Research Paper

Named Entity Recognition in Vietnamese documents

Tri Tran Q.[1], Thao Pham T. X.[2], Hung Ngo Q.[3], Dien DINH[4] and Nigel COLLIER[5]
[1,2,3]*Faculty of Computer Sciences, University of Information Technology-VNU of HCMC*
[4]*Faculty of Information Technology, University of Natural Sciences - VNU of HCMC*
[5]*National Institute of Informatics*

ABSTRACT

Named Entity Recognition (NER) aims to classify words in a document into pre-defined target entity classes and 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 the results of an experiment in which a Support Vector Machine (SVM) based NER model is applied to the Vietnamese language. Though this state of the art machine learning method has been widely applied to NER in several well-studied languages, this is the first time this method has been applied to Vietnamese. In a comparison against Conditional Random Fields (CRFs) the SVM model was shown to outperform CRF by optimizing its feature window size, obtaining an overall F-score of 87.75. The paper also presents a detailed discussion about the characteristics of the Vietnamese language and provides an analysis of the factors which influence performance in this task.

KEYWORDS

Named Entity Recognition (NER), Support Vector Machine (SVM), text mining

1 Introduction

Named Entity Recognition (NER) aims to identify and classify certain proper nouns into some predefined target entity classes such as *person*, *organization*, *location*, *numeral expressions*, *temporal expressions*, *monetary values*, and *percentage*. Much previous work in NER has been done in languages such as English, [3], [9] Japanese, [2] and Chinese, [4], [12] and NER systems have been developed using supervised learning methods such Decision Tree, [2] Maximum Entropy model, [12] and Support Vector Machine [6] which gained high performance. However, Vietnamese NER appears to present a significant challenge in a number of important respects. 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 is considerably fewer available extant resources such as lexicons, parsers, word nets, etc. for Vietnamese which have been used in previous studies.

In this study, we investigate the effectiveness of NER for Vietnamese free texts using currently available lexical resources and a word based tokenization approach. This is in contrast to a previous study that employed morphosyllables [7]. We compare the use of support vector machines using varying window sizes against a conditional random fields model. Both of these models have been seen to achieve state of the art performance in previous NER tasks.

The remainder of this paper describes the details of our approach. Firstly we describe some features of the Vietnamese language which make it particularly challenging. We then present our SVM based NER model along with experimental results and discussion in sec-

Received October 30, 2006; Revised January 16, 2007; Accepted January 17, 2007.

[1] *tritq@uit.edu.vn*, [2] *xuanthao@nii.ac.jp*, [3] *hungnq@uit.edu.vn*, [4] *ddien@fit.hcmuns.edu.vn*, [5] *collier@nii.ac.jp*

DOI: 10.2201/NiiPi.2007.4.2

An example of ambiguous Vietnamese word segmentation.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Một	**luật**	**gia**	**cầm**	**cự**	**với**	**tình**	**hình**	**hiện**	**nay**
	A, an	law	home, *increase*	hold, *animal*	resist	With	love, *state*	image	appear	now
#1	Một	***luật gia***		***cầm cự***		với	tình hình		hiện nay	
	A	**lawyer**		**contends**		With	situation		present	
#2	Một	luật	***gia cầm***		***cự***	với	tình hình		hiện nay	
	A	law	**poultry**		**resists**	With	situation		present	

tion 3 and section 4, respectively. Conclusion and future work appear in section 5.

2 Vietnamese features

2.1 Word boundary

Vietnamese writing is monosyllabic in nature. Every “syllable” is written as though it were a separate dictation-unit with a space before and after. This unit is called morphosyllable or “tiếng” in Vietnamese. Each morphosyllable tends to have its own meaning and thus a strong identity. However, these morphosyllables are not automatically combined into ‘words’ as the linguistic notion of word commonly applies for European languages. This is a key issue in many endless controversies among Vietnamese linguists who form two schools of thought: (1) “every morphosyllable is a word” and (2) “not every morphosyllable is word”. In this paper, we follow the latter school. We consider morphosyllables in Vietnamese to have the status of morphemes, i.e. one or many (up to 4) morphosyllable(s) combine together to form a single word, which can be identified grammatically or semantically correct by its context. For example, in a sample Vietnamese sentence “Một luật gia cầm cự với tình hình hiện nay.”, there are 10 morphosyllables.

In a dictionary, these ten morphosyllables are 10 words with their own meanings. But in this sentence, some of them are only morphemes. There are many different ways to perform word-segmentation, but only two of them are grammatically correct and one of them (#1) is more reasonable in terms of its semantics as can be seen in the following gloss (here we use the underscore “_” to link morphosyllables of a Vietnamese word together)

1) “A lawyer contends with the present situation”
(“Một luật_gia cầm_cự với tình_hình hiện_nay”)
2) “A law poultry resists the present situation”
(“Một luật gia_cầm cự với tình_hình hiện_nay”)

2.2 Loanwords

More than 50% of Vietnamese vocabulary originated from Chinese which we refer to as Sino-Vietnamese words which are usually used in writing texts, especially in science and politics. We consider these Sino-Vietnamese words to be inherited-words, not loanwords. These inherited-words are similar in status to Chinese-originated words in Japanese (Kanji) or Latin-originated words in European languages. Within the Vietnamese vocabulary, most loanwords are actually from French (words borrowed during the French colonization from 1789 till 1945) and English (words borrowed recently, since 1945 until the present).

Because Vietnamese writing is based on Latin characters and is an exact phonetic transcription (phoneme transcription) of the spoken language, it is convenient to form the foreign words by phonetic transliteration (with or without hyphens between syllables) or keeping it unchanged (especially in recent high-level textual materials), e.g. cinéma (French) → “xi–nê” or “xinê”; virus (English) → “vi-rút” or “virút” or “virus”; Albert Einstein → “An-be Anh-xtanh”; White House → “Bạch_Ốc” (an old Sino-Chinese word was translated from Chinese with the meaning “a house with white color”) or “Nhà_Trắng” (a pure Vietnamese word with the meaning “a house with white color”); Tokyo → “Đông_Kinh” (an old Sino-Chinese word was translated from Chinese with the meaning “East Capital”) or “Tô-ki-ô” or “Tô-ky-ô” or “Tokyo” (current usage).

In addition, there are many compound words used in mixed combination with Vietnamese (pure Vietnamese and/or Sino-Vietnamese) and loanwords. E.g. “tiêm vắc-xin” (to vaccinate); “vi-rút cúm gia_cầm” (bird flu virus); etc.

2.3 Vietnamese word formation

In terms of typology, Vietnamese is an isolating language, words have no inflection and word formation is a combination of isolated morphosyllables and changes in the word-order. All syntactic aspects will be represented by grammatical particles (another word). For example: “viết” (write) → “(đã) viết” (wrote); “sẽ viết” (will write); “đang viết” (be writing); “vắc-xin” (vaccine); “tiêm vắc-xin” (vaccinate); “đã tiêm vắc-xin” (vaccinated); “bệnh”(disease) → “(nhiều) bệnh” (diseases); etc.

2.4 Vietnamese spelling variation

There are variations (with the same meaning) in Vietnamese spelling due to the following reasons:

- The use of hyphenation in loanwords or compound words: inside the loanwords or compound words, each syllable may be hyphenated or non-hyphenated. E.g. “Tô-ki-ô” or “Tôkiô” (Tokyo). In which, the latter form is recently more acceptable.

- The tone-mark in transliterations: in transliterating foreign words, each syllable may be tone-marked or not. E.g. “Đề-các” or “Đê-cac” (Descartes). In which, the latter form is recently more acceptable.

- The similar sound in transliterations of loanwords: since Vietnamese writing is the phoneme transcription of the spoken language, foreign words will be transcribed in sound-like monosyllable or polysyllable. E.g.: “Indonesia” (English) may be “In-đô-nê-xi-a” or “In-đô-nê-si-a” because the sound of /si/ and /xi/ are similar.

- The capitalization of proper names: in proper names (Vietnamese or foreign names) which more than one morphosyllable, all or first or some of morphosyllables are capitalized. E.g. “the National University of HCM city” has seven variants such as “Trường Đại_học Quốc_gia TP.HCM”, “Trường Đại_Học Quốc_Gia TP.HCM”, “Trườ ng đại_học quốc_gia TP.HCM”, ...; Descartes → “Đề-các” or “Đề-Các”, ...

3 Named entity recognition

3.1 Outline of NER model

We developed our model using SVM [11], [13] which performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories.

Suppose we have a set of training data of the form $\{(x_1,y_1), \ldots, (x_N,y_N)\}$ where $x_i \in R^D$ is a feature vector of the i-th sample in the training data and $y_i \in \{+1, -1\}$ is the class to which x_i belongs. The goal is to find a decision function that accurately predicts class y for an input vector x. A non-linear SVM classifier gives a decision function f(x)=sign(g(x)) for an input vector where

$$g(x) = \sum_{i=1}^{m} w_i K(x, z_i) + b$$

Here f(x) = 1 means x is a member of a certain class and f(x) = −1 means x is not a member. z_is are called support vectors and are representatives of training examples. *m* is the number of support vectors. Therefore, the computational complexity of g(x) is proportional to *m*. Support vector and other constants are determined by solving a certain quadratic programming problem. K(x, z) is a kernel that implicitly maps vectors into a higher dimensional space. Typical kernels use dot products: K(x, z) = k(x.z).

Our general NER system includes two main phases:

- ■ Training
- ■ Classification

Both the training and classification processes were done by YamCha[1] [10] toolkit, an SVM-based tool for detecting classes in documents and formulating the NER task as a sequential labeling problem. Here, the pairwise multi-class decision method and the *second polynomial kernel function* were selected.

In the training phase (see Fig. 1), with the information from the gazetteer, we extracted features of words in the gold standard corpus (training data with correct labels). Then we used SVM to train this corpus and got a result model.

In the classification phase (see Fig. 2), before an unannotated article is processed for NER, it needs to

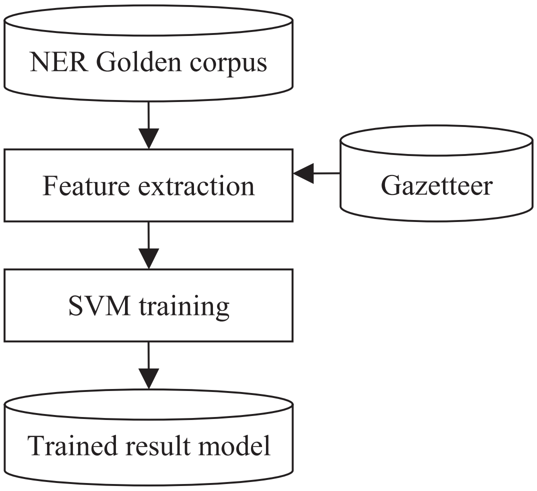

Fig. 1 Architecture of NER training.

[1] http://chasen.org/~taku/software/yamcha/

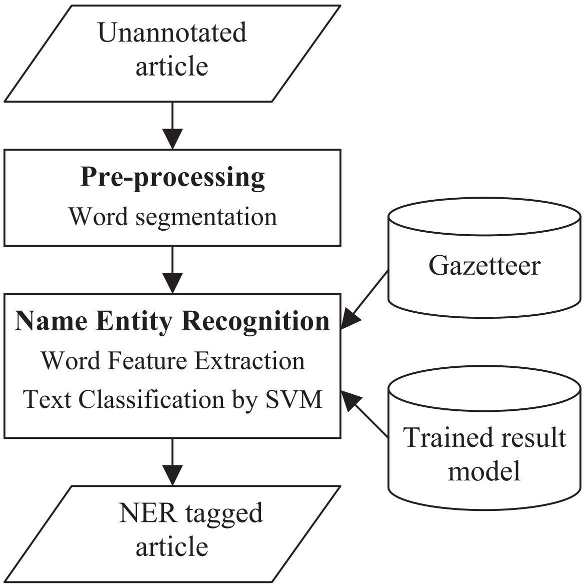

Fig. 2 Architecture of NE classification.

undergo word segmentation. Following from this the words in the article are featurized. Finally, named entities in the article will be identified and classified by the SVM model constructed in the training phase.

We used another toolkit called CRF++[2], a Conditional Random Fields (CRFs) based tool for segmenting and labeling sequence data, to compare against our model. This is briefly described below. Let G = (V, E) be a graph such that $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$, so that **Y** is indexed by the vertices of G. Then (**X**, **Y**) is a **conditional random field** in case, when conditioned on **X**, the random variables $\mathbf{Y}_v$ obey the Markov property with respect to the graph: $p(\mathbf{Y}v \mid \mathbf{X}, \mathbf{Y}_w, w \neq v) = p(\mathbf{Y}v \mid \mathbf{X}, \mathbf{Y}_w, w \sim v)$, where w∼v means that w and v are neighbors in G [5].

3.2 Experiment details

3.2.1 Data set

The named entity tags are provided in IOB notation:

I Current word is inside of a named entity

O Current word is outside of a named entity

B Current word is the beginning of a named entity

The IOB notation is used when named entities are not nested and therefore do not overlap. For instance, we can identify the whole phrase "*Bộ Nông_nghiệp và Phát_triển Nông_thôn*" (Ministry of Agriculture and Rural Development) as an organization name, thus this phrase can be labeled as "B-ORG I-ORG I-ORG I-ORG I-ORG".

Given a word and seven target entity classes including PER (Person), LOC (Location), ORG (Organization), TIM (Date/Time), PCT (Percentage), NUM (Number) and O (Other) class, for any particular named entity class C (except for the O class), the class label could be one of the two forms B-C (Beginning of the named entity C) or I-C (Inside of the named entity C). The labeling problem for NER can therefore be reduced to the problem of assigning one label in 7*2+1=15 labels to each word.

One prohibitive factor when constructing a supervised model based on labeled training examples is the high cost of human annotation. To reduce this we constructed a graphical user interface (GUI) tool to make the manual annotation more convenient. This tagging tool, helped the annotator save time and reduce mistakes while annotating. The process of making the training data can be described as follows: Firstly, we collected thousands of articles from *VnExpress*[3] and *Tuoi Tre*[4] newspapers within the last 6 months of the year 2005. These are two of the most popular online newspapers in Vietnam and cover various topic fields. We picked 500 articles in seven fields: society, health, entertainment, sport, politics, business and sci-tech and used a word segmentation tool [1] whose precision is evaluated at 94.87% to perform word segmentation for these articles before human NE annotation. Following this one hundred articles were manually annotated and we trained an SVM model based on this corpus. This model was used to bootstrap the annotation of a hundred articles among the remaining four hundred articles. All of the automatically annotated articles were then corrected manually. We then added these corrected articles into the pool of the current corpus and retrained the SVM on this expanded corpus. This process was iteratively repeated until the corpus had 500 annotated articles and this is the data used to train in our model.

The entire training data are annotated by one person and contain 156,031 tokens, 109,255 words and 13,603 named entities. Table 1 provides the detailed breakdown of entities in each class.

The training data are in four-column format with one Vietnamese word per line, basic input features in the first three columns and hand-annotated named entity tags in the last one. The tags have to obey certain rules which are based on the definition of each target entity class [8]. For example, a Vietnamese sentence in the corpus "Thủ_tướng Trung_Quốc Ôn_Gia_Bảo đã đến thăm Việt_Nam vào năm 2004." will be annotated as "Thủ_tướng `<NAME cl=''LOCATION''>` Trung_Quốc`</NAME>` `<NAME cl=''PERSON''>` Ôn_Gia_Bảo`</NAME>` đã đến thăm `<NAME`

[2] http://chasen.org/~taku/software/CRF++/

[3] http://vnexpress.net/Vietnam/Home/

[4] http://www.tuoitre.com.vn/Tianyon/Index.aspx

cl=’’LOCATION’’> Việt_Nam </NAME> vào <NAME cl=’’TIME’’> năm 2004 </NAME>.”
and its English translation is
“The Prime Minister of <NAME cl=’’LOCATION’’> China</NAME> <NAME cl=’’PERSON’’>Wen Jiabao </NAME> visited <NAME cl=’’LOCATION’’> Vietnam</NAME> in year <NAME cl=’’TIME’’> 2004</NAME>.”.

Table 2 shows its four-column format in the training data.

3.2.2 Feature set

The basic input features we used were:

1) The current word and two consecutive words before and after the current word (surface word)

2) Orthographic feature of the current word as well as 2 words preceding and succeeding the current word (see Table 3)

3) Gazetteer feature of the current word and words in a window of 2 words before and after the current word

4) The true answer tags (gold standard named entity labels) in the history of the focus word

To improve the performance, we used a gazetteer which contains about 17,500 proper names of Vietnamese people, 7,400 locations and organizations names.

Table 1 No. of entities in each class.

Class	No. of entities
PERSON	1356
LOCATION	4948
ORGANIZATION	2949
CURRENCY	791
NUMBER	2619
PERCENT	252
TIME	1588

4 Experimental results and Discussions

Our first experiment used SVM with a context window of three previous words and three features and is evaluated based on precision, recall and F-measure. The detailed results of every class in the experiment are shown in Table 4.

- Precision (P): number of correctly assigned labels divided by the total number of labeled items.
- Recall (R): number of correctly assigned labels divided by the number of items that should have been assigned a particular label.
- Balance F-measure (F): F=2PR/(P+R)

One important advantage of the TinySVM package is that it lets the user dynamically assign a feature from the history of previous class assignments. We tried a window size of three previous words (SVM1) and this window size gave a better result than that of two previous words (SVM2) and that of four previous words (SVM3). The results of the entire set of 10-fold cross-validation, using four models: SVM1, SVM2, SVM3 and CRF are shown in Table 5. We found that our model achieved the best result with the F-score of 87.75.

Table 2 A sentence in the training corpus.

English word	Surface word	Orthographic feature	Gazetteer feature	Label
The	Thủ_tướng	InitCap	NON	O
Prime Minister				
Of				
China	Trung_Quốc	InitCap	LOC	B-LOC
Wen Jiabao	Ôn_Gia_Bảo	InitCap	NON	B-PER
Visited	đã	LowerCases	NON	O
	đến	LowerCases	NON	O
	thăm	LowerCases	NON	O
Vietnam	Việt_Nam	InitCap	LOC	B-LOC
In	vào	LowerCases	NON	O
Year	năm	LowerCases	NON	B-TIM
2004	2004	Number	NON	I-TIM
.	.	Mark	NON	O

Table 3 Orthographic feature.

Feature	Meaning	Examples
InitCap	The initial letter of the current word is capitalized	Chủ_tịch
AllCaps	All the letters of the current word are capitalized	HCM
LowerCase	All the letters of the current word are uncapitalized	hoạt_động
CapsAndHyphen	All the letters of the current word are capitalized and there is a hyphen in the current word	NN-PTNT
CapsAndPeriod	All the letters of the current word are capitalized and there is a period in the current word	TP.HCM
CapsAndDigits	All the letters of the current word are capitalized and there are numbers in the current word	TCVN3
LettersAndDigits	The current word includes letters and numbers	34/2005/CT/TTg
InitCapAndDigits	The initial letter of the current word is capitalized and there are numbers in the current word	Q3, F4
InitCapAndPeriod	The initial letter of the current word is capitalized and there is a period in the current word	Tp.
InitCapAndHyphen	The initial letter of the current word is capitalized and there is a hyphen in the current word	X-quang
OpenParen	Open bracket	(
CloseParen	Close bracket	)
Brace	Quotation marks, square brackets, etc.	", [,], {, }
Number	The current word is a valid number	5.7, 143
Mark	Colon, question mark, exclamation mark, etc.	:, ?, !, .
Percent	The current word contains % sign	%
Hyphen	The current word contains hyphen	-
Slash	The current word contains slash	/
Date	Day, dates, year, etc.	3-11, 20/07/2006
Time	Time expression	20:30, 14h21’

Table 4 Detailed performance of NER.

Class	P(%)	R(%)	F
PERSON	92.91	87.09	89.90
ORGANIZATION	85.16	77.11	80.93
LOCATION	89.13	88.75	88.93
TIME	87.32	85.01	86.14
NUMBER	89.56	92.74	91.12
CURRENCY	94.52	87.23	90.72
PERCENT	98.80	98.80	98.80
Overall	**89.05**	**86.49**	**87.75**

Table 5 Comparison among models.

	P(%)	R(%)	F
SVM1	89.05	86.49	87.75
SVM2	88.00	85.42	86.69
SVM3	88.82	86.13	87.45
CRF	88.65	84.41	86.48

To evaluate contribution of each feature, we combined each of the features in turn to make sets of features and to evaluate our model according to these feature sets. The conjoined features we used were:

F1: (1) + (2) + (4)
F2: (1) + (3) + (4)

Table 6 Detailed performance.

	F1	F2	F3
SVM1	86.15	85.28	87.75
SVM2	86.40	84.92	86.69
SVM3	86.04	85.13	87.45

F3: (1) + (2) + (3) + (4)

Owing to the F-scores of conjoined features (see Table 6), we found that the combination between surface word and gazetteer feature gave a lower performance than the combination between surface word and orthographic feature. Thus gazetteer feature is useful only when it is combined with the orthographic feature.

The overall result was comparatively high; nevertheless, some classes such as ORGANIZATION, PERSON, LOCATION and TIME were not well recognized. The experiment on NER for only three classes (ORGANIZATION, PERSON, LOCATION) achieved an overall F-measure of 86.70. One possible reason which can explain this is errors of word segmentation. For example, suppose that we have a phrase "Thủ_tướng Trung Quốc Ôn Gia Bảo" (Chinese Prime Minister Wen Jiabao)

("Thủ_tướng" in English is "Prime Minister"
"Trung Quốc" in English is "Chinese"
"Ôn Gia Bảo" in English is "Wen Jiabao")

The correct word segmentation is "Thủ_tướng Trung_Qưốc Ôn_Gia_Bảo" and we have two named entities which are "Trung_Quốc" (LOC) and "Ôn_Gia_Bảo" (PER). However, if the result after word segmentation is "Thủ_tướng **Trung Quốc_Ôn Gia_Bảo"**, we cannot correctly recognize any named entity because in this case, "Quốc_Ôn" is considered an inseparable word.

The recognition is made more little difficult due to variations in Vietnamese spelling, i.e. there are many instances of the same named entity. For example, "Indonesia", "In-đô-nê-xi-a" and "In-đô-nê-si-a" all indicate a country. In addition, there are some cases of nested entities, which influence on the named entity recognition. For instance, "cầu chữ Y" is a Y-shaped bridge while "chợ **Cầu chữ Y**" is a market named after the "Y-shaped bridge". At last, misidentification between LOCATION and ORGANIZATION also affects the result of NER.

We calculated statistics on the number of misidentified cases and found that the errors focus on misclassification between O and ORGANIZATION (presented by B-ORG and I-ORG) classes. Table 7 presents the total number of times that a class is identified as another and the error rate over total errors of the most misidentified cases among classes. For instance, the number of times that B-LOC was misidentified as O (B-LOC vs. O) is 198, and the error rate is 5.47%.

Table 7 Confusion matrix of instances.

Requested tag vs. Tagged tag	No. of errors	Error rate (%)
I-ORG vs. O	342	9.46
B-ORG vs. O	342	9.46
B-LOC vs. O	198	5.47
O vs. ORG-I	160	4.42
I-TIM vs. O	155	4.28
O vs. B-NUM	151	4.17
B-TIM vs. O	141	3.90
B-NUM vs. O	133	3.67
O vs. B-ORG	128	3.54
B-LOC vs. B-ORG	126	3.48

Table 8 Average length of entities in each class.

Class	Term length mean
PERSON	2.38
ORGANIZATION	3.97
LOCATION	2.03
TIME / DATE	4.1
NUMBER	1.35
CURRENCY	2.45
PERCENTAGE	1.85

It is relatively hard to recognize a whole phrase as an ORGANIZATION entity since the term length mean of this class is quite large (see Table 8). The TIME class also gets low performance because of a larger average length (4.1).

5 Conclusion and Future work

This paper introduced an SVM-based system for recognizing named entities in Vietnamese documents. One of the most important components in our NER system is that bearing the responsibility for selecting features. The better and finer the selected features are, the better the NER system is. The experimental results on a moderate number of Vietnamese documents show that this method is not only significantly accurate but also effective.

In previous published work, we are only aware of one similar study [7] which used a CRF based model based on morphosyllable segmentation. This study however made no comparison to other models so we are unable to directly compare our result. In future work however we would like to compare the effects of different segmentation strategies.

Our work is being applied within the BioCaster project[5] which is developing a core multi-lingual text mining system for Internet news to generate real-time summaries of emerging/re-emerging disease outbreaks worldwide. In this project, our task is to help recognize Vietnamese named entities, then find relations to diseases, symptoms, and so on, and keep track of those named entities.

Though experimental results are satisfactory, there are some factors affecting these results such as corpus size, selected features, etc. and the changes of these factors have a certain influence on the accuracy of the system. Hence our future work will focus on the following to improve the application of our method:

- Corpus: Corpus plays an important part in machine learning approach. In the future time, the corpus needs expanding in every aspect. We will extract articles not only from VnExpress and Tuoi Tre, but also from other newspapers to have more various fields and diversified contents.
- Knowledge resources: We will never have a full list of named entities, but the larger the list is, the more accurate the NER system is. So we will supplement the gazetteer as much as possible.
- Feature selection: The currently selected features are based on experience of some NER systems in some languages as well as our experiments. However, it is unsure if they are the best features. So in the future, we will spend more time studying and choosing the best features for our NER model.
- Combination with other methods: SVM is an effective classification method but it has some restrictions. We intend to combine SVM with one or more other methods to increase the accuracy. These combinative methods can be rule-based method or machine learning methods.

Acknowledgments

We thank Mr. Taku Kudo for his Support Vector Machines tool, YamCha. We would like to thank the Global Liason Office of National Institute of Informatics in Tokyo for granting us the travel fund to research this problem. We also sincerely thank colleagues in the Vietnamese Computational Linguistics Group (VCL) for their invaluable and insightful comments.

[5] The **BioCaster** project: http://biocaster.nii.ac.jp/

References

[1] Dinh Dien and Vu Thuy, “A maximum entropy approach for Vietnamese word segmentation,” *Proc. of the 4th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future 2006*, HCM City, Vietnam, pp.247–252, 2006.

[2] Hideki Isozaki, “Japanese named entity recognition based on a simple rule generator and decision tree learning,” *Proc. of the Association for Computational Linguistics*, pp.306–313, 2001.

[3] James Mayfield, Paul Mc Namee and Christine Piatko, “Named Entity Recognition using Hundreds of Thousands of Features,” *Proc. of CoNLL-2003*, Edmonton, Canada, pp.184–187, 2003.

[4] Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, and Changning Huang, “Chinese Named Entity Identification Using Class-based Language Model,” *COLING 2002*, Taipei, Taiwan, 2002.

[5] John Lafferty, Andrew McCallum, and Fernando Pereira, “Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data,” *Proc. Of the ICML’01*.

[6] Koichi Takeuchi and Nigel Collier, “Use of Support Vector Machines in Extended Named Entity Recognition,” *Proc. of the 6th Conference on Natural Language Learning*, pp.119–125, 2002.

[7] N.C. Tu, T.T. Oanh, P.X. Hieu and H.Q. Thuy, “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, 2005.

[8] N. Chinchor, “MUC-7 named entity task definition,” *Proc. of the 7th Message Understanding Conference*.

[9] Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang, “Named Entity Recognition through Classifier Combination,” *Proc. of Conference on Natural Language Learning-2003*, Edmonton, Canada, pp.168–171, 2003.

[10] Taku Kudo and Yuji Matsumoto, “Chunking with Support Vector Machines,” *NAACL 2001*, 2001.

[11] T. Joachims, “Making large-scale SVM learning practical.” In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[12] Tzong-Han Tsai, S.H. Wu, C.W. Lee, Cheng-Wei Shih and Wen-Lian Hsu, “A Chinese Named Entity Recognizer Using the Maximum Entropy-based Hybrid Model,” *International Journal of Computational Linguistics and Chinese Language Processing*, vol.9, no.1, 2004.

[13] Vladimir N. Vapnik, “The Nature of Statistical Learning Theory.” Springer, 1995.

Tri Tran Q.

Tri Tran Q. is a teaching assistance, University of Information Technology, VNU-HCMC, Vietnam. His research interests include Machine Learning, Natural Language Processing, and Information Extraction. He is now internship at the National Institute of Informatics (NII) in Tokyo in the BioCaster project.

Thao Pham T. X.

Thao Pham T. X. is a teaching assistance, University of Information Technology, VNU-HCMC, Vietnam. Her research interests include Natural Language Processing, and Named Entity Recognition. She is now internship at the National Institute of Informatics (NII) in Tokyo in the BioCaster project.

Hung Ngo Q.

Hung Ngo Q. is a lecturer in Faculty of Computer Science of University of Information Technology, VNU-HCMC, Vietnam. His research interests include Machine Translation, Text Mining, Text Summarization, and Named Entity Recognition. He is now internship at the National Institute of Informatics (NII) in Tokyo in the BioCaster project.

Dien DINH

Dien Dinh is a senior lecturer in Department of Knowledge Engineering 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. He is now visiting researcher at the National Institute of Informatics (NII) in Tokyo in the BioCaster project.

Nigel COLLIER

Nigel Collier is associate professor in the Principals of Informatics Research Divisions at the National Institute for Informatics (NII) in Japan. Before coming to NII he received a B.Sc. in Computer Science from Leeds University (UK) in 1992, an M.Sc. in Machine Translation from UMIST (UK) in 1994 and a Ph.D. in Language Engineering from UMIST (now merged with Manchester University) in 1996. From 1996 to 1998 he was a Toshiba Fellow working in Toshiba's human interface laboratories on machine translation, and from 1998 to 2000 he worked on information extraction in the molecular-biology domain at the Tsujii laboratory of the University of Tokyo as a JSPS research associate. He is a member of various organizations including ACL, IEEE Computer Society, ACM and IPSJ. His research interests include NLP - primarily using empirical methods, machine learning of NL knowledge from corpora, and artificial intelligence. He currently manages the BioCaster Project with the aim of detecting and tracking rumors about disease outbreaks from news text.