# Building English-Vietnamese Named Entity Corpus with Aligned Bilingual News Articles

**Quoc Hung Ngo**
University of Information Technology
Vietnam National Universiry - HCM City
Ho Chi Minh City, Vietnam
hungnq@uit.edu.vn

**Dinh Dien**
University of Sciences
Vietnam National Universiry - HCM City
Ho Chi Minh City, Vietnam
ddien@fit.hcmus.edu.vn

**Werner Winiwarter**
University of Vienna
Research Group Data Analytics and Computing
Währinger Straße 29, 1090 Wien, Austria
werner.winiwarter@univie.ac.at

## Abstract

Named entity recognition aims to classify words in a document into pre-defined target entity classes. It is now considered to be fundamental for many natural language processing tasks such as information retrieval, machine translation, information extraction and question answering. This paper presents a workflow to build an English-Vietnamese named entity corpus from an aligned bilingual corpus. The workflow is based on a state of the art named entity recognition tool to identify English named entities and map them into Vietnamese text. The paper also presents a detailed discussion about several mapping errors and differences between English and Vietnamese sentences that affect this task.

## 1 Introduction

Named entity recognition (NER) is a basic task in natural language processing and one of the most important subtasks in Information Extraction. It is really essential to identify objects and extract relations between them. Moreover, recognizing proper names from news articles or newswires is also useful in detecting events and monitoring them. The NER task aims to identify and classify certain proper nouns into some pre-defined target entity classes such as person (PER), organization (ORG), location (LOC), temporal expressions (TIME), monetary values (MON), and percentage (PCT).

Several previous works in NER have been done on languages such as English (J. Sun et al., 2002; C.W. Shih et al., 2004), Japanese (R. Sasano and S. Kurohashi, 2008), Chinese (J. Sun et al., 2002; C.W. Shih et al., 2004), and Vietnamese (N.C. Tu et al., 2005; T.X. T. Pham et al., 2007; Q.T. Tran et al., 2007); and NER systems have been developed using supervised learning methods such as Decision Tree, Maximum Entropy model (D. Nadeau and S. Sekine, 2007), and Support Vector Machine (Q.T. Tran et al., 2007), which achieved high performance. Moreover, there are several studies for bilingual named entity recognition (C.J. Lee et al., 2006; D. Feng et al., 2004; F. Huang and S. Vogel, 2002). However, for the English-Vietnamese pair, this task still presents a significant challenge in a number of important respects (R. Sasano and S. Kurohashi, 2008). Firstly, words in Vietnamese are not always separated by spaces, so word segmentation is necessary and segmentation errors will affect the level of NER performance. Secondly, some proper names of foreign persons and locations are loanwords or represented by phonetic symbols, so we can expect wide variations in some Vietnamese terms. Thirdly, there are considerably fewer available existing resources such as lexicons, parsers, word nets, etc. for Vietnamese that have been used in previous studies.

*Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics*, pages 85–93, Dublin, Ireland, August 23-29 2014.

In this study, we suggest a process to build a bilingual named entity corpus from aligned news express articles. In fact, this process is applied to build an English-Vietnamese Named Entity Corpus by using available English named entity recognition to tag entities in the English text, and then map them into Vietnamese text based on word alignments. The mapping results are also corrected manually by using a visualization tool.

The remainder of this paper describes the details of our work. Firstly we address the data source for building the corpus in Section 2. Next, we present a procedure to build an English-Vietnamese Named Entity Corpus by using a bilingual corpus and mapping English entities into Vietnamese entity tags in Section 3. Experimental results and conclusion appear in Sections 4 and 5, respectively.

## 2 Tagset and Data Source

### 2.1 Tagset for Named Entities

There are many tagsets for the NER task, such as the hierarchical named entity tagset with 150 types (S. Sekine et al., 2002), the biological named entity tagset (Y. Tateisi et al., 2000; J-D Kim et al., 2003), or common named entity tagsets with 3 types and 7 tags (N. Chinchor and P. Robinson, 1998). According to the definition of the MUC-7 conference (N. Chinchor and P. Robinson, 1998), we will identify six types of named entities:

- **PERSON (PER)**: Person entities are limited to humans identified by name, nickname or alias.
- **ORGANIZATION (ORG)**: Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.
- **LOCATION (LOC)**: Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.
- **TIME (TIM)**: Date/Time entities are complete or partial expressions of time of day, or date expressions.
- **PERCENTAGE (PCT)**: Percentage entities are percentage expressions, including percentage range expressions.
- **MONEY (MON)**: Money entities include monetary expressions.

We have developed a guide for bilingual named entity tagging and published it as EnVnNEguide at http://code.google.com/p/evbcorpus/downloads/list.

### 2.2 Data Source

The data source for building the English-Vietnamese named entity corpus is a part of the EVBCorpus[1], 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 (as shown in Table 1) (Q.H. Ngo et al., 2013). This corpus is also aligned semi-automatically at the word level.

In particular, each article was translated one to one at the whole article level, so we align sentence to sentence. Then, sentences are aligned semi-automatically at the word level, including automatic alignment by class-based method (D. Dien et al., 2002) and use of the BiCAT tool (Q.H. Ngo and W. Winiwarter, 2012) to correct the alignments manually. The details of the corpus are listed in Table 1.

Parallel documents are also chosen and classified into categories, such as economy, entertainment (art and music), health, science, social, politics, and technology (percentage of each category is shown in Table 2 and Figure 1). The wide range of categories ensures that named entities in the corpus are diversified enough for other following tasks.

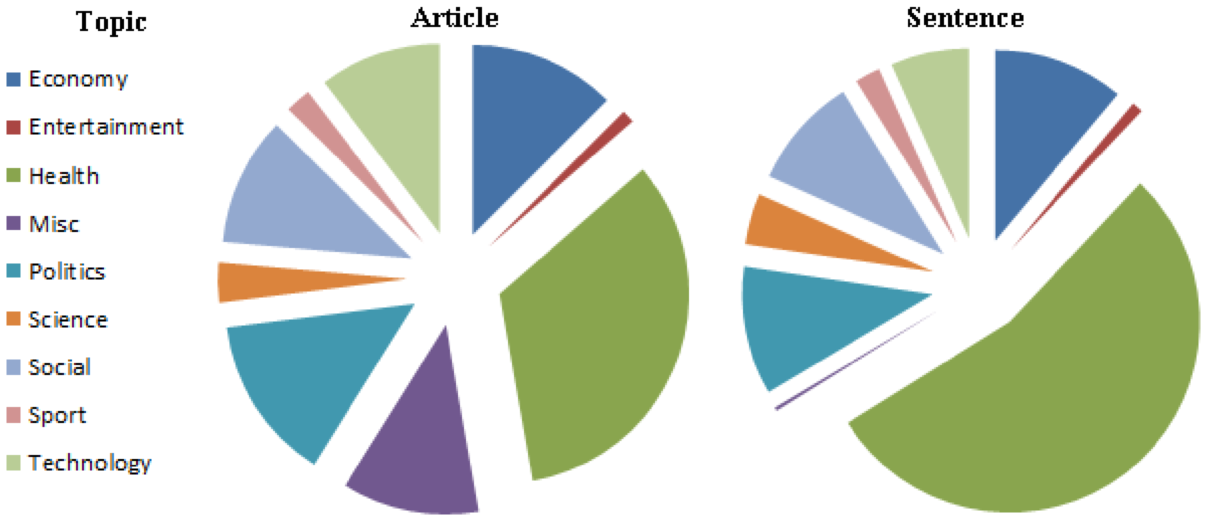


Figure 1: The distribution of articles and sentences in each topic in EVBNews

Table 1: 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 |

## 3 Building Named Entity Corpus

### 3.1 Model of Building EVNECorpus from EVBCorpus

Figure 2 shows the main modules of bilingual named entity corpus building, including three main modules: pre-processing, named entity recognition, bilingual entity mapping, and bilingual entity correction. According to this workflow, the bilingual corpus will be tagged with named entities on the English text, then, named entities are mapped from English to Vietnamese text. Finally, annotators will correct both English and Vietnamese named entities by using the BiCAT tool (Q.H. Ngo and W. Winiwarter, 2012).

At the first stage, a Named Entity Recognition system is used to tag English entities in the English sentence. Several Named Entity Recognition systems for English text are available online. For traditional NER, the most popular publicly available systems are: OpenNLP NameFinder [2] , Illinois NER[3] system by Lev Ratinov (L. Ratinov and D. Roth, 2009), Stanford NER[4] system by Jenny Rose Finkel (J.R. Finkel et al., 3005), and Lingpipe NER[5] system by Baldwin, B. and B. Carpenter (B. Carpenter, 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.

The following example is the result of the Stanford NER for the English sentence “Prime Minister Gordon Brown resigned as Britain ’s top politician on Tuesday evening, making way for Conservative

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

[2] http://sourceforge.net/apps/mediawiki/opennlp/

[3] http://cogcomp.cs.illinois.edu/page/software_view/4

[4] http://nlp.stanford.edu/ner/index.shtml

[5] http://alias-i.com/lingpipe/index.html

Table 2: Number of files and sentences for each topic

| Topic | File | Sentence |
|---|---|---|
| Economy | 125 | 4,326 |
| Entertainment | 11 | 365 |
| Health | 336 | 21,107 |
| Politics | 141 | 4,253 |
| Science | 34 | 1,692 |
| Social | 110 | 3,699 |
| Sport | 22 | 838 |
| Technology | 104 | 2,609 |
| Misc | 117 | 117 |
| **Total** | **1,000** | **45,531** |

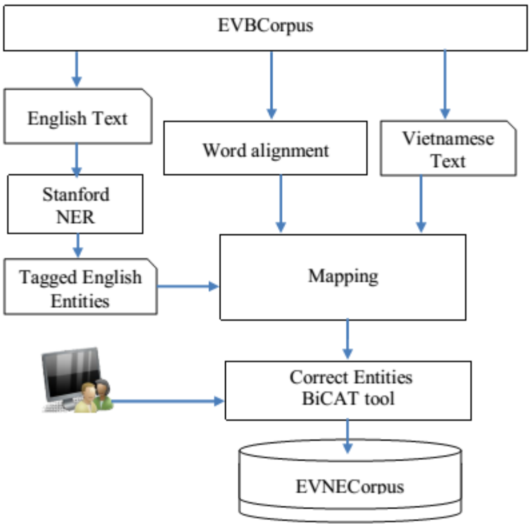


Figure 2: Architecture of building EVNECorpus from EVBCorpus

leader David Cameron." :

Prime Minister [Gordon Brown]$_{PER}$ resigned as [Britain]$_{LOC}$ 's top politician on [Tuesday]$_{TIM}$ evening, making way for [Conservative]$_{ORG}$ leader [David Cameron]$_{PER}$ .

and its mapped named entities in the Vietnamese sentence:

Thủ tướng [Gordon Brown]$_{PER}$ đã từ giã chức vụ cao nhất trên chính trường [Anh]$_{LOC}$ vào tối [thứ ba]$_{TIM}$ , nhường chỗ cho [David Cameron]$_{PER}$ nhà lãnh đạo [Đảng Bảo thủ]$_{ORG}$ .

### 3.2 Mapping English to Vietnamese Named Entities

At the next stage, every alignment will be mapped from English into Vietnamese tokens. Every named entity tag on linked English words is mapped to Vietnamese tokens on the target sentence. Left and right boundaries are also detected to re-build named entity chunks on the Vietnamese sentence (as shown in Figure 3):

- Remove all alignments which are not related to English named entity tokens
- Map English named entity tokens to Vietnamese text by using alignments

- Identify the boundaries of named entities and rebuild named entity script text.

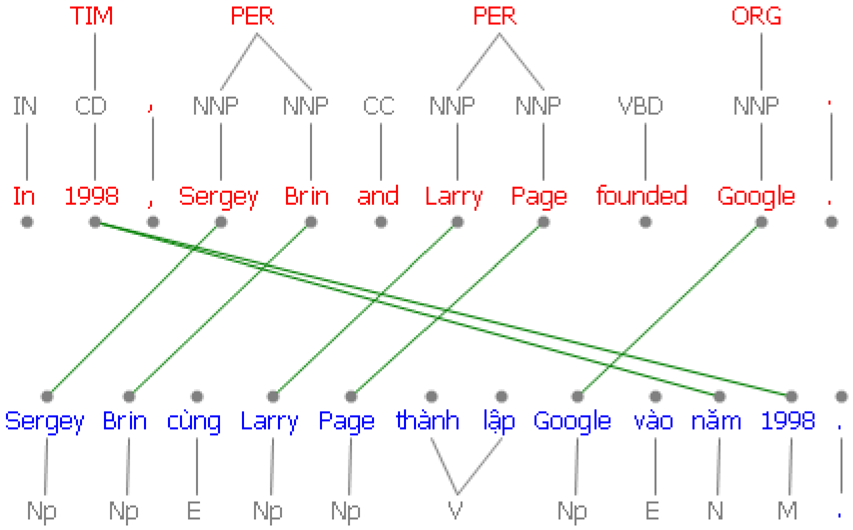


Figure 3: Mapping named entities

However, there is a difference between the number of tagged English named entities and mapped Vietnamese named entities (as shown in Table 3). Several English entities in the English sentences are not translated into the Vietnamese text, therefore, these entities are not mapped.

Table 3: Number of entities at the first stage

| Tag | Name | Tagged English Entity | Mapped Vietnamese Entity | Unmapped Entity |
|---|---|---|---|---|
| LOC | Location | 10,418 | 10,354 | 64 |
| ORG | Organization | 8,197 | 8,120 | 77 |
| PER | Person | 7,217 | 7,153 | 64 |
| TIM | Time | 4,474 | 4,437 | 37 |
| MON | Money | 1,003 | 992 | 11 |
| PCT | Percent | 1,201 | 1,193 | 8 |
| **Total** | | **32,510** | **32,249** | **261** |

There are several common errors of the mapping stage. The most frequent error is caused by the separation of entities from English text in Vietnamese. It means that there are several cases of one tagged English named entity being separated into two distinct Vietnamese named entities in the Vietnamese text. On the other hand, for example, the phrase "President [Bill Clinton]$_{\mathrm{PER}}$ between [1994]$_{\mathrm{TIM}}$ and [1997]$_{\mathrm{TIM}}$" has one PER entity and two TIM entities whereas the Vietnamese translated text "Tổng thống [Bill Clinton]$_{\mathrm{PER}}$ trong giai đoạn [1994-1997]$_{\mathrm{TIM}}$" has one PER entity and only one TIM entity.

Three common reasons which lead to mapping errors (these cases are discussed in the next section by analysing unmatched cases) are:

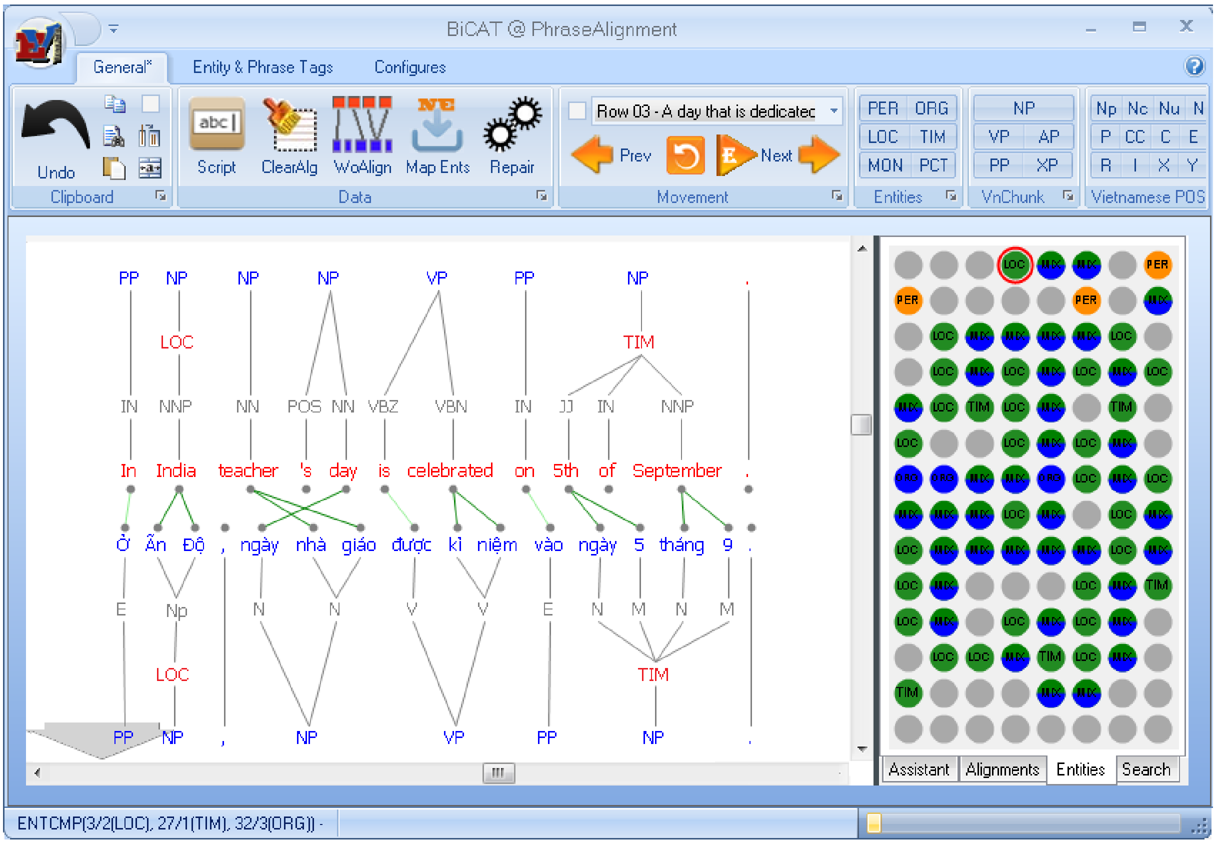
- The differences between English and Vietnamese characteristics.

- The splitting of an English named entity to two Vietnamese entities.

- The entities are replaced by pronouns and possessive pronouns in the target sentences or the other way around.

### 3.3 Correcting Named Entities

As shown in Figure 4, we use the BiCAT tool (Q.H. Ngo and W. Winiwarter, 2012) for correcting named entities in both English and Vietnamese sentences. The BiCAT tool is a visualization tool based on drag, drop, and edit label operations (actions) to correct the sentence pairs. It is designed for annotators to review whole phrase structures of English and Vietnamese sentences. They can compare the English named entity result with the Vietnamese named entity result and correct them in both sentences. The comparison is also used to detect incorrect named entities in both English and Vietnamese text.



Row 03 - A day that is dedicatec

ENTCMP(3/2(LOC), 27/1(TIM), 32/3(ORG)) -

Figure 4: Screenshot of BiCAT with the named entity map

Moreover, several additional information, such as POS tagger, chunker, is also shown for building further linguistic tags. Several features are implemented on the entity matrix at the right panel of the BiCAT tool:

- Show the sentence where named entities occur.
- Highlight the pairs which have imbalance in number of entities between source sentence and target sentence.
- Quick jump to the sentence pair on which the user clicks.

## 4 Experiment and Results

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, the English-Vietnamese Named Entity Corpus (EVNECorpus) has 32,454 English named entities and 33,338 Vietnamese named entities in the EVBNews corpus (see Table 4 for details and its comparison in Figure 5).

Table 4: Number of entities in the EVNECorpus

| Tag | Name | English Entity | Vietnamese Entity | Unmatched Entity |
|---|---|---|---|---|
| LOC | Location | 10,406 | 11,343 | 998 |
| ORG | Organization | 8,177 | 8,218 | 189 |
| PER | Person | 7,201 | 7,205 | 199 |
| TIM | Time | 4,408 | 4,417 | 136 |
| MON | Money | 1,003 | 993 | 32 |
| PCT | Percent | 1,194 | 1,170 | 27 |
| **Total** | | **32,454** | **33,338** | **1,581** |

There are several common unmatched named entities in English-Vietnamese named entity corpus: the English-Vietnamese Named Entity Corpus (EVNECorpus) has 32,454 English named entities and 33,338 Vietnamese named entities in the EVBNews corpus (see Table 5). Moreover, to classify the unmatched named entities, we also tag part-of-speech for English sentences by the POS Tagger[6] of the Stanford Natural Language Processing Group (K. Toutanova and C. D. Manning, 2000).

As shown in Table 5, a large number of English adjectives (tagged JJ tag) are not tagged as named entities while their translations are tagged as locations. Most of them are coming from country names, such as French, English, and Vietnamese, and they refer to people or languages. In the English text ”Cuban missile crisis”, the word Cuban is not tagged as a location because Cuban is an adjective (”Cuban/JJ missile/NN crisis/NN”), while, in its Vietnamese translation, ”Khủng khoảng tên lửa [Cu Ba]$_{\textbf{LOC}}$”, ”Cu Ba” is tagged as a location. Moreover, there are several English named entities that are split into two entities in the Vietnamese sentences. For example, ”[Thailand]$_{\textbf{LOC}}$ ’s [Ministry of Public Health]$_{\textbf{ORG}}$” has two entities while its translation is [Bộ Y tế Thái Lan]$_{\textbf{ORG}}$. Finally, there are several named entities that are replaced by pronouns and possessive pronouns (30 cases for PRP and 11 cases for PRP$) in the translated sentences and the inverse direction (38 cases).

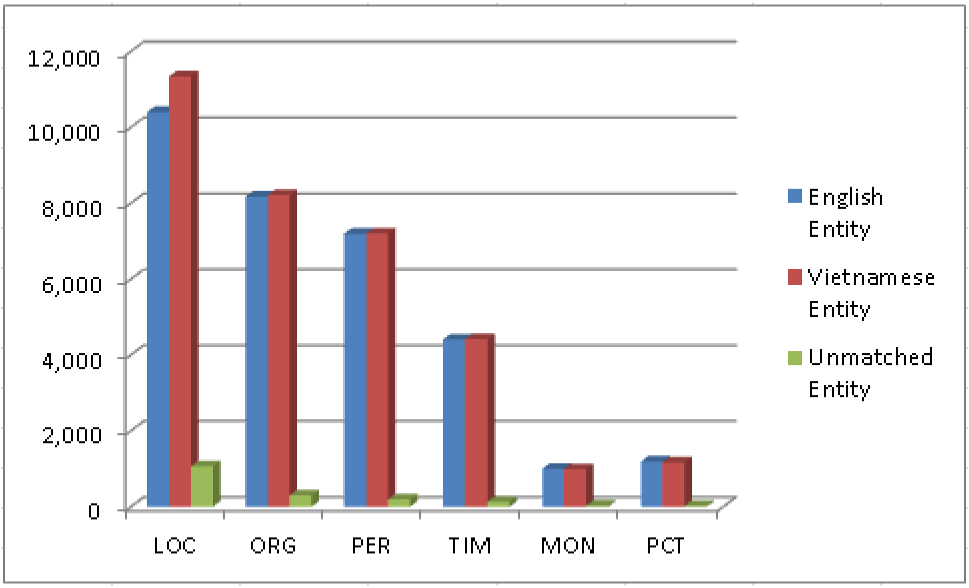


Figure 5: English entities and Vietnamese entities in the EVNECorpus

[6] http://nlp.stanford.edu/software/tagger.shtml

Table 5: Common Unmatched Named Entities

| | Description/Examples | POS | Count |
|---|---|---|---|
| 1 | English named entities without alignments to Vietnamese words | | 190 |
| 2 | Vietnamese named entities without alignments to English source words | | 106 |
| 3 | English named entities with alignments to Vietnamese words rather than Vietnamese entities | NNP | 38 |
| 4 | Vietnamese named entities with alignments to English words rather than English entities | | |
| | - Eurobond, | NN | 10 |
| | - Democrats, Eurobonds, Socialists | NNS | 45 |
| | - Kenyan, French, | NNP | 229 |
| | - Russians, Danish, Philippines | NNPS | 107 |
| | - They, he, she, it | PRP | 36 |
| | - Him, His, her, them | PRP$ | 11 |
| | - French, English, and Fuji-based | JJ | 797 |
| | - other cases | Others | 12 |
| | **Total** | | **1,581** |

## 5 Conclusion

In this paper, we have shown a workflow of building an English-Vietnamese named entity corpus. This workflow is based on an aligned bilingual corpus. In addition, we built a Vietnamese word segmentation corpus for training and evaluating the system. As result, the corpus is built semi-automatically with over 45,000 sentences, and totally 32,454 English named entities and 33,338 Vietnamese named entities. Moreover, we also pointed out several differences in named entity tagging between English and Vietnamese text. These differences can be used to map named entity tags, linguistic information, and in machine translation systems.

However, adding to the six common named entity types additional names (such as product, disease, and event names) is also necessary, and we need further research for identifying them in bilingual corpora because they affect the named entity recognition process as well as the corpus.

## References

Bob Carpenter 2006. *Character Language Models for Chinese Word Segmentation and Named Entity Recognition*, In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 169-172.

Chun-Jen Lee, Jason S. Chang, and Jyh-Shing R. Jang. 2006. *Alignment of Bilingual Named Entities in Parallel Corpora Using Statistical Models and Multiple Knowledge sources*, ACM Transactions on Asian Language Information Processing (TALIP) 5, no. 2 (2006): 121-145.

Cheng-Wei Shih, Tzong-Han Tsai, Shih-Hung Wu, Chiu-Chen Hsieh, and Wen-Lian Hsu. 2004. *The Construction of a Chinese Named Entity Tagged Corpus: CNEC1.0.*, In Proceedings of Conference on Computational Linguistics and Speech Processing (ROCLING). Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Dinh Dien, Hoang Kiem, Thuy Ngan, Xuan Quang, Nguyen V. Toan, Hung Ngo, and Phu Hoi. 2002. *Word Alignment in English – Vietnamese Bilingual Corpus*, In Proceedings of the 2nd East Asian Language Processing and Internet Information Technology (EALPIIT'02), pp. 3-11.

Donghui Feng, Yajuan Lü, and Ming Zhou. 2004. *A New Approach for English-Chinese Named Entity Alignment*, In Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EMNLP), vol. 2004, pp. 372-379. Association for Computational Linguistics.

David Nadeau, and Satoshi Sekine. 2007. *A Survey of Named Entity Recognition and Classification*, Lingvisticae Investigationes 30, no. 1 (2007): 3-26.

Fei Huang, and Stephan Vogel. 2002. *Improved Named Entity Translation and Bilingual Named Entity Extraction*, In Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces, pp. 253-258. IEEE.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. *GENIA Corpus - a Semantically Annotated Corpus for Bio-Textmining*, Bioinformatics 19 (suppl. 1), pp. 80-82. Oxford University Press.

Jenny R. Finkel, Trond Grenager, and Christopher Manning. 2005. *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*, In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. Association for Computational Linguistics.

Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, and Changning Huang. 2002. *Chinese Named Entity Identification Using Class-based Language Model*, In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, pp. 1-7. Association for Computational Linguistics.

Kristina Toutanova, and Christopher D. Manning, 2000. *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger*, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70. Association for Computational Linguistics.

Lev Ratinov, and 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. Association for Computational Linguistics.

Nguyen C. Tu, Tran T. Oanh, Phan X. Hieu, and Ho Q. Thuy. 2005. *Named Entity Recognition in Vietnamese Free-Text and Web Documents Using Conditional Random Fields*, The 8th Conference on Some Selection Problems of Information Technology and Telecommunication. Hai Phong, Vietnam.

Nancy Chinchor, and Patricia Robinson. 1997. *MUC-7 named entity task definition*, In Proceedings of the 7th Conference on Message Understanding.

Quoc Hung Ngo, and 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. Association for Computational Linguistics.

Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. 2013. *EVBCorpus - A Multi-Layer English-Vietnamese Bilingual Corpus for Studying Tasks in Comparative Linguistics*, In Proceedings of the 11th Workshop on Asian Language Resources (11th ALR within the IJCNLP2013), pp. 1-9. Asian Federation of Natural Language Processing Associations.

Quoc Tri Tran, TX. Thao Pham, Quoc Hung Ngo, Dien Dinh, and Nigel Collier. 2007. *Named Entity Recognition in Vietnamese Documents*, Progress in Informatics, No.4, March 2007, pp. 5-13.

Ryohei Sasano, and Sadao Kurohashi. 2008. *Japanese Named Entity Recognition Using Structural Natural Language Processing*, In Proceedings of the International Joint Conference on Natural Language Processing (IJC-NLP), pp. 607-612. Asian Federation of Natural Language Processing Associations.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. *Extended Named Entity Hierarchy*, In Proceedings of the Language Resources and Evaluation Conference, pp. 1818-1824. Association for Computational Linguistics.

TX. Thao Pham, Quoc Tri Tran, Dinh Dien, and Nigel Collier. 2007. *Named Entity Recognition in Vietnamese Using Classifier Voting*, ACM Transactions on Asian Language Information Processing (TALIP) 6, no. 4 (2007): 3.

Yuka Tateisi, Tomoko Ohta, Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. 2000. *Building an Annotated Corpus in the Molecular-Biology Domain*. In Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content, pp. 28-36. Association for Computational Linguistics.