Combining Proper Name-Coreference with Conditional Random Fields for Semi-supervised Named Entity Recognition in Vietnamese Text

Rathany Chan Sam, Huong Thanh Le,
Thuy Thanh Nguyen, and Thien Huu Nguyen

Hanoi University of Science and Technology,
1 DaiCoViet street, Hanoi, Vietnam
rathany_cam@yahoo.com,
{huonglt,thuynt}@it-hut.edu.vn,
nguyenhuuthien88bk@yahoo.com

Abstract. Named entity recognition (NER) is the process of seeking to locate atomic elements in text into predefined categories such as the names of persons, organizations and locations. Most existing NER systems are based on supervised learning. This method often requires a large amount of labelled training data, which is very time-consuming to build. To solve this problem, we introduce a semi-supervised learning method for recognizing named entities in Vietnamese text by combining proper name coreference, named-ambiguity heuristics with a powerful sequential learning model, Conditional Random Fields. Our approach inherits the idea of Liao and Veeramachaneni [6] and expands it by using proper name coreference. Starting by training the model using a small data set that is annotated manually, the learning model extracts high confident named entities and finds low confident ones by using proper name coreference rules. The low confident named entities are put in the training set to learn new context features. The F-scores of the system for extracting "Person", "Location" and "Organization" entities are 83.36%, 69.53% and 65.71% when applying heuristics proposed by Liao and Veeramachaneni. Those values when using our proposed heuristics are 93.13%, 88.15% and 79.35%, respectively. It shows that our method is good in increasing the system accuracy.

Keywords: information extraction, named entity extraction, entity coreference, semi-supervised learning, CRFs.

1 Introduction

Named Entity Recognition is a subtask of information extraction. Its purpose is to identify and classify certain proper nouns into some predefined target entity classes such as person, organization, location and temporal expressions.

Much previous work on NER followed the supervised learning approach [2], [3], [9], [12], [15] which requires a large hand-annotated corpus. Such approaches

J.Z. Huang, L. Cao, and J. Srivastava (Eds.): PAKDD 2011, Part I, LNAI 6634, pp. 512–524, 2011.

can achieve good performances. However, annotating such a corpus requires a lot of human effort. This problem can be solved by using a sequence-based semi-supervised method that trains a classification model on an initial set of labelled data, makes predictions on a separate set of unlabelled data, and then iteratively attempts to create an improved model using predictions of the previously generated model (plus the original labelled data). Based on this method, we propose a semi-supervised learning method for recognizing named entities in Vietnamese text by combining proper name coreference, named-ambiguity heuristics with a powerful sequential learning model, Conditional Random Fields (CRFs). Our approach inherits the idea of Liao and Veeramachaneni [6] and expands it by using proper name coreference. Starting by training the model using a small data set that is tagged manually, the learning model extracts high confident named entities with and finds low confident NEs by using proper name coreference rules. The low confident NEs are put in the training data set to learn new context features.

Example 1.

(a) *Hôm nay có mưa lớn ở* ***Thành phố Hồ Chí Minh*** /It rains heavily in Hochiminh city today.

(b) *Chính phủ đang tìm giải pháp chống tắc đường ở* ***TP HCM*** /The government is finding a method to solve traffic jam in Hochiminh city.

In Example 1, both "*Thành phố Hồ Chí Minh*/Hochiminh city" and "*TP HCM*" are Location entities and refer to one location. However, the system can only find one Location entity with high confident score, which is "*Thành phố Hồ Chí Minh*/Hochiminh city". The phrase "*TP HCM*" is not recognized as a Location entity by the system since the confidence score of this phrase is smaller than the threshold. Based on the coreferent rules, the system discovers that "*Thành phố Hồ Chí Minh*/Hochiminh city" and "*TP HCM*" refer to the same location. From that point of view, "*TP HCM*" is considered as a low confidence part of "*Thành phố Hồ Chí Minh*/Hochiminh city". "*TP HCM*" is then forced to be a Location entity. It is put in the training set to relearn the new feature context.

In addition, based on our empirical study, several named entity (NE) rules are manually added to the system, in order to create new training data from unlabelled text.

The rest of this paper is organized as follows. Section 2 introduces recent studies on semi-supervised NER methods and works that inspire our research. Section 3 briefly introduces include CRF and the training and inference of the CRF. Section 4 discusses the semi-supervised NER problem for Vietnamese text and our solution to this problem. Section 5 analyzes our experimental results. Finally, our conclusions and future work are given in Section 6.

2 Related Works

The term "semi-supervised" (or "weakly supervised") is relatively recent. The main technique for semi-supervised learning is called "bootstrapping" and involves a small degree of supervision for starting the learning process.

Niu et al [13] present a bootstrapping approach for NER. This approach only requires a few common noun/pronoun seeds that correspond to the concept for the target NE type, e.g. *he/she/man/woman* for Person entity. The entire bootstrapping procedure is implemented as training two successive learners: (i) a decision list is used to learn the parsing-based high precision NE rules; (ii) a Hidden Markov Model is then trained to learn string sequence-based NE patterns. The second learner uses the training corpus automatically tagged by the first learner.

Mohit and Hwa [11] used Expectation Maximization (EM) algorithm along with their Naïve Bayes classifier to form a semi supervised learning framework. In this framework, the small labelled dataset is used to do the initial assignments of the parameters for the Naïve Bayes classifier. After this initialization step, in each iteration the Naïve Bayes classifier classifies all of the unlabelled examples and updates its parameters based on the class probability of the unlabelled and labelled NE instances. This iterative procedure continues until the parameters reach a stable point. Subsequently, the updated Naïve Bayes classifies the test instances for evaluation.

Perrow and Barber [14] take advantage of the simple idea that if a term is annotated with a label in one name, it is highly likely that this term should be annotated with the same label everywhere in the data. Then they annotated this label everywhere in the corpus, assuming that the annotations are the same unless explicitly told otherwise (by further annotations). In this way, they used all the data in the corpus, containing largely only partially annotated records. To learn the parameters (the transition and emission probability tables) they use the Expectation Maximization (EM) algorithm which is an iterative algorithm that increases the likelihood of the corpus given the model parameters in each iteration.

The Yarowsky algorithm [17], originally proposed for word sense disambiguation, makes the assumption that it is very unlikely for two occurrences of a word in the same discourse to have different senses. This assumption is exploited by selecting words classified with high confidence according to sense and adding other contexts of the same words in the same discourse to the training data, even if they have low confidence. This allows the algorithm to learn new contexts for the senses leading to higher accuracy.

Wong and Hwee [16] use the idea of multiple mentions of a token sequence being to the same named entity for feature engineering. They use a named entity recognition model based on the maximum entropy framework to tag a large unlabelled corpus. Then the majority tags of the named entities are collected in lists. The model is then retrained by using these lists as extra features. This method requires a sufficient amount of manually tagged data initially to work.

Liao and Veeramachaneni [6] repeated learning to improve training corpus and the feature set by selecting unlabelled data that has been classified with low confidence by the classifier trained on the original training data, but whose labels are known with high precision from independent evidence. They propose two strategies of obtaining such independent evidence for NER. The first strategy

is based on the fact that multiple mentions of capitalized tokens are likely to have the same label and occur in independently chosen context and call that the multi-mention property. The second strategy is based on the fact that entities such as organizations, persons, etc., have context that is highly indicative of the class, yet is independent of the other context (e.g. company suffixes like Inc., Co., etc.; person titles like Mr., CEO, etc.). They use two heuristics to find the low confidence sequence tokens. In the first heuristics, if the sequence of tokens has been classified as (Organization, Person, Location) with high confidence score (larger than a threshold T), their system forces the labels of other occurrences of the same sequence in the same document, to be (Organization, Person, Location) and adds all such duplicate sequences classified with low confidence (smaller than T) to the training data for the next iteration. The second heuristics is removing company suffix or person title from the sentence. Then the system reclassifies the sentence after removing the company suffix or person title and checks whether the labels have low confidence score or not. If it has low confidence score, the sequence will be added to the training data.

Our research bases on [6] and expands it by combining proper name coreference, named-ambiguity heuristics with a powerful sequential learning model, Conditional Random Fields. This approach will be discussed in detailed in Section 4.

3 Conditional Random Field

Conditional random fields are undirected graphical models trained to maximize a conditional probability [7].

A linear-chain CRF with parameters $\lambda = \{\lambda_1, \ldots, \lambda_N\}$ defines a conditional probability for a state (or label) sequence $\mathbf{y} = y_1, \ldots, y_T$ given an input sequence $\mathbf{x} = x_1, \ldots, x_T$ to be

$$P_\lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathbf{Z_x}} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{N} \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}, t)\right) \quad . \tag{1}$$

where T is the length of sequence, N is the number of features, $\mathbf{Z_x}$ is the normalization constant that makes the probability of all state sequences sum to one, $f_k(y_{t-1}, y_t, \mathbf{x}, t)$ is a feature function which is often binary-valued, but can be real-valued, and λ_k is a learned weight associated with feature f_k. Large positive values for λ_k indicate a preference for such an event, while large negative values make the event unlikely.

The weights of a CRF, $\lambda = \{\lambda_1, \ldots, \lambda_N\}$, are a set to maximize the conditional log-likelihood of labelled sequences in some training set, $D = \{(\mathbf{x}^1, \mathbf{l}^1), (\mathbf{x}^2, \mathbf{l}^2), \ldots, (\mathbf{x}^M, \mathbf{l}^M)\}$:

$$L_\lambda = \sum_{j=1}^{M} \log(P_\lambda(\mathbf{l}^j|\mathbf{x}^j)) - \sum_{k=1}^{N} \frac{\lambda_k^2}{2\sigma^2} \quad . \tag{2}$$

where the second sum is a Gaussian prior over parameters (with variance σ) that provides smoothing to help cope with sparsity in the training data.

When the training labels make the state sequence unambiguous (as they often do in practice), the likelihood function in exponential models such as CRF is convex, so there are no local maxima, and thus finding the global optimum is guaranteed. It has recently been shown that quasi-Newton methods, such as L-BFGS, are significantly more efficient than traditional iterative scaling and even conjugate gradient [10].

Inference in CRF is to find the most probable state sequence $\mathbf{y}^*$ corresponding to the given observation sequence $\mathbf{x}$.

$$\mathbf{y}^* = \text{argmax}_{\mathbf{y}^*} p(\mathbf{y}|\mathbf{x}) \ . \tag{3}$$

In order to find $\mathbf{y}^*$, one can apply the dynamic programming technique with a slightly modified version of the original Viterbi algorithm for HMMs.

4 Named Entity Recognition in Vietnamese Text

The training module of our NER takes as input a small set of Vietnamese documents that have been annotated for three types of named entities including Person, Organization and Location, e.g., *Ông* <Person>*Nguyễn Cảnh Lương*</Person> *hiện giữ chức vụ phó hiệu trưởng* <Orgnization>*trường Đại học Bách khoa Hà Nội*</Orgnization > /Mr.Nguyen Canh Luong currently keeps the position of vice-president of Hanoi University of Science and Technology.

The model after one training step is used to repredict unlabelled training data. The confident score is computed to find the high confidence NEs. The low confident NEs are then found based on the confident score and heuristics.

Section 4.1 presents Vietnamese characteristics that involve the organization of named-entities, in order to create heuristics for finding low confidence NEs. From that point of view, the heuristics are used in our system are proposed. Section 4.2 introduces the semi-supervised learning algorithm for recognizing named entities.

4.1 Characteristics of Vietnamese Proper Names

There are some cases of ambiguities between entities in Vietnamese text, as shown below.

1. **Case 1:** One name or one NE is a part of another NE. More specifically, one NE may be the middle part or the last part of another NE. In this case, the NE that covers the larger text should be tagged instead of the smaller one. For example, the phrase "*Công ty Phát triển Phần mềm Tạ Quang Bửu*/The Ta Quang Buu software development company" is the name of an organization. This phrase contains the word "*Tạ Quang Bửu*" which is the name of a Vietnamese famous scientist. In this case, this phrase should be tagged as an organization name instead of tagging "*Tạ Quang Bửu*" as a person name.
2. **Case 2:** The recognition of a name entity depends on its context. For example:

Example 2.

(a) *Hôm nay,* ***công ty FPT*** *tổ chức liên hoan cho các thành viên* /Today, the FPT company organizes a party for its employees.

(b) *Hôm nay, chúng tôi sẽ tổ chức liên hoan ở* ***công ty FPT*** /Today, we organize a party at the FPT company.

The FPT company is an organization in the first sentence, whereas it is a location in the second one.

These ambiguities will be considered to create NER rules, as reported in Section 4.2.

Beside the above rules, the forms of named entities in Vietnamese are also considered to recognize NE. These forms are shown below:

- Person Names: [prefix] + [family name] + [middle name] + name
- Organization Names: [prefix] + [category] + [category of business] + name + [location]
- Location Names: [prefix] + name

In the above forms,the words in the square brackets are optional and the name can sometimes be abbreviated. This abbreviation can be placed in round brackets or not. Based on above forms, Nguyen and Cao [5] created the proper name coreference rules for Vietnamese text as shown in Table 1.

Some rules in Table 1 can be applied suitably with other languages. For example, rule 2 can be used to find name coreference in English, e.g, Steven Paul Jobs and Jobs. Specific rules for a certain language are also welcoming to make our system more efficient when being operated for that language.

4.2 Semi-supervised Learning Algorithm

Based on the idea of Liao and Veeramachaneni [6] , our system starts by training a model from a small labelled data L. This model is used to find new training data from unlabelled text. After extracting NEs by using the model getting from the training process, the low confidence NEs in unlabelled texts are detected by using heuristics for proper name coreference, some special rules, and rules for resolving ambiguity problems in labeling entities. The system is then retrained on the new data which includes low confidence NEs above. This process is repeated until the system cannot be improved. The algorithm is shown in Table 2 below.

In the training process (Step 1), training documents are first parsed by a Part of Speech (POS) tagger [8] to split documents into words and to get the syntactic roles of words in the sentences. Then features used in the CRF algorithm are calculated and a model C is build. The features are common features that are widely applied in other NER systems. These features are the word itself, the orthographic feature, the gazetteer feature and the POS of the current word, two words before and two words after the current word. In other words, all above mentioned features are calculated for the window size of five.

Table 1. Proper Name Coreferent Rules[1]

Rule	Content
1	Two names are similar.
2	One name is a part of another name, e.g., "*Nguyễn Chí Mai*" and "*Mai*".
3	One name is an alias of another name, e.g., "*Sài Gòn*" and "*TP Hồ Chí Minh*".
4	One name is an abbreviation of another name, e.g., "*TP HCM*" and "*Thành Phố Hồ Chí Minh*".
5	The first k words and the last m words of the two names are similar, in which $k + m$ is the total number of words of one name, e.g., "*Công ty Cổ phần Đại An*" and "*Công ty Đại An*".
6	Except the prefix, all words of N2 appear in N1 or are abbreviations of N1, e.g., "*Công ty TNHH Apave Việt Nam*", "*Cty Apave Việt Nam*", and "*Công ty Pave*" are names of a company.
7	One name is the postfix of another name, e.g., "*Nguyễn Chí Mai*" and "*Chí Mai*".
8	The postfix of one name is the abbreviation of words in the postfix of another word; the remaining parts of the two names are similar. For example, with two names "*Bộ Giáo dục và Đào Tạo*" and "*Bộ GD&ĐT*", the string "*GD&ĐT*" is the abbreviation of "*Giáo dục và Đào tạo*".
9	The last k words of two names are similar, the prefix of N2 is an abbreviation of the prefix of N1, in which N2 has $k + 1$ words. For example, "*Công ty HP VN*" and "*Cty HP VN*".
10	All abbreviations of N2 abbreviate for phrases in N1 and all the remaining words in N2 are appeared in N1. For example, all phrases "*Công ty TNHH Hewlett Packard Việt Nam*", "*Cty HP VN*", "*HP VN*", "*HP Việt Nam*", and "*Công ty HP Việt Nam*" are names of one company.
11	Two names appear continuously in a document in the format N1(N2), in which N2 has only one word and is recognized as a NE. For example, "*Phòng Thương mại và Công nghiệp Việt Nam (VCCI)*", "*Liên đoàn Bóng đá Việt Nam (VFF)*", and "*Tổng công ty Cao su VN (Geruco)*".

In the extracting process (Step 2), the unlabelled documents are parsed by a POS tagger and calculated all features like Step 1. Then, these documents are labelled by the classifier that is produced by using the model C received from Step 1.The documents after being labelled in this process are called labelled documents. After the labeling process, the confidence scores are calculated by using constrained forward-backward algorithm [4] (Step 2.1). This algorithm calculates the sum of probabilities of all paths passing through the constrained segment (constrained to be the assigned labels).

Then, all NEs (Person, Location, Organization) that have high confidence scores (the score is larger than a threshold t_1) will be combined with the proper name coreference rules in Table 1 to create searching patterns. The NE with a high confidence score is called a high-confident NE. Each searching pattern is accompanied by the label of the high-confident NE that produces that pattern. For examples, if the old C finds that the phrase "*Nguyễn Chi Mai*/Nguyen Chi

[1] In Table 1, N1 and N2 are two names that are being compared.

Table 2. The semi-supervised NER algorithm

Given:
- L - a small set of labelled training data
- U - unlabelled data

Loop for k iterations:
- **Step 1**: Train a model C based on L
- **Step 2**: Extract new data D based on C
  - 2.1 Classify kth portion of U and compute confidence scores
  - 2.2 Find high-confidence NE segments and use them to find proper name coreference to tag other low-confidence words
  - 2.3 Tag named entities corresponding to rules that the system cannot detect
  - 2.4 Find qualified O words
  - 2.5 Shuffle part of the NEs in the extracted data
  - 2.6 Add extracted data to D
- **Step 3**: Add D to L

Mai" is a Person entity with high confidence score, after applying proper name conference rules, the patterns "*Mai*", "*Nguyễn*/Nguyen", "*Chi Mai*" and "*Nguyễn Chi Mai*/Nguyen Chi Mai" are detected. These patterns are also tagged as Person. With these produced patterns, we perform a process of pattern matching on the labelled documents to extract matched sequences of tokens, called entity candidates. Each entity candidate is assigned a label which is similar to the label associated with the pattern it matches (Step 2.2). The candidates which have low confidence scores (the score is smaller than t_2) or have no label in terms of the old model will be added to the training data with their assigned labels (Step 2.5 and 2.6). These entity candidates are considered as "good entity candidates". This heuristics is called heuristic Group 1.

The reason for using two thresholds t_1[2] and t_2[3] is to ensure only the new knowledge that the old model does not have is added to the training data. The NEs whose confidence score is in the range of t_1 and t_2 are not used for finding new NEs and are not added to the training data to avoid potential ambiguities.

As mentioned in Section 4.1, there are several ambiguities in Vietnamese text. Due to these ambiguities, some good entity candidates found above may not be real entities, but are parts of other entities. They may have been assigned with a label different than their correct label. To solve this problem, the good entity candidates are processed further as indicated below:

1. **Post process 1:** In the labelled documents, for each of the good entity candidates, the smallest Noun Phrase (NP) containing this candidate is checked to see whether the first words of this NP is a prefix of Person, Organiza-

[2] $t\_1 = 0,95$

[3] $t\_2 = 0,85$

These thresholds are chosen based on our experimental results.

tion, or Location or not. If yes, the entity candidate will be replaced by this NP and the label of this NP is determined by the prefix mentioned above. This method allows us to find NEs with complex structures in Vietnamese documents.

Example 3.
(a) *Hôm nay, chị* ***Nguyễn Chí Mai*** *đi Sài Gòn* /Ms. Nguyen Chi Mai goes to Saigon today.
(b) *Hôm nay,* ***công ty Nguyễn Chí Mai*** *mở cửa* /The Nguyen Chi Mai company opens today.

The phrase "*Nguyễn Chí Mai*/Nguyen Chi Mai" in Example 3a is a high-confident NE that has the Person label in the old model. By applying the process of pattern matching, the system finds that "*Nguyễn Chí Mai*/Nguyen Chi Mai" in Example 3b is a good entity candidate with the Person label, although it is actually just a part of a bigger entity - "*Công ty Nguyễn Chí Mai*/The Nguyen Chi Mai Company". In this case, the system finds the smallest NP that contains the phrase "*Nguyễn Chí Mai*/Nguyen Chi Mai", which is "*Công ty Nguyễn Chí Mai*/The Nguyen Chi Mai Company". Since the first word of "*Công ty Nguyễn Chí Mai*/The Nguyen Chi Mai Company" is a prefix of Organization, this NP is used to replace our good entity candidate "*Nguyễn Chí Mai*/ Nguyen Chi Mai" in the training data with the Organization label.

2. **Post process 2:** if a good entity candidate or the NP that replaces an NE candidate in the Post process 1 is preceded by a location adverb (*ở*/in, *tại*/at, *gần*/near, etc.), this candidate or NP will be reannotated with Location.

Example 4.
(a) *Hôm nay,* ***công ty FPT*** *tổ chức liên hoan cho các thành viên* /Today, the FPT company organizes a party for its employees.
(b) *Hôm nay, chúng tôi sẽ tổ chức liên hoan ở* ***công ty FPT*** /Today, we organize a party at the FPT company.

The entity *công ty FPT*/FPT company in Example 4a is the Organization entity. It is the Location entity in Example 4b.

Besides the above heuristics, the system also uses other NER rules that are proposed by us, based on our empirical study. These rules are called Group 2 and are shown in Table 3 below (Step 2.3). Again, some modifications and/or additions that are specific for different languages can help the system adapt successfully to other languages.

In addition, to balance the training data, the word which is predicted as O by C with high confidence score and satisfies one of three conditions (contains no capitalized token, is not a number, is not in a stopword list) will be added to the training data (Step 2.4).

Since the window size is five (two words on the left and two words on the right of the current token), features of two consecutive tokens before and after the low confidence NEs are also added to the training data.

Table 3. Rules in Group 2

Rule	Definition
1	If the NP has a prefix in one of three prefix dictionaries Person, Location and Organization, it will be labelled based on the prefix.
The following rules deal with NPs that have two properties: +Having no prefix in three prefix dictionaries of Person, Location, and Organization. +Containing only one word and all letters of this word are capitalized.	
2	If a NP is followed by a chain of words conforming the following form: [**added word**][**definition word**][**numeral word**] [**word belonging to one of Person, Location, Organization dictionaries**] in which: +**Added word**: *đã*/already, *đang*/in the process of, *vẫn*/still, *rồi*/already, *sẽ*/will, etc +**Definition word**: *là*/is, *chính là*/be, *làm*/do, *chỉ*/only, etc +**Numeral word**: *các*/many, *những*/many, *mọi*/all, *một*/one, *vài*/some, etc Definition word is obligation, whereas added words and numeral words are optional. The label of this NP will be the type of the dictionary that contains it. *Example 5.* (a) ***Andrew Grove*** *là một giám đốc công ty* /Andrew Grove is a director of the company. (b) ***Hồ Chí Minh*** *là con đường huyền thoại* /Ho Chi Minh is the legendary road. In Example 5, Andrew Grove is a Person entity, whereas *Hồ Chí Minh*/Ho Chi Minh is a Location entity.
3	If a NP is preceded by a word belonging to one of two kinds: verbs which are often followed by a location (*đến*/come, *đi*/go, *tới*/reach, etc) or adverbs which often indicates a place (*tại*/at, *ở*/at, *gần*/near, etc), this NP will be labelled as the Location entity.
4	If a NP is followed by a chain of words conforming the following form: [**definition mark**][**numeral word**] [**word belonging to one of Person, Location, Organization dictionaries**] Definition marks include comma "," , figure dash "-", open parentheses "(" (these marks can usually be an indicator to a definition for its previous word in Vietnamese) Then, the NP is labelled by the type of the dictionary (Person, Location, Organization) *Example 6.* 1. ***Vinamilk****, công ty sữa lớn nhất Việt Nam, được thành lập năm 1976* /Vinamilk, the biggest dairy company in Vietnam, is established in 1976. *Vinamilk* is an Organization entity in Example 6.
5	if a NP is preceded by a chain of words conforming the following form: [**numeral word**][**word belonging to one of Person, Location, Organization Markers**][**word that supplements the meaning of the previous word**][**mark ":" or listing word**] in which: + Listing words includes: *như*/like, *gồm*/include, *gồm có*/include, etc + Supplemental word is often an adjective Then, the NP and all of the words following this NP are labelled according to the Markers (Person, Location, Organization) in the form above (the words must contain only capitalized tokens and punctuations such as comma or semi-colon are ignored) *Example 7.* 1. *Các nước tiên tiến như:* ***Mỹ, Nhật, Pháp****,... đều quan tâm đến vấn đề này* /Advanced countries like USA, Japan, France ... are all concerned about this issue. In Example 7, *Mỹ*/USA, *Nhật*/Japan, *Pháp*/France are Location entities (since the word "*nước*/country" is listed in the Location Markers).

5 Experiments and Discussion

Table 5 show the result of nine iterations of the process when the heuristics in Group 1 and Group 2 are used.

Our experiments use 900 unlabelled documents, and 50 documents labelled manually. Each document contains about 750 tokens. All of these documents are taken from newspaper websites on economics, politics, cultures and education.

50 labelled documents are taken as initial training data; and 900 unlabelled documents are used as testing data. After each round running the training process, 100 documents from those 900 unlabelled documents are taken to find low confidence NEs.

Three experiments were carried out: (i) using the two heuristics in [6]; (ii) using the heuristics in Group 1; and (iii) using the heuristics in Group 1 and Group 2. The results are shown in Table 4 below.

These experiments are evaluated based on Precision, Recall, and F-measure, in which:

- Precision (P): number of correctly assigned labels divided by the total number of labelled items.
- Recall (R): number of correctly assigned labels divided by the number of items that should have been assigned a particular label.
- F-measure: $F = \frac{2 \times P \times R}{P+R}$

Table 4 shows that when the proper name coreference heuristics are used, the results are better than when using the heuristics in [6], especially for the Location

Table 4. Experimental results

Method	Person			Location			Organization		
	P	R	F	P	R	F	P	R	F
Heuristic[6]	79.61	87.48	83.36	65.39	74.23	69.53	67.35	64.15	65.71
Group 1	86.62	90.82	88.67	78.80	85.32	81.93	72.62	80.59	76.40
Group 1 + Group 2	93.53	92.73	93.13	85.32	91.17	88.15	77.10	81.74	79.35

Table 5. Results of 9 interations when heuristics of Group 1 and Group 2 are used

Time	Person			Location			Organization		
	P	R	F	P	R	F	P	R	F
1	69.88	73.51	71.65	48.27	65.95	55.74	46.03	53.75	49.16
2	73.06	77.38	75.16	58.49	69.85	63.67	58.22	65.33	61.57
3	76.13	79.21	77.64	63.34	75.07	68.71	65.66	71.07	68.26
4	80.17	80.89	80.53	69.11	76.97	72.83	67.67	73.60	70.51
5	85.65	84.30	84.97	71.20	78.31	74.59	69.55	75.13	72.23
6	87.45	87.25	87.35	74.83	81.28	77.92	71.83	75.92	73.82
7	91.31	88.57	89.92	75.42	83.78	79.38	75.20	78.85	76.98
8	91.80	91.14	91.47	79.18	86.63	82.74	76.01	80.16	78.03
9	93.53	92.73	93.13	85.32	91.17	88.15	77.10	81.74	79.35

and Organization entities. This is because the Location and Organization entity in Vietnamese text are very complicated. When the heuristics in Group 1 and Group 2 are used, the system also receives higher F-scores.

Table 5 shows that the more new data is added to the training data, the more accurate the system is. If a bigger corpus is used, the system promises to provide a higher accuracy.

6 Conclusions

This paper presents a semi-supervised learning method for recognizing named entities in Vietnamese text. The system starts by training a model with a small labelled data set using CRFs algorithm, then the received model is used to find new training data from unlabelled text. After extracting NEs by using the model getting from the training process, the low confidence NEs in unlabelled text are detected by using heuristics for proper name coreference, some special rules, and rules for resolving ambiguity problems in labeling entities. The system is then retrained on the new data which includes these low confidence NEs. In evaluating the system, our experiments are carried out with the heuristics mentioned above and the heuristics in [6]. The experimental results show that our heuristics outperform the heuristics in [6]. Our future work includes: (i) carrying out experiments with a larger corpus; (ii) investigating other rules that can improve the accuracy of the system; and (iii) experimenting the system with other entity types.

References

1. Blum, A., Mitchell, T.: Combining Labelled and Unlabelled Data with Co-training. In: Proceedings of the Workshop on Computational Learning Theory, pp. 92–100 (1998)
2. Bikel, D., Schwartz, R., Weischedel, R.: An Algorithm that Learns What's in a Name. Machine Learning 34(1-3), 211–231 (1999)
3. Borthwick, A.: Maximum Entropy Approach to Named Entity Recognition. Ph.D. Thesis, New York University (1999)
4. Culotta, A., McCallum, A.: Confidence Estimation for Information Extraction. In: Proceeding of HLT-NAACL, pp. 109–112 (2004)
5. Nguyen, T.H., Cao, H.T.: An Approach to Entity Coreference and Ambiguity Resolution in Vietnamese Texts. Vietnamese Journal of Post and Telecommunication 19, 74–83 (2008)
6. Liao, W., Veeramachaneni, S.: A Simple Semi-supervised Algorithm for Named Entity Recognition. In: Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, pp. 28–36 (2009)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of ICML Conference, pp. 282–290
8. Le, H.P., Roussanaly, A., Nguyen, T.M.H., Rossignol, M.: An Empirical Study of Maximum Entropy Approach for Part of Speech Tagging of Vietnamese Texts. In: Proceedings of TALN 2010 Conference, Canada (2010)

9. McCallum, A., Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. In: Proceedings of CoNLL, Canada, pp. 188–191 (2003)
10. Malouf, R.: A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In: Sixth Workshop on Computational Language Learning, CoNLL (2002)
11. Mohit, B., Hwa, R.: Syntax-based semi-supervised Named Entity Tagging. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Michigan, pp. 57–60 (2005)
12. Nguyen, C.T., Tran, T.O., Phan, X.H., Ha, Q.T.: Named Entity Recognition in Vietnamese Free-Text and Web Documents Using Conditional Random Fields. In: Proceedings of the 8th Conference on Some Selection Problems of Information Technology and Telecommunication, Hai Phong, Vietnam (2005)
13. Niu, C., Li, W., Ding, J., Rohini, K.S.: A Bootstrapping Approach to Named Entity Classification Using Successive Learner. In: Proceedings of the 41st Annual Meeting of the ACL, pp. 335–342 (2003)
14. Perrow, M., Barber, D.: Tagging of Name Record for Genealogical Data Browsing. In: Proceedings of the 6th ACM/IEEE JCDL, Chapel Hill, NC, USA, pp. 316–325 (2006)
15. Tran, Q.T., Pham, T.X.T., Ngo, Q.H., Dinh, D., Collier, N.: Named Entity Recognition in Vietnamese Using Classifier Voting. Proceedings of ACM Transactions on Asian Language Information Processing, TALIP (2007)
16. Wong, Y., Ng, H.T.: One Class per Named Entity: Exploiting Unlabelled Text for Named Entity Recognition. In: Proceedings of IJCAI, pp. 1763–1768 (2007)
17. Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In: Proceedings of Meeting of the Association for Computational Linguistics, pp. 189–196 (1995)