is a lecturer in Department of Computer Science in Industrial University of Ho Chi Minh, Vietnam. He has graduated Master in 2008 from the University of Natural Science of Ho Chi Minh, Vietnam. His current research interests include Machine Learning, Machine Translation, Named Entity Recognition.

Dien DINH is associate professor in Computer Science. He is actually deputy head of Knowledge Engineering Department in University of Natural Sciences, VNU-HCMC, Vietnam. He received the Ph.D. degree in Linguistics in 2005 from the University of Social Sciences & Humanity, VNU-HCMC and Ph.D. degree in Computer Sciences in 2002 from the University of Natural Sciences, VNU-HCMC. His research interests include Vietnamese-related NLPs using machine learning of linguistics knowledge from English-Vietnamese bilingual corpora.