Building an English-Vietnamese Bilingual Corpus for Machine Translation

Conference Paper · November 2012

DOI: 10.1109/IALP.2012.30

Hung Quoc Ngo
University of Information Technology, Vietnam National University - HCMCity

Building an English-Vietnamese Bilingual Corpus for Machine Translation

Quoc Hung Ngo
Faculty of Computer Science
University of Information Technology
HoChiMinh City, Vietnam
hungnq@uit.edu.vn

Werner Winiwarter
University of Vienna
Research Group Data Analytics and Computing
Universitätsstraße 5, 1010 Vienna, Austria
werner.winiwarter@univie.ac.at

***Abstract*—Bilingual corpora are critical resources for machine translation research and development since parallel corpora contain translation equivalences of various granularities. Manual annotation of word alignments is of significance to provide a gold-standard for developing and evaluating both example-based machine translation models and statistical machine translation models. This paper presents research on building an English-Vietnamese parallel corpus, which is constructed for building a Vietnamese-English machine translation system. We describe the specification of collecting data for the corpus, linguistic tagging, bilingual annotation, and the tools specially developed for the manual annotation. An English-Vietnamese bilingual corpus of over 800,000 sentence pairs and 10,000,000 English words as well as Vietnamese words has been collected and aligned at the sentence level, and over 45,000 sentence pairs of this corpus have been aligned at the word level.**

Keywords—English-Vietnamese corpus, word alignment, linguistic tagging, bilingual annotation

I. INTRODUCTION

In natural language processing, a bilingual corpus is a valuable resource. A huge bilingual corpus is not only used to train natural language processing (NLP) tasks effectively but also to evaluate NLP systems objectively, such as chunking in bilingual text, bilingual comparison, bitext transfer, and machine translation.

Because of the importance of bilingual corpora, there are many projects on corpus acquisition for language pairs and even multi-languages, such as Multilingual Parallel Corpus JRC-ACQUIS[1], for European languages, the Chinese-English corpus of Xu Xunfeng (HK PolyU), the JEFLL Corpus Project[2], of Yukio Tono (Tokyo University of Foreign Studies), the English-Thai Bitext Corpus of Doug Cooper[3], etc.

For the English-Vietnamese language pair, there are several projects for building an English-Vietnamese corpus for particular purposes, such as building a bilingual corpus for word sense disambiguation by Dinh Dien [2,3], building a bilingual corpus through web mining by Van Bac Dang and Bao Quoc Ho [14], and a downloadable bilingual corpus of the VLSP project[4]. However, most of these corpora are not available for download or just at the aligned sentence level.

In this paper, we demonstrate a procedure to build an English-Vietnamese Bilingual Corpus (EVBCorpus). More specifically, the goal is to build and annotate a large bilingual corpus which is tagged with linguistic information, such as part-of-speech, chunks, and bitext alignment at the word level and more. This bilingual corpus can then be used in the automatic training of machine translation systems and for evaluating bitext alignment.

Fig. 1 shows the main modules of bilingual corpus building, including three main modules: pre-processing, linguistic tagging, and bilingual annotation. In particular, the pre-processing module contains formatting data, paragraph alignment, and sentence alignment. The result of the preprocessing module is English-Vietnamese sentence pairs. These bilingual pairs are tagged linguistically in the tagging modules, including English chunking, Vietnamese chunking, and English-Vietnamese word alignment. The aligned source and target chunks can be corrected as chunking result, alignment result as well as Vietnamese word segmentation result at the bilingual annotation stage.

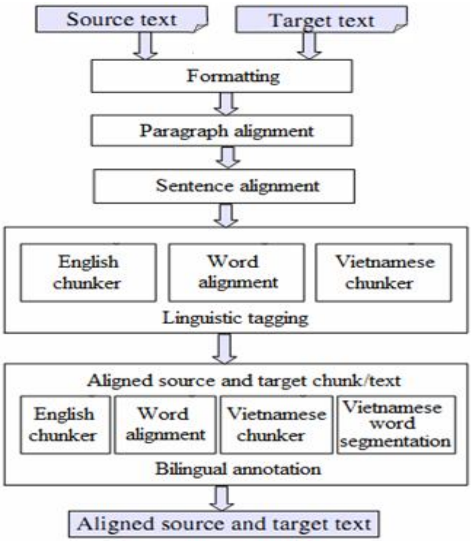

Figure 1: Modules for corpus building

II. DATA

A. Data source

The EVBCorpus consists of both original English text and its Vietnamese translations, and original Vietnamese text and its English translations. The original data is from books, fictions or short stories, law documents, and

[1] http://optima.jrc.it/Acquis/

[2] http://jefll.corpuscobo.net/

[3] http://crcl.th.net/bitext/

[4] http://vlsp.vietlp.org:8080/demo/?page=resources

978-0-7695-4886-9/12 $26.00 © 2012 IEEE
DOI 10.1109/IALP.2012.30

newspaper articles. The original articles were translated by skilled translators or by contribution authors and were checked again by skilled translators. Parallel documents are also chosen and classified into categories, such as economy, entertainment (art and sport), health, science, social and politics, and technology.

Each article was translated one to one at the whole article level, so we first need to align paragraph to paragraph and then sentence to sentence. At the paragraph stage, aligning is simply moving the sentences up or down and detecting the separator position between paragraphs of both articles. At the sentence stage, however, aligning is more complex and it depends on the translated articles which are translated by one-by-one method or a literal meaning-based method. In many cases (as common in literature text), several sentences are merged into one sentence to create the one-by-one alignment of sentences. The details of the corpus are listed in Table 1.

TABLE 1: DETAILS OF DATA SOURCES OF EVBCORPUS

Source	Document	Paragraph	Sentence	Word[a]
En-Vn Books	15	13,980	80,323	1,375,492
En-Vn Fictions	100	192,723	590,520	6,403,511
En-Vn Laws	250	86,803	98,102	1,912,055
En-Vn News	1,000	24,523	45,531	740,534
Total	**1,365**	**318,029**	**814,476**	**10,431,592**

a. Number of words is counted on the English text.

An important feature of the corpus is that it has been pre-processed at the basic linguistic level, namely that of words. Especially, in Vietnamese, tokens are not words and a word can be a token or a group of tokens. Therefore, the first important step in pre-processing is a Vietnamese word segmentation which is just done to evaluate the corpus, whereas this step used for later processing is included the Vietnamese chunking module. In our project, we use vnTokenizer[5] of L. H. Phuong et al [9] to segment words in Vietnamese text.

TABLE 2: CHARACTERISTICS OF THE ENGLISH-VIETNAMESE BILINGUAL CORPUS

	English	Vietnamese
Sentences	814,476	
Words	10,431,592	10,272,514
Tokens	(10,431,592)	12,973,600
Vocabulary	32,881	23,725
Aver. Sentence length	12.5	12.5

There are 10,431,592 English words and 10,272,514 Vietnamese words (containing 12,973,600 Vietnamese tokens) in our bilingual corpus (see Table 2). The vocabulary used in this whole corpus makes up nearly 50 percent of the number of English words in an English-Vietnamese dictionary (containing 65,000 words). Similarly, the Vietnamese vocabulary makes up 40 percent of the number of Vietnamese words in the Vietnamese-English dictionary. Vietnamese words are counted based on the result of using the vnTokenizer module on the Vietnamese text.

B. Data Standardization

One task in pre-processing a corpus is standardizing data. The important punctuation marks for conversation sentences (direct speech) in English and Vietnamese are not similar. Usually, Vietnamese conversation sentences start with a minus sign while English ones start and end with a quotation mark, for example:

English sentence: "Very pretty." he said.

Vietnamese sentence: - Màu đẹp lắm. - anh ta nói.

In our experiment, we change the Vietnamese format for conversations to the English format, which means using quotation marks for conversations in both languages.

Standardizing Vietnamese punctuation stress marks is also a necessary step of pre-processing. In several Vietnamese words with two vowel sounds, the stress mark can be put on the first sound or the second sound. For example, *hòa* and *hoà* are correct and acceptable for Vietnamese language and they are only one word, however, for computing, they are different, which makes the corpus more ambiguous. Usually, stress marks are put on the middle characters, so we put the stress marks according to this rule.

III. LINGUISTIC TAGGING

For linguistic tagging, we tag chunks for both English and Vietnamese text. English-Vietnamese sentence pairs are also aligned word-by-word to create the connections between the two languages.

A. Chunking for English

There are several available chunking systems for English text, such as CRFChunker[6] by Xuan-Hieu Phan, OpenNLP[7] (which is an open source NLP project and one of SharpNLP's modules) of Jason Baldridge et al. However, we focus on parser modules to build an aligned bilingual treebank in future. Based on Rimell's evaluation of five state-of-the-art parsers [13], the Stanford parser is not the parser with the highest score. However, the Stanford parser supports both parse trees in bracket format and dependencies representation [1, 10]. We chose the Stanford parser not only for this reason but also because it is updated frequently, and to provide for the ability of our corpus for semantic tagging in future.

In our project, the full parse result of an English sentence is considered to extract phrases as chunking result for the corpus. For example, for the English sentence "*Products permitted for import, export through Vietnam's border-gates or across Vietnam's borders.*", the extracted chunks based on the Stanford parser result are:

[5] http://www.loria.fr/~lehong/tools/vnTokenizer.php

[6] http://crfchunker.sourceforge.net/

[7] http://sharpnlp.codeplex.com/

[Products]$_{NP}$ [permitted]$_{VP}$ [for]$_{PP}$ [import]$_{NP}$, [export]$_{NP}$ [through]$_{PP}$ [Vietnam's border-gates]$_{NP}$ [or]$_{PP}$ [across]$_{PP}$ [Vietnam's borders]$_{NP}$.

B. Chunking for Vietnamese

There are several chunking systems for Vietnamese text, such as noun phrase chunking of Le Minh Nguyen et al [7] or full phrase chunking of Nguyen Huong Thao et al [11]. In our system, we use the phrase chunker of Le Minh Nguyen and Hoang Tru Cao [8] to chunk Vietnamese sentences. This is module SP8.4 in the VLSP project.

The VLSP project is a KC01.01/06-10 national project named Building Basic Resources and Tools for Vietnamese Language and Speech Processing. This project involves active research groups from universities and institutes in Vietnam and Japan, and focuses on building a corpus and toolkit for Vietnamese language processing, including word segmentation, part-of-speech tagger, chunker, and parser.

The chunking result also includes the word segmentation and the part-of-speech tagger result. These results are based on the result of word segmentation by Le Hong Phuong, Nguyen Thi Minh Huyen et al [9]. The tagset of chunking includes 5 tags: NP, VP, ADJP, ADVP, and PP.

C. Word Alignment in Bilingual Corpus

In a bilingual corpus, word alignment is very important because it demonstrates the connection between two languages. In our corpus, we apply a class-based word alignment approach to align words in the English-Vietnamese pairs. Our approach is based on the result of Dinh Dien et al [4]. This approach originates from the English-Chinese word alignment approach of Ker and Chang [6]. The class-based word alignment approach uses two layers to align words in a bilingual pair, dictionary-based alignment and semantic class-based alignment. The dictionary used for the dictionary-based stage is a general machine-readable bilingual dictionary while the dictionary used for the class-based stage is the Longman Lexicon of Contemporary English (LLOCE) dictionary, which is a type of semantic class dictionary.

The result of the word alignment is indexed based on token positions in both sentences. For example:

English: I had rarely seen him so animated .
Vietnamese: Ít khi tôi thấy hắn sôi nổi như thế .

The word alignment result is [1-3], [3-1,2], [4-4], [5-5], [6-8,9], [7-6,7], [8-10] (visualized in Fig. 2).

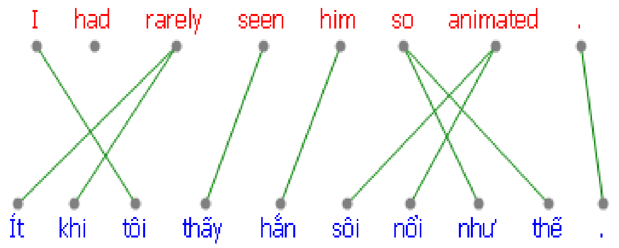
I had rarely seen him so animated .
Ít khi tôi thấy hắn sôi nổi như thế .

Figure 2: An example of word alignment in bilingual corpus

IV. BILINGUAL ANNOTATION

In our project, we build an annotation application, BiCA2012, which is a tool for annotating a corpus visually, quickly, and effectively [12]. This tool has two main annotation stages:

- **Bitext Alignment**: This first stage of annotation is a bitext alignment, which aligns paragraph by paragraph and then sentence by sentence.
- **Word Alignment:** This stage based on combining English chunking, Vietnamese chunking, and word alignment results in aligning a parallel sentence pair at the chunk level (see Fig. 3).

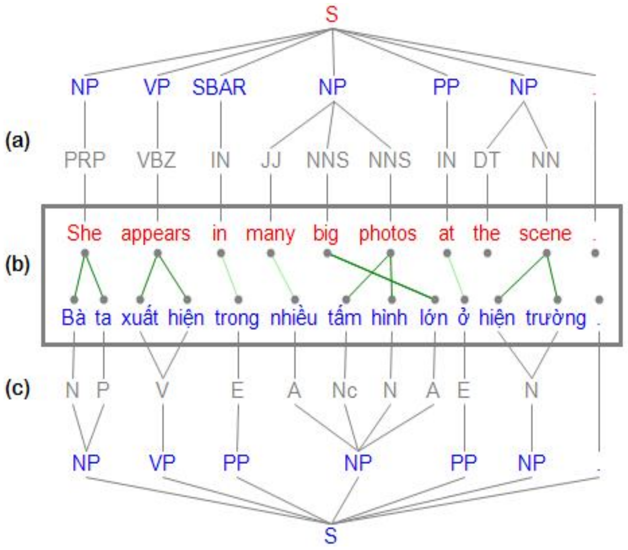
S
NP VP SBAR NP PP NP .
(a)
PRP VBZ IN JJ NNS NNS IN DT NN
(b)
She appears in many big photos at the scene .
Bà ta xuất hiện trong nhiều tấm hình lớn ở hiện trường .
(c)
N P V E A Nc N A E N
NP VP PP NP PP NP .
S

Figure 3: Combine English chunking (a), Vietnamese chunking(c), and word alignment (b)

With the visualization provided by our tool, annotators review whole phrase structures of English and Vietnamese sentences. They can compare the English chunking result with the Vietnamese result and correct them in both sentences. Moreover, mistakes regarding word segmentation for Vietnamese, POS tagging for English and Vietnamese, and English-Vietnamese word alignment can be detected and corrected through drag, drop, and edit label operations (actions) of our tool. Based on drag and drop on labels and tags, annotators can change the results of the tagging modules visually, quickly, and effectively.

V. RESULTS AND ANALYSIS

A. Bilingual Corpus

As a main result of our project, we have built an English-Vietnamese bilingual corpus with 1,365 documents, over eight hundred thousand sentences, and over ten million words from four resources: books, literal novels, law documents, and news articles. As mentioned in Section II, all of these documents are collected and aligned as chapter-to-chapter (for books, novels, and laws), or article-to-article (for news articles) first. Next, they are semi-automatically separated to align at the paragraph level, and at the sentence level at last. However, we still keep the context of paragraphs and sentences, which is very useful for other tasks in several machine translation models, such as document classification before translating or detecting the context of words in documents. This corpus also standardizes data at the sentence level, such as punctuation marks for conversation sentences or stress marks for Vietnamese text.

The English-Vietnamese bilingual corpus is stored in both standard formats, the HTML format for processing at next steps by our tool or previewing on web browsers, and the SGML format for sharing this data with other tasks. A part of this corpus and the annotation tool are published at http://code.google.com/p/evbcorpus/ .

B. Aligned Bilingual Corpus

The annotation process costs a lot of time and effort, especially with a corpus of over 10 million words of each language. In our evaluation, we annotated 1000 news articles of EVBNews with 45,531 sentence pairs, and 740,534 English words (832,441 Vietnamese words and 1,082,051 Vietnamese tokens), as shown in Table 3. The data is tagged and aligned automatically at the word level between English and Vietnamese.

In this evaluation, we just focus on the set of alignments and amount of annotation rather than evaluate the quality of the Word Alignment module. The number of corrected alignments is 10% higher than the number of original alignments, which are aligned automatically by the Word Alignment module. However, apart from 10% added alignments as new alignments, automatically linked alignments are also modified and corrected by annotators.

TABLE 3: NUMBER OF ALIGNMENTS IN 1,000 NEWS ARTICLES

	English	**Vietnamese**
Files	1,000	1,000
Sentences	45,531	45,531
Words	740,534	832,441
Sure Alignments	447,906	447,906
Possible Alignments	560,215	560,215
Words in Alignments	654,060	768,031

Alignments are annotated with both sure alignments S and possible alignments P, with S $\subseteq$ P. These two types of alignments are annotated to evaluate the alignment models with the Alignment Error Rates (AER) [5]. In 1,000 aligned news articles, there are 447,906 sure alignments, accounting for 80% of 560,215 possible alignments (as shown in Table 3). These sure alignments mainly come from nouns, verbs, adverbs, and adjectives which are meaningful words in sentences. On the other hand, the 20% remaining possible alignments are mainly from prepositions in both English words and Vietnamese words.

VI. CONCLUSION

In this paper we have introduced a complete workflow to build an English-Vietnamese bilingual corpus, from collecting data, tagging chunks, aligning words in bilingual text, and developing an annotation tool for bilingual corpora. We showed that the size of our corpus with over 800,000 English-Vietnamese aligned pairs at sentence level is a valuable contribution to build a good corpus in the future. We pointed out that linguistic information tagging based on our procedure, including tagging and annotation, so far, stops at the chunk level.

However, one potential model of full parser alignment is to combine full parse trees and word or chunk alignments; we plan to address this in future work. We also plan to completely prepare the corpus with all sentence pairs aligned and corrected semi-automatically at the chunk level. In addition, 45,531 aligned sentence pairs have been also used to map linguistic tags (such as named entities, co-reference chunks) from English to Vietnamese text to build Vietnamese corpora semi-automatically.

REFERENCES

[1] Dan Klein and Christopher D. Manning (2003). Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

[2] Dinh Dien, (2002). Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation. In Proceedings of Workshop on Machine Translation in Asia, pp. 26-32, 2002.

[3] Dinh Dien, Hoang Kiem (2003). POS-tagger for English-Vietnamese bilingual corpus. In Proceedings of the workshop on building and using parallel texts: data driven machine translation and beyond, Edmonton, Canada, pp 88–95.

[4] Dinh Dien, Hoang Kiem, Thuy Ngan, Xuan Quang, Van Toan, Quoc Hung-Ngo, Phu Hoi (2002). Word alignment in English – Vietnamese bilingual corpus. Proceedings of EALPIIT'02, HaNoi, Vietnam, pg 3-11.

[5] Franz Josef Och, Hermann Ney (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29:19–51.

[6] Ker Sue J. and Jason S. Chang (1997). A class-based approach to word alignment. Computational Linguistics, 23(2):313–343.

[7] Le Minh Nguyen, Huong Thao Nguyen, Phuong Thai Nguyen, Tu Bao Ho and Akira Shimaz (2009). An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models. The 7th Workshop on Asian Language Resources (In Conjunction with ACL-IJCNLP).

[8] Le Minh Nguyen, Hoang Tru Cao (2008), Constructing a Vietnamese Chunking System. In Proc. of the 4th National Symposium on Research, Development and Application of Information and Communication Technology, Science and Technics Publishing House, 249-257.

[9] Le Hong Phuong, Nguyen Thi Minh Huyen, Roussanaly Azim, H. T. Vinh (2008). A hybrid approach to word segmentation of Vietnamese texts. Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, LATA 2008, Springer LNCS 5196, Tarragona, Spain.

[10] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In LREC 2006.

[11] Nguyen Huong Thao, Nguyen Phuong Thai, Le Minh Nguyen, and Ha Quang Thuy (2009). Vietnamese Noun Phrase Chunking based on Conditional Random Fields. In Proceedings of The First International Conference on Knowledge and Systems Engineering (KSE 2009).

[12] Quoc Hung-Ngo, Werner Winiwarter (2012). A Visualizing Annotation Tool for Semi-Automatically Building a Bilingual Corpus, In Proceedings of the 5th Workshop on Building and Using Comparable Corpora, LREC2012 Workshop, pp. 67-74.

[13] Rimell, L., S. Clark, and M. Steedman (2009). Un-bounded dependency recovery for parser evaluation. In Proceedings EMNLP, pp. 813–821.

[14] Van Bac Dang, Bao Quoc Ho (2007). Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining. Research, Innovation and Vision for the Future (RIVF), IEEE International Conference. pp. 261-266.