Semi-supervised Learning for Vietnamese Named Entity Recognition using Online Conditional Random Fields

Quang H. Pham
University of Science, Vietnam
quangg2012@gmail.com

Minh-Le Nguyen
Japan Advanced Institute of
Science and Technology
nguyenml@jaist.ac.jp

Binh T. Nguyen
University of Science, Vietnam
ngtbinh@hcmus.edu.vn

Nguyen Viet Cuong
National University of Singapore
nvcuong@nus.edu.sg

Abstract

We present preliminary results for the named entity recognition problem in the Vietnamese language. For this task, we build a system based on conditional random fields and address one of its challenges: how to combine labeled and unlabeled data to create a stronger system. We propose a set of features that is useful for the task and conduct experiments with different settings to show that using bootstrapping with an online learning algorithm called Margin Infused Relaxed Algorithm increases the performance of the models.

1 Introduction

Named Entity Recognition (NER) is an important problem in natural language processing and has been investigated for many years (Tjong Kim Sang and De Meulder, 2003). There have been a lot of works on this task, especially for major languages such as English, Chinese, etc. (McCallum and Li, 2003; Gao et al., 2005; Ritter et al., 2011). For the Vietnamese language, several authors have attempted to tackle the NER problem using both supervised and semi-supervised methods (Tu et al., 2005; Tran et al., 2007; Nguyen et al., 2010; Pham et al., 2012; Le Trung et al., 2014). However, previous works for NER in the Vietnamese language mainly used offline supervised learning methods, where all the training data are gathered before a model is trained.

In this paper, we report preliminary results for a Vietnamese NER system trained by using conditional random fields (CRFs) (Lafferty et al., 2001). Unlike previous works for NER in the Vietnamese language, we use an online learning algorithm, the Margin Infused Relaxed Algorithm (MIRA) (Crammer and Singer, 2003), to train the CRFs. Furthermore, due to the fact that the number of labeled data is small while that of unlabeled data is very large, we treat this problem under the semi-supervised learning framework. In particular, we use the bootstrapping method on top of the CRF models to gradually increase the number of labeled data. Using bootstrapping, a small number of new labeled training data are available after each round and then can be used to update the CRF model.

We demonstrate that using MIRA to learn CRFs instead of the traditional offline method would increase the performance of our system. We also propose a set of features that is useful for this task and gives competitive performance. In contrast to previous works such as (Tran et al., 2007), we do not use features from outside sources, e.g. gazetteer features; so our feature set does not require human effort to create such resources and therefore, is easy to build.

The rest of this paper is organized as follows. In Section 2, we review some previous works for the NER task, especially for the Vietnamese language. A brief introduction to CRF and MIRA is given in Section 3. This will be followed by a description of our feature set in Section 4. In Section 5, we describe our semi-supervised learning approach for the Vietnamese NER problem. We show our experimental setup and results in Section 6. In Section 7, we give some discussions about the problem. Finally, we conclude our paper and discuss some future works in Section 8.

2 Related Works

NER is an important problem that was first introduced at the Sixth Message Understanding Conference (MUC–6) (Grishman and Sundheim, 1996)

Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP, pages 50–55,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

and since then has attracted many researchers to investigate the problem with new methods as well as different languages. Over the years, researchers have tried to solve the problem under supervised learning (McCallum and Li, 2003), semi-supervised learning (Ji and Grishman, 2006), and unsupervised learning (Etzioni et al., 2005) frameworks. One dominant approach for NER is supervised learning with conditional random fields (McCallum and Li, 2003). However, semi-supervised learning approaches are also attractive for this task because it is expensive to get a large amount of labeled data. Notably, Riloff et al. (1999) introduced the mutual bootstrapping method that proved to be highly influential. Besides, using bootstrapping methods, Ji and Grishman (2006) were able to improve the performance of existing NER systems.

For the Vietnamese language, using supervised learning, Tu et al. (2005) built an NER system with CRFs and reported 87.90% F_1 score as their highest performance. Using SVMs, Tran et al. (2007) achieved 87.75% F_1 score for the task. For semi-supervised learning, Pham et al. (2012) achieved 90.14% F_1 score using CRFs with the generalized expectation criteria (Mann and McCallum, 2010), while Le Trung et al. (2014) reported an accuracy of 95% for their system that uses bootstrapping and rule-based models.

3 Margin Infused Relaxed Algorithm for CRFs

3.1 Conditional Random Fields

Linear-chain conditional random field (CRF) is a sequence labeling model first introduced by Lafferty et al. (2001). This model allows us to define a rich set of features to capture complex dependencies between a structured observation $\mathbf{x}$ and its corresponding structured label $\mathbf{y}$. Throughout this paper, we will use the term CRF to refer to linear-chain CRF, a widely used type of CRFs in which $\mathbf{x}$ and $\mathbf{y}$ have linear-chain structures.

Formally, let $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ be the input sequence, $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ be the label sequence, $\mathcal{F} = \{f_k(y_t, y_{t-1}, \mathbf{x})\}_{k=1}^{K}$ be a set of real-valued functions (features) over two consecutive labels y_t, y_{t-1} and the input sequence $\mathbf{x}$, and $\Lambda = \{\lambda_k\}_{k=1}^{K}$ be the set of parameters associated with the features that we want to learn. A linear-chain CRF defines the conditional distribution $p(\mathbf{y}|\mathbf{x})$ as:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{k=1}^{K}\sum_{t=1}^{T} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x})\right)$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left(\sum_{k=1}^{K}\sum_{t=1}^{T} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x})\right)$ is the normalization constant, also called the partition function.

Normally, training a CRF is an iterative process where all the parameters are updated after each iteration to maximize the conditional log-likelihood of the training data. During testing, the label sequence for a new test instance is determined by a Viterbi-like algorithm, which returns the label sequence with the highest probability according to the trained model (Sutton and McCallum, 2006).

3.2 Margin Infused Relaxed Algorithm

MIRA is an online learning algorithm developed by Crammer and Singer (2003). In this algorithm, at each round, the model receives a training example, makes a prediction on the example, and suffers a loss. Then the training algorithm updates the weight vector so that the norm of the change to the weight vector is as small as possible while keeping a margin at least as large as the loss of the incorrect examples.

Details of the single-best MIRA (Crammer, 2004; McDonald et al., 2005) for the sequence labeling task are given in Algorithm 1. In the update step at line 4 of the algorithm, $s(\mathbf{x}, \mathbf{y})$ is a scoring function and $L(\mathbf{y}, \mathbf{y}')$ is a loss function. The difference between MIRA and offline training for CRFs is that MIRA processes one data example at a time while the offline algorithm processes all the data at each iteration. However, the features and the prediction algorithm are identical regardless of the learning algorithms.

4 Features for CRFs

We model NER as a sequence labeling task where each word in a sentence is associated with a tag to indicate which type of named entities it belongs to. There are 5 possible tags that we are interested in: *person*, *organization*, *location*, *miscellaneous* (proper names), and *none*. The *none* tag indicates that the corresponding word is not a part of any named entity. For instance, it may be a verb or an adjective.

We build a set of features that is useful for the Vietnamese NER task. Recall that a feature

Algorithm 1 MIRA for Sequence Labeling

```
INPUT: Training data $\mathcal{D} = \{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^{|\mathcal{D}|}$, number of iterations $N$.
$\mathbf{w}_0 \leftarrow \mathbf{0}$;   $\mathbf{v} \leftarrow \mathbf{0}$;   $i = 0$;
for $n = 1$ to $N$ do
    for $t = 1$ to $|\mathcal{D}|$ do
        $\mathbf{w}_{i+1} \leftarrow \arg\min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}_i\|$ such that $s(\mathbf{x}_t, \mathbf{y}_t) - s(\mathbf{x}_t, \mathbf{y}) \geq L(\mathbf{y}_t, \mathbf{y}), \forall \mathbf{y}$;
        $\mathbf{v} \leftarrow \mathbf{v} + \mathbf{w}_{i+1}$;
        $i \leftarrow i + 1$;
    end for
end for
return $\mathbf{v}/(N \times T)$;
```

in CRFs is a function over the observation $\mathbf{x}$ and two consecutive labels y_t and y_{t-1}. In this paper, we use as features the binary functions that can be fully defined based on the observation sequence.

Particularly, the first type of features we use is the identity of words in a window of size 5 and their combinations. Besides, information about capitalization plays a notably important role for this task. For example, a person's name always has its first letter capitalized, and an abbreviation of a company's name or a place is all capitalized (e.g. Ho Chi Minh city is abbreviated as HCM). Thus, we add orthographic features to feed this information into the CRFs. This type of features describes whether a word is in lower case, whether it has the first letter capitalized, whether all of its letters are capitalized, and whether it contains numeric letters or not. For this type of features, we also take the orthographic information of words in a window of size 5. Finally, we include as features the part-of-speech of the word and the combination of the word's identity and its part-of-speech to better describe the context of the sentence.

We note that not all of the features described above are used since there are possibly redundant features that do not increase the performance. Therefore, we conduct a feature selection step for choosing which features to be utilized for later experiments. We first start with the current word's identity and orthographic features. Then, we add several features, build an appropriate model, and measure its performance on a validation set, which contains 150 sentences extracted from the total training data. If the performance increases, we keep those features; otherwise, they are discarded. The process of adding and discarding features is repeated until there is no more feature left to be added.

Group	Features
Single	W_0, W_{-1}, W_1, W_{-2}, W_2, $W_{-1,0}$, $W_{0,1}$, $W_{-2,-1}$, $W_{1,2}$, $W_{-1,0,1}$, $W_{-2,-1,0}$, $W_{0,1,2}$, , O_0 , P_{-1}, P_0, P_1, $P_{-1,0}$, $P_{0,1}$, $P_{-2,-1}$, $P_{1,2}$
Complex	W_0P_1, $W_0P_0O_0$, W_0O_0, $W_0W_{-1}O_0O_{-1}$, $W_0W_1O_0O_1$

Table 1: Features for training the CRFs. The subscripts indicate the position of the features relatively to the current position.

In Table 1, we give the final set of our features. This set includes 2 groups: single and complex features. The first group contains features about word identity (W), part-of-speech (P), and orthographic information (O). Complex features are formed by combining the single features. From Table 1, possible word identity features such as $W_{-1,1}$ and $W_{-2,2}$ are not listed because they were eliminated during the feature selection step.

5 Bootstrapping with CRFs

One main difficulty of the Vietnamese NER task is the lack of labeled data. Since texts from news, books, etc. naturally do not come with named entity labels, we have to manually label the data set. This is tedious and time consuming when the data size becomes very large. One way to address this problem is to gradually create more labeled data with just a small amount of labeled data via semi-supervised learning.

More specifically, we use the bootstrapping method in this paper. First, we build a model on a labeled corpus and use it to label the data from a data set that has not been labeled. After that, we select some newly labeled instances (sentences in our case), remove them from the unlabeled data set, and add them to the labeled data set. The criteria for choosing instances may vary and depend on the task. Then, the next model is trained on the new labeled set and it will also get an amount of new labeled data from the unlabeled data set. This process is repeated until we satisfy with the amount of labeled data that have been received.

We provide our CRF training procedure with bootstrapping in Algorithm 2. The criterion for choosing the sentences from the unlabeled data

Algorithm 2 Bootstrapping with CRFs

```
INPUT: Labeled data set $L$, unlabeled data set $U$, number of iterations $n$, the amount of sentences per round $k$.
for $i = 0$ to $n$ do
    Train CRF model $M_i$ on data set $L$.
    Use $M_i$ to label $U$.
    Choose $k$ labeled sentences $X = \{\mathbf{x}_j\}_{j=1}^{k}$ with highest confidence from $U$.
    $L \leftarrow L \cup X; \quad U \leftarrow U \setminus X$.
end for
return $M_n$.
```

set is to choose the sentence whose best label sequence got the highest probability assigned by the model.

6 Experimental Results

6.1 Setup

We build a corpus of 1,911 sentences from law news articles and manually tag their named entity labels. To build the unlabeled data set, we collect another 17,500 sentences, which also come from law articles .Both data sets are collected from online newspaper articles. The labeled data set is annotated using the IBO label format (Tjong Kim Sang and De Meulder, 2003) with the 5 labels mentioned in Section 4.

For the bootstrapping experiments, we split our corpus into two parts: the first part contains a fixed set of 411 sentences for testing, and the second part contains 1500 sentences for training. We train 3 initial models using 500, 1000, and 1500 sentences respectively from the second part and apply the bootstrapping algorithm to each trained model, with the maximum number of iterations n being 15. In each iteration, the model selects the top $k = 10$ highest confidence (i.e., highest value of $p(\mathbf{y}|\mathbf{x}, \Lambda)$) sentences to add into its training set. Finally, we compare the results of these models after 5, 10, and 15 rounds of bootstrapping with the initial models. To evaluate the performance of the models, we use the micro-averaged precision (P), recall (R), and F_1 score (F).

In our experiments, we use the CRF++ toolkit[1] which comes with MIRA training option to build our models. Regarding the tasks of Vietnamese word segmentation and part-of-speech tagging, we use a standard tool for Vietnamese language processing provided by Nguyen et al. (2005).

6.2 Results

In Table 2, we depict the highest F_1 score (in %) of the models for every 5 rounds of bootstrapping. For all the initial training sizes, the best CRF trained using MIRA outperforms the best normal CRF in the semi-supervised learning scenario. With 1000 initial training sentences, we achieve the highest increase in F_1 score (which is 2.43%) after 5 rounds of bootstrapping with MIRA compared to not using bootstrapping. Our highest performance is 89.16%, obtained by training with 1500 initial sentences and after 15 rounds of bootstrapping with MIRA.

It is interesting to note that the performance does not always increase after every round. From our error analysis, whenever a model makes a mistake at a round, it affects all the following models and makes them more inaccurate. This leads to a decrease of F_1 score for the later models on the fixed test set.

7 Discussions

When inspecting the best model in Table 2 (CRF model using MIRA with 1500 training sentences and 10 rounds of bootstrapping), we find several cases that may be difficult for the model to predict. In the examples below, every two consecutive words are separated by a white space, the syllables in each word are connected by underscores, and the bold phrases include one word and its wrongly predicted label. All words having the *none* label or having been correctly classified are neither in bold nor followed by any label.

For the Vietnamese language, we find that the model may easily confuse a person name with a location name and vice versa. For instance, the model may mistake a person name for a location name as in the following sentence:

> Họ nói rằng lượng hàng hoá họ nhận được có nguồn từ **Trần_Thế_Luân/location**.
>
> (They said that all the goods they received originated from Tran_The_Luan.)

Here, the word "Trần_Thế_Luân" refers to a person name rather than a location name as predicted above. In this case, the confusion may be caused

[1] The toolkit is available at: `http://taku910.github.io/crfpp`.

#Data	CRF with MIRA				Normal (offline) CRF			
	0	5	10	15	0	5	10	15
500	84.78	**85.69**	83.73	84.60	83.24	83.22	82.93	82.90
1000	85.91	**88.34**	87.11	87.11	87.67	87.59	87.67	87.14
1500	88.12	88.15	**89.16**	87.96	88.58	88.70	88.74	88.16

Table 2: Results of bootstrapping with different initial training sizes after 0, 5, 10, and 15 rounds of bootstrapping. The bold figures are the best F_1 scores with respect to a training size.

by the similar sentence structures when using a person name or a location name. For example, we can replace the word "Trần_Thế_Luân" in the sentence above by a location name and the sentence is still correct. Furthermore, in Vietnamese, many person names are used to name the locations. This also makes it more difficult to distinguish these two labels.

Another source of mistakes is the confusion between an organization name and a person name. For example, the following sentence was added during bootstrapping:

> Trong_khi_đó, **ACB/none** đang dư tiền nên đã chuyển cho **Vietbank/person** và **Kienlongbank/person**.
>
> (In the meantime, ACB is having a lot of extra money, so they transfer some to both Vietbank and Kienlongbank.)

In this example, the model could not recognize "ACB" as an organization name, and it also misclassified "Vietbank" and "Kienlongbank" as person names (ACB, Vietbank, and Kienlongbank are in fact three major banks in Vietnam). This is a difficult case since the English word "bank" is concatenated with the word "Viet" and "Kienlong", and thus it is harder to classify these words without using an external dictionary. Moreover, the sentence structure also cannot help to distinguish the two labels in this case because we can replace the three words "ACB", "Vietbank", and "Kienlongbank" by three person names and the sentence is still correct.

8 Conclusions and Future Works

We have presented preliminary results for a Vietnamese NER system trained using the CRF with MIRA and bootstrapping. We also proposed a set of useful features, which are easy to compute and do not need human work for processing unlabeled data. Our experiments showed that combining CRFs trained by MIRA with bootstrapping increases our system's performance.

For future works, we will focus on how to choose more meaningful sentences from the unlabeled data set and how to enhance the bootstrapping algorithm for the NER task. Since there are many algorithms to build our model, investigating how to combine these models in the semi-supervised learning framework to achieve better results is also a promising direction.

Acknowledgments

This work was carried out when the first author participated in an internship program at the Japan Advanced Institute of Science and Technology.

References

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Yacov Shlomo Crammer. 2004. *Online learning of complex categorical problems*. Ph.D. thesis, Hebrew University of Jerusalem.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, 31(4):531–574.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *International Conference on Computational Linguistics (COLING)*, volume 96, pages 466–471.

Heng Ji and Ralph Grishman. 2006. Data selection in semi-supervised learning for name tagging. In *Workshop on Information Extraction Beyond The Document*, pages 48–55.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

Hieu Le Trung, Vu Le Anh, and Kien Le Trung. 2014. Bootstrapping and rule-based model for recognizing Vietnamese named entity. In *Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, pages 167–176.

Gideon S. Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Jounral of Machine Learning Research*, 11:955–984.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Conference on Natural Language Learning at HLT-NAACL (CoNLL)*, pages 188–191.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 91–98.

Cam-Tu Nguyen, Xuan-Hieu Phan, and Thu-Trang Nguyen. 2005. JVnTextPro: A tool to process Vietnamese texts.

Dat Ba Nguyen, Son Huu Hoang, Son Bao Pham, and Thai Phuong Nguyen. 2010. Named entity recognition for Vietnamese. In *Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, pages 205–214.

Thi-Ngan Pham, Le Minh Nguyen, and Quang-Thuy Ha. 2012. Named entity recognition for Vietnamese documents using semi-supervised learning method of CRFs with generalized expectation criteria. In *International Conference on Asian Language Processing (IALP)*, pages 85–88.

Ellen Riloff, Rosie Jones, et al. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI Conference on Artificial Intelligence*, pages 474–479.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1524–1534.

Charles Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. *Introduction to Statistical Relational Learning*, pages 93–128.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Conference on Natural Language Learning at HLT-NAACL (CoNLL)*, pages 142–147.

Q. Tri Tran, T.X. Thao Pham, Q. Hung Ngo, Dien Dinh, and Nigel Collier. 2007. Named entity recognition in Vietnamese documents. *Progress in Informatics*, 5:14–17.

Nguyen Cam Tu, Tran Thi Oanh, Phan Xuan Hieu, and Ha Quang Thuy. 2005. Named entity recognition in Vietnamese free-text and web documents using conditional random fields. In *Conference on Some Selection Problems of Information Technology and Telecommunication*.