EVBCorpus - A Multi-Layer English-Vietnamese Bilingual Corpus for Studying Tasks in Comparative Linguistics

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
Währinger Straße 29, 1090 Wien, Austria
werner.winiwarter@univie.ac.at

Bartholomäus Wloka
University of Vienna, Research Group Data Analytics and Computing
Austrian Academy of Sciences, Institute for Corpus Linguistics and Text Technology
Währinger Straße 29, 1090 Wien, Austria
bartholomaeus.wloka@univie.ac.at

Abstract

Bilingual corpora play an important role as resources not only for machine translation research and development but also for studying tasks in comparative linguistics. Manual annotation of word alignments is of significance to provide a gold-standard for developing and evaluating machine translation models and comparative linguistics tasks. 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. Moreover, the 45,000 sentence pairs have been tagged using other linguistics tags, including word segmentation for Vietnamese text, chunker and named entity tags.

1 Introduction

Recent years have seen a move beyond traditionally inline annotated single-layered corpora towards new multi-layer architectures, deeper and more diverse annotations. There are several studies which are background for building multi-layer corpora. These studies include building tools (A. Zeldes et al., 2009; C. Muller and M. Strube, 2006; Q. Hung and W. Winiwarter, 2012a), annotation progress (A. Burchardt et al., 2008; Hansen Schirra et al., 2006; Ludeling et al., 2005), and data representation (A. Burchardt et al., 2008; Stefanie Dipper, 2005). Despite intense work on data representations and annotation tools, there has been comparatively less work on the development of architectures affording convenient access to such data.

Moreover, several research works have been carried out to build English-Vietnamese corpora at many different levels, for example, a study on building POS-tagger for bilingual corpora or building a bilingual corpus for word sense disambiguation of Dinh Dien and co-authors (D. Dien, 2002a; D. Dien et al., 2002b; D. Dien and H. Kiem, 2003). Other research efforts for this language pair are building English-Vietnamese corpora (B. Van et al., 2007; Q. Hung et al., 2012b; Q. Hung and W. Winiwarter, 2012c).

The present paper shows the process of building a multi-layer bilingual corpus, including four main modules: (1) bitext alignment, (2) word alignment, (3) linguistic tagging, and (4) mapping and annotation (as shown in Figure 1). In particular, the bitext alignment (1) includes paragraph and sentence matching. This step also needs annotation to ensure that the result of this step are English-Vietnamese sentence pairs. These bilingual sentence pairs are aligned at the word

International Joint Conference on Natural Language Processing, pages 1–9,
Nagoya, Japan, 14-18 October 2013.

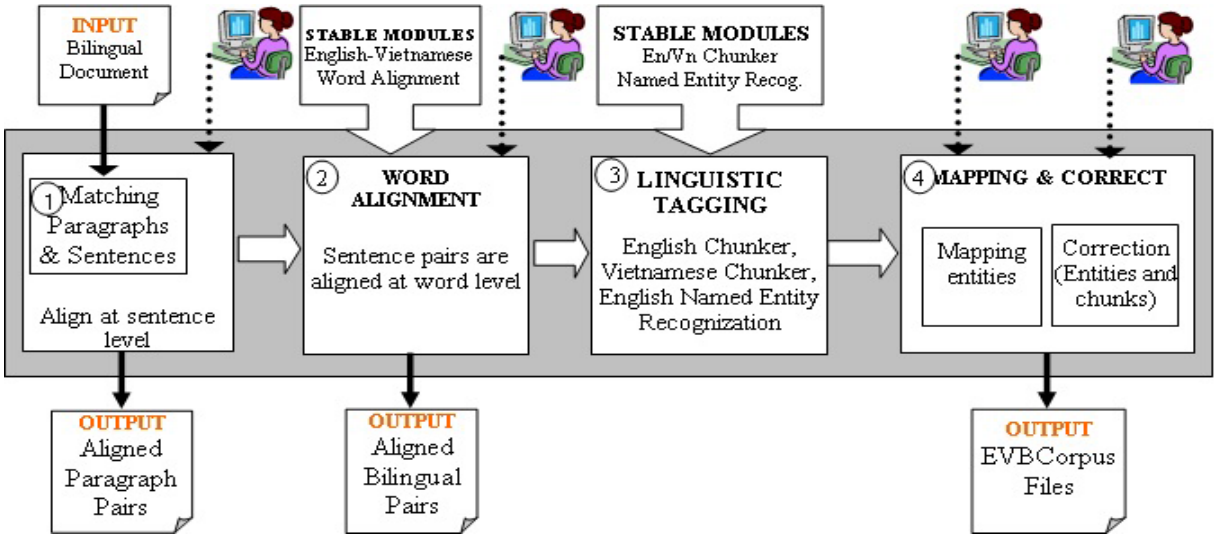

Figure 1: Overview of building EVBCorpus

level by a word alignment module (2). Then, these bilingual sentences are tagged linguistically and independently by the specific tagging modules (3), including English chunking, Vietnamese chunking, and Named Entity recognition. Finally, the aligned source and target text can be corrected as alignment result, word segmentation, chunking result, as well as named entity recognition result at the mapping and correction stage (4).

Moreover, we also suggest that annotating factors in a multi-layer corpus can afford corpus designers several advantages:

- Linguistics tagging for the corpus has to be carried out layer-by-layer based on specific tags and existing tagging tools.
- Distributing annotation work collaboratively, so that annotators can specialize on specific subtasks and work concurrently.
- Using different level annotation tools suited to different tasks in tagging linguistics tags.
- Allowing multiple annotations of the same type to be created and evaluated, which is important for controversial layers with different possible tag sets or low inter-annotator agreement.

The remainder of this paper describes the details of our approach to build a multi-layer bilingual corpus. Firstly we describe the data source for corpus building in Section 2. Next, we demonstrate a procedure for linguistic tagging and mapping English linguistic tags into Vietnamese tags in Section 3. Section 4 addresses the annotation process with the BiCAT tool. Conclusion and future work appear in Section 5.

2 Data Sources

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 newspaper articles. The original articles were translated by skilled translators or by contribution authors and were checked again by skilled translators. The details of the EVBCorpus corpus are listed in Table 1.

Table 1: Details of data sources of EVBCorpus

Source	Doc.	Sentence	Word
EVBBooks	15	80,323	1,375,492
EVBFictions	100	590,520	6,403,511
EVBLaws	250	98,102	1,912,055
EVBNews	1,000	45,531	740,534
Total	**1,365**	**814,476**	**10,431,592**

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 by using the BiCAT[1]

[1] https://code.google.com/p/evbcorpus/

tool, an annotation tool for building bilingual corpora (see Section 4 and Figure 7) (Q. Hung and W. Winiwarter, 2012a).

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 data source for multi-layer linguistic tagging is a part of the EVBCorpus which consists of both original English text and its Vietnamese translations. It contains 1,000 news articles defined as the EVBNews part of the EVBCorpus. This corpus is also aligned semi-automatically at the word level.

Table 2: Characteristics of EVBNews part

	English	Vietnamese
Files	1,000	1,000
Paragraphs	25,015	25,015
Sentences	45,531	45,531
Words	740,534	832,441
Words in Alignments	654,060	768,031

In particular, each article was translated one to one at the whole article level, so we align sentence to sentence. Then, sentences are aligned at the word level semi-automatically, including automatic alignment by class-based method and use of the BiCAT tool to correct the alignments manually. The details of the corpus are listed in Table 1 and Table 2.

Parallel documents are also chosen and classified into categories, such as economics, entertainment (art and music), health, science, social, politics, and technology (details of each category are shown in Table 3).

3 Linguistic Tagging

In our project, the corpus has four information layers, (1) word segmentation, (2) part-of-speech, (3) chunker, and (4) named entity tags (as shown in Figure 2).

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 (as shown in Figure 3).

Table 3: Number of files and sentences in each field

	File	Sentence
Economics	156	6,790
Entertainment	27	1,639
Health	253	13,835
Politics	141	4,520
Science	47	2,544
Social	108	4,075
Sport	22	962
Technology	137	4,778
Miscellaneous	109	6,388
Total	**1,000**	**45,531**

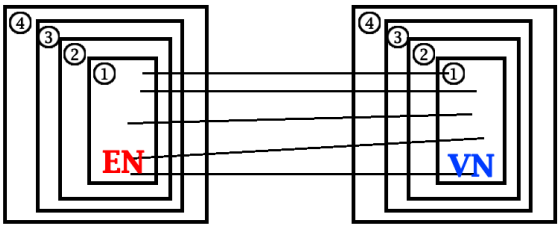

Figure 2: Multi-layer structure of aligned corpus files

3.1 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 and co-authors (D. Dien et al., 2002b). This approach originates from the English-Chinese word alignment approach of Ker and Chang (Sue Ker and Jason Chang, 1997). 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] and these alignments

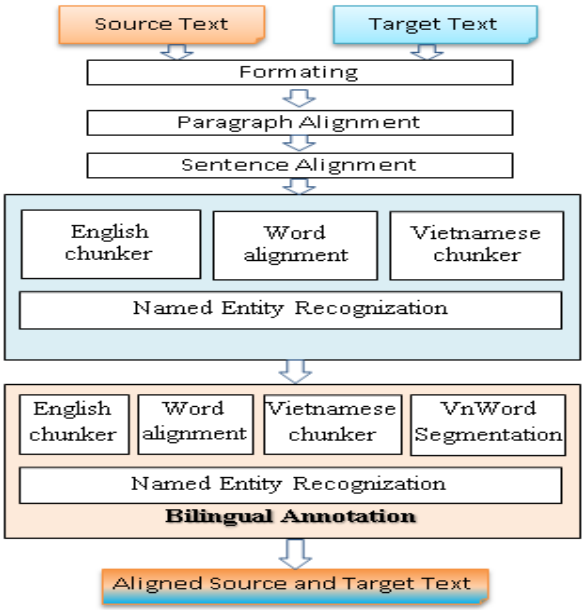

Figure 3: Modules for multi-layer corpus building

can be visualized word by word in Figure 4.

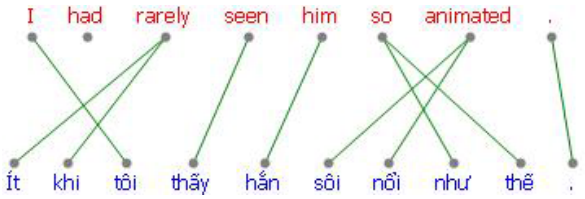

Figure 4: Example of word alignment

3.2 Chunking for English

There are several available chunking systems for English text, such as CRFChunker[2] by Xuan-Hieu Phan or OpenNLP[3] (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 tree-bank in future. Based on Rimell 's evaluation of 5 state-of-the-art parsers (Rimell et al., 2009), the Stanford parser is not the parser with the highest score. However, the Stanford parser[4] supports both parse trees in bracket format and dependencies representation (Dan Klein, 2003; Marneffe et al., 2006). 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.

[2] http://crfchunker.sourceforge.net/
[3] http://opennlp.apache.org/
[4] http://nlp.stanford.edu/software/lex-parser.shtml

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:

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

3.3 Chunking for Vietnamese

There are several chunking systems for Vietnamese text, such as noun phrase chunking of (Le Nguyen et al., 2008) or full phrase chunking of (Nguyen H. Thao et al., 2009). In our system, we use the phrase chunker of (Le Nguyen et al., 2009) to chunk Vietnamese sentences. This is module SP8.4 in the VLSP project.

The VLSP project[5] 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 H. Phuong et al., 2008). The tagset of chunking includes 5 tags: NP, VP, ADJP, ADVP, and PP.

For example, the chunking result for the sentence "Các sản phẩm được phép xuất khẩu, nhập khẩu qua cửa khẩu, biên giới Việt Nam." is [Các sản phẩm]$_{VP}$ [được]$_{VP}$ [phép]$_{NP}$ [xuất_khẩu]$_{VP}$, [nhập_khẩu qua]$_{VP}$ [cửa_khẩu]$_{NP}$, [biên_giới Việt_Nam]$_{NP}$." (see Figure 5).

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

3.4 Named Entity Recognition

Several Named Entity recognition systems for English text are available online. For traditional

[5] http://vlsp.vietlp.org:8080/demo/

Figure 5: Result of the Vietnamese chunking

NER, the most popular publicly available systems are: OpenNLP NameFinder[6], Illinois NER[7] system (Ratinov and Roth, 2009), Stanford NER[8] system by the NLP Group at Stanford University (Finkel et al., 2005), and Lingpipe NER[9] system by Aspasia Beneti and co-authors (A. Beneti et al., 2006). The Stanford NER reports 86.86 F1 on the CoNLL03 NER shared task data. We chose the Stanford NER to provide for the ability of our corpus for tagging with multi-type, such as 3 classes, 4 classes, and 7 classes.

For Vietnamese text, there are also several studies on Named Entity Recognition, such as Nguyen Dat and co-authors (Nguyen Dat et al., 2010) or Tri Tran and co-authors (Tran Q. Tri et al., 2007). However, there is no available system to download for tagging on Vietnamese text. In this project, therefore, we carry out mapping English named entities into Vietnamese text based on corrected English-Vietnamese word alignments to get basic Vietnamese named entities. These entities will be corrected by annotators in the next stage.

4 Annotation

In our project, we use an annotation tool, BiCAT, which is a tool for tagging and correcting a corpus visually, quickly, and effectively (Q. Hung and W. Winiwarter, 2012a). This tool has the following 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 allows annotators to modify word alignments between English tokens/words and Vietnamese tokens in each sentence pair at the chunk level (see Figure 6).
- ***Word Segmentation***: In general, only Vietnamese text is considered for correcting word segmentation.
- ***POS Tagger***: The annotation tool supports annotating and correcting POS tags for both English and Vietnamese text as shown in Figure 6. However, in our project, we use the POS result of chunking modules as the final results for our corpus.
- ***Chunker***: This stage is based on combining English chunking, Vietnamese chunking, and word alignment results in the comparison between English and Vietnamese structures (as shown in Figure 6).
- ***Named Entity Recognition***: This stage is based on combining English NER and mapping English entities into Vietnamese text to get Vietnamese entities.

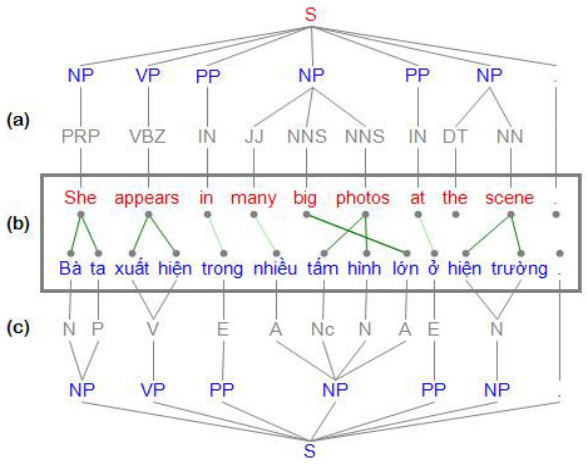

Figure 6: Combine English chunking (a), Vietnamese chunking(c), and word alignment (b)

With the visualization provided by the BiCAT 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

[6] http://sourceforge.net/apps/mediawiki/opennlp/
[7] http://cogcomp.cs.illinois.edu/page/software_view/4
[8] http://nlp.stanford.edu/ner/index.shtml
[9] http://alias-i.com/lingpipe/index.html

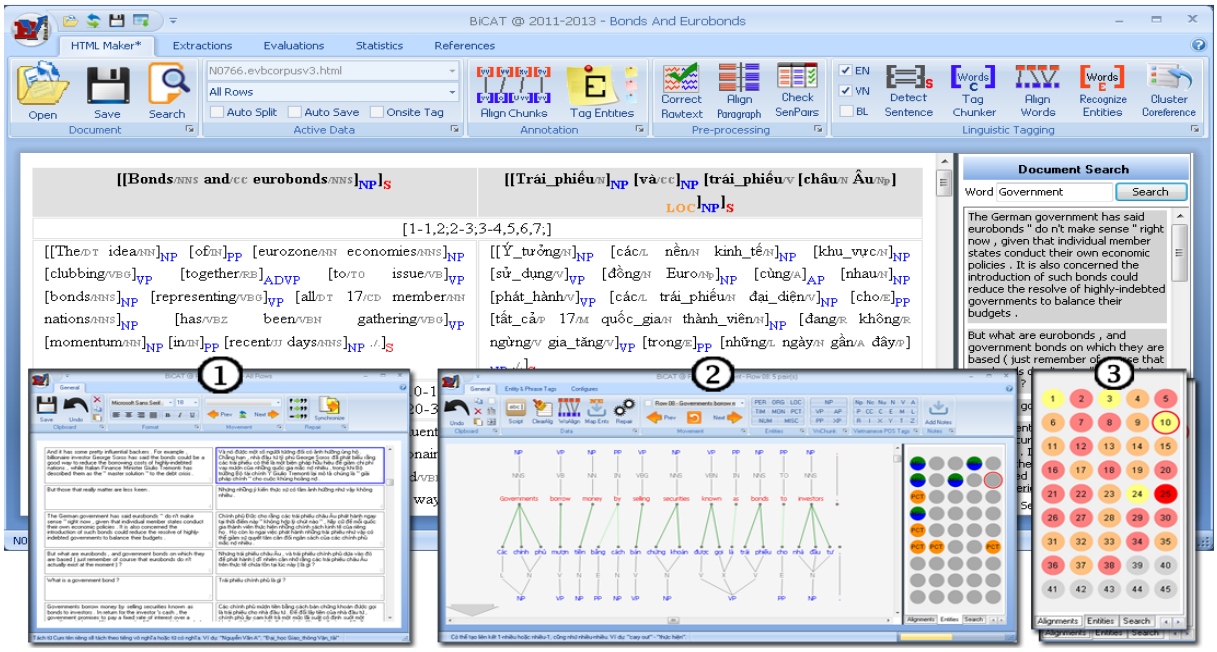

Figure 7: Screenshot of BiCAT with (1) bitext alignment, (2) word alignment, linguistic tagging, and (3) assistant panels

English and Vietnamese, and English-Vietnamese word alignment can be detected and corrected through drag, drop, and edit label operations (actions). Based on drag and drop on labels and tags, annotators can change the results of the tagging modules visually, quickly, and effectively.

As shown in Figure 7, the annotation includes forms for (1) bitext alignment, (2) word alignment, POS/Chunk tagging. This tool also has several (3) assistant panels based on context of tagging words and tags. Assistant panels of the annotation tool are:

- Looking up the bilingual dictionary for meanings and part-of-speech of words to correct translation text and word alignments.
- Searching similar phrase for suggesting and correcting translation text and word alignments.
- State of the word alignment of sentences in whole document for detecting sentence pairs with less alignments.
- Statistics of named entities as a named entity map for detecting unbalanced number of named entities between English and Vietnamese text in the document.

5 Results and Analysis

5.1 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 1,000 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 4. The data is tagged and aligned automatically at the word level between English and Vietnamese.

Table 4: 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. These two types of alignments are annotated to evaluate the alignment models with the Alignment Error Rates (AER) (Och and Ney, 2003). In 1,000 aligned news articles, there are 447,906 sure

alignments, accounting for 80% of 560,215 possible alignments (as shown in Table 4). 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.

5.2 Bilingual Corpus with Linguistic Tags

The first step of linguistic tagging for bilingual corpus is Vietnamese word segmentation. In general, the EVBNews corpus is chosen to practise for building the multi-layer bilingual corpus. This corpus is aligned at the word level as mentioned in Section 5.1.

For Vietnamese, the word segmentation module and the part-of-speech tagger module are packaged into the chunking module. We used vnTokenizer[10] tool (a Vietnamese word segmentation based on a hybrid approach between maximal matching strategy and the linear interpolation smoothing technique) (Le H. Phuong et al., 2008), and vnTagger[11] tool (an automatic part-of-speech tagger for tagging Vietnamese texts) (Le H. Phuong et al., 2010). On the other hand, part-of-speech tagger and chunker of English text can be extracted from the Stanford Parser module as mentioned in Section 3.1. All tagged texts, then, are corrected manually by annotators with the BiCAT tool.

Table 5: Top 5 chunks of EVBNews corpus

Chunk Tags	En. Chunks	Vn. Chunks
NP	238,134	239,286
VP	101,234	138,413
ADJP	9,604	16,196
ADVP	20,681	563
PP	88,722	77,906
Total	**458,375**	**472,364**

The tagset of English chunking includes 9 chunk tags[12] while the Vietnamese chunk tagset has 5 tags: NP, VP, ADJP, ADVP, and PP. Table 5 shows top 5 English and Vietnamese chunks of 1,000 news articles of the EVBNews corpus. In general, the number of English and Vietnamese chunks are nearly equal, however, there is a slight difference between the adjective and adverb chunk of English and Vietnamese. The number of adverb phrases is twice as much as the number of adjective phrases in English text while Vietnamese text mainly uses adjectives to subordinate nouns and verbs.

5.3 Bilingual Named Entity Corpus

As a next layer of the EVBCorpus, Vietnamese named entity tags are tagged for the 1,000 news articles of the EVBNews. Named entities include six tags, Location (LOC), Person (PER), Organization (ORG), Time including date tags (TIM), Money (MON), and Percentage (PCT). English text is tagged with English NER tags by Stanford NER and then mapped to Vietnamese text. Next, Vietnamese entity tags are corrected manually.

In total, there are 32,454 English named entities and 33,338 Vietnamese named entities in the EVBNews corpus (see Table 6). We just focus on the set of alignments and amount of annotation rather than evaluate the quality of the Word Alignment module.

Table 6: Number of entities at each stage

Entity	En. Entities	Vn. Entities
LOC	10,406	11,343
PER	7,201	7,205
ORG	8,177	8,218
TIM	4,478	4,417
MON	998	985
PCT	1,194	1,170
Total	**32,454**	**33,338**

There is a difference between the number of English entities and the number of Vietnamese entities. This difference occurs because several English words are not considered as entities while a part of their translation in Vietnamese is considered as entities. For example, the word *"Vietnamese"* in the sentence *"Nowadays, Vietnamese food is more popular."* is not an entity in the English sentence, while in its Vietnamese translation *"Thức ăn Việt Nam ngày càng được biết đến nhiều hơn."*, the word *"Việt Nam"* is a LOC entity.

6 Conclusions

In this paper, we have introduced a complete workflow to build a multi-layer English-

[10] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer
[11] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTagger
[12] ftp://ftp.ims.uni-stuttgart.de/pub/corpora/chunker-tagset-english.txt

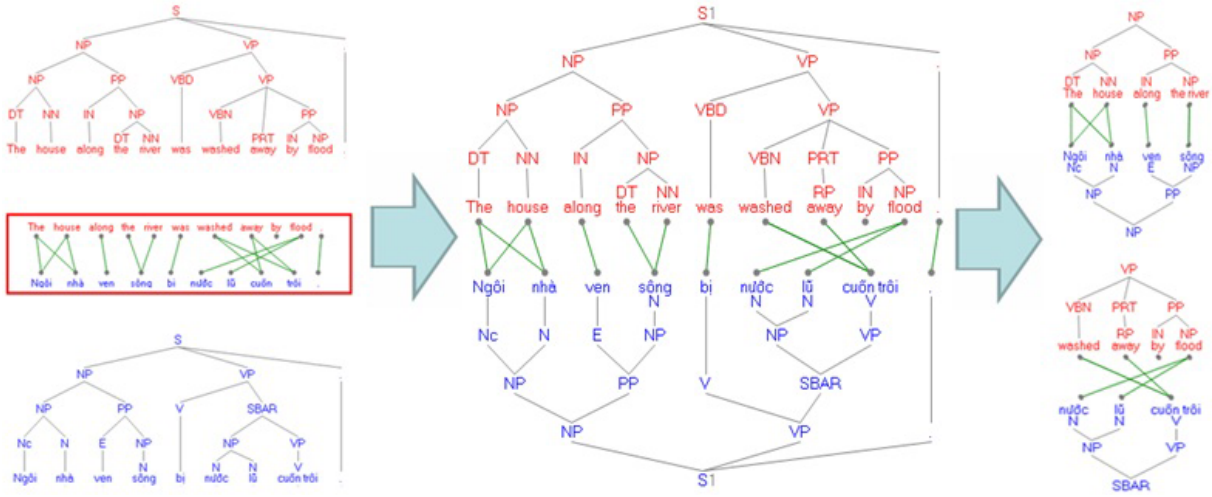

Figure 8: Combine and align full English-Vietnamese parse trees

Vietnamese bilingual corpus, from collecting data, aligning words in bilingual text, tagging chunks and named entities, and developing an annotation tool for bilingual corpora. We showed that the size of the EVBCorpus with over 800,000 English-Vietnamese aligned pairs at the sentence level and 45,531 aligned sentence pairs at the word level is a valuable contribution to study other tasks in comparative linguistics. We pointed out that linguistic information tagging based on our procedure, including tagging and annotation, so far, stops at the chunk level. A part of this corpus and the annotation tool are published at http://code.google.com/p/evbcorpus/.

However, one potential model of full parser alignment is to combine full parse trees and word or chunk alignments as shown in Figure 8. In addition, 45,531 aligned sentence pairs with tagged named entities have been also used to map other linguistic tags (such as co-reference chunks and semantic tags) from English to Vietnamese text.

References

Aljoscha Burchardt, Sebastian Padó, Dennis Spohr, Anette Frank, and Ulrich Heid. 2008. *Formalising multi-layer corpora in OWL/DL–Lexicon modelling, querying and consistency control.* In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008), pp. 389-396.

Amir Zeldes, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. *Annis: A search tool for multi-layer annotated corpora.* In Proceedings of Corpus Linguistics, vol. 9, 2009, pp. 20–23.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. *Multi-level error annotation in learner corpora.* In Proceedings of Corpus Linguistics 2005 Conference, United Kingdom, July, 2005.

Aspasia Beneti, Woiyl Hammoumi, Eric Hielscher, Martin Müller, and David Persons. 2006. *Automatic generation of fine-grained named entity classifications.* Technical report, University of Amsterdam.

Christoph Muller and Michael Strube. 2006. *Multi-level annotation of linguistic data with MMAX2.* Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, 2006, pp. 197-214.

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

Dinh Dien. 2002a. *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.

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

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.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. *Incorporating Non-local Information into Information Extraction Systems by*

Gibbs Sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

Franz Josef Och, Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics 29, 2003, pp. 19–51.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. *Unbounded dependency recovery for parser evaluation.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 813–821.

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

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.* In Proceedings of the 7th Workshop on Asian Language Resources (In Conjunction with ACL-IJCNLP), pp. 9-16.

Le Hong Phuong, Nguyen Thi Minh Huyen, Roussanaly Azim, H. T. Vinh. 2008. *A hybrid approach to word segmentation of Vietnamese texts.* In Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, LATA 2008, Springer LNCS 5196, Tarragona, Spain, 2008, pp. 240-249.

Le Hong Phuong, Azim Roussanaly, Nguyen Thi Minh Huyen, and Mathias Rossignol. 2010. *An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts.* In Proceedings of the Traitement Automatique des Langues Naturelles (TALN2010), Canada, 2010.

Lev Ratinov, Dan Roth. 2009. *Design challenges and misconceptions in named entity recognition.* In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL '09), pp. 147-155.

Jochen L. Leidner, Tiphaine Dalmas, Bonnie Webber, Johan Bos, and Claire Grover. 2003. *Automatic Multi-Layer Corpus Annotation for Evaluating Question Answering Methods: CBC4Kids.* In Proceedings of the 3rd International Workshop on Linguistically Interpreted Corpora, 2003, pp. 39-46.

Hilda Hardy, Kirk Baker, Laurence Devillers, Lori Lamel, Sophie Rosset, Tomek Strzalkowski, Cristian Ursu, and Nick Webb. 2002. *Multi-layer dialogue annotation for automated multilingual customer service.* In Proceedings of the ISLE Workshop, 2002, pp. 90-99.

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. *Generating Typed Dependency Parses from Phrase Structure Parses.* In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), 2006, pp. 449-454.

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), pp. 172-178.

Nguyen Dat, Son Hoang, Son Pham, and Thai Nguyen. 2010. *Named entity recognition for Vietnamese.* Intelligent Information and Database Systems, 2010, pp. 205-214.

Quoc Hung Ngo, Werner Winiwarter. 2012a. *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.

Quoc Hung Ngo, Dinh Dien, Werner Winiwarter. 2012b. *Automatic Searching for English-Vietnamese Documents on the Internet.* In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Languages Processing (3rd SSANLP within the COLING2012), pp. 211-220, Mumbai, India.

Quoc Hung Ngo, Werner Winiwarter. 2012c. *Building an English-Vietnamese Bilingual Corpus for Machine Translation.* In Proceedings of the International Conference on Asian Language Processing 2012 (IALP 2012), IEEE Society, pp. 157-160, Ha Noi, Vietnam.

Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2006. *Multi-dimensional annotation and alignment in an English-German translation corpus.* In Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing, pp. 35-42, ACL 2006.

Stefanie Dipper. 2005. *XML-based stand-off representation and exploitation of multi-level linguistic annotation.* In Proceedings of Berliner XML Tage, 2005, pp. 39-50.

Sue J. Ker and Jason S. Chang. 1997. *A class-based approach to word alignment.* Computational Linguistics 23, No. 2, 1997, pp. 313–343.

Tran Quoc Tri, Xuan Thao Pham, Quoc Hung Ngo, Dien Dinh, and Nigel Collier. 2007. *Named entity recognition in Vietnamese documents.* Progress in Informatics Journal, No. 4, March 2007, pp. 5-13.

Van Bac Dang, Bao Quoc Ho. 2007. *Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining.* In Proceedings of Research, Innovation and Vision for the Future (RIVF'07), IEEE Society, pp. 261-266.