Named Entity Recognition in Vietnamese Text Using Label Propagation

Huong Thanh Le, Rathany Chan Sam,
Hoan Cong Nguyen
Hanoi University of Science and Technology, Vietnam
huonglt@soict.hut.edu.vn, rathany_cam@yahoo.com,
nguyenconghoan.105@gmail.com

Thuy Thanh Nguyen
University of Engineering and Technology, Vietnam
nguyenthanhthuy@vnu.edu.vn

Abstract - **This paper presents our named entity recognition system for Vietnamese text using labeled propagation. In here we propose: (i) a method of choosing noun phrases as the named entity candidates; (ii) a method to measure the word similarity; and (iii) a method of decreasing the effect of high frequency labels in labeled documents. Experimental results show that our labeled propagate method achieves higher accuracy than the old one [12]. In addition, when the number of the labeled data is small, its accuracy is higher than when using conditional random fields.**

Keywords- Named entity recognition, labeled propagation, semi-supervised learning, words similarity.

I. INTRODUCTION

Named Entity Recognition is the process of locating and classifying tokens and phrases in text into predefined categories such as the names of persons, organizations, locations, etc. Most research on NER followed the supervised learning approach [1, 2, 7, 9] which require a large hand-annotated corpus. Such approaches can achieve good performances. However, annotating such a corpus requires a lot of human effort. To resolve this problem, semi-supervised learning approach has been used in the recent years [6, 8, 10]. The idea of this approach is to train the system by using both small labeled data and big unlabeled data.

Recently, Sam et al [11] present a semi-supervised learning method for recognizing named entities in Vietnamese text by combining proper name co-reference, named-ambiguity heuristics with a powerful sequential learning model, Conditional Random Fields. 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 co-reference rules. The low confident named entities are added to the training set to learn new context features. The system is then retrained on the new data which includes low confidence NEs above. This process is repeated until the system performance cannot be improved. Their result showed that the semi-supervised learning method is better than the supervised learning method which uses only CRF. However, their method only gives a better result when the corpus contains named co-reference words. If we can not find the new training data by using name co-reference rules, the method will become supervised learning one. Furthermore, the iteration of learning process makes the complexity of algorithm high and takes a lot of time.

Recently, labeled propagation was used as the semi-supervised method to find the relation between two entities [3] and got a good result. However, as far as we known, there are no work that use labeled propagation for named entity recognition.

From that point of view, we propose to use labeled propagation in recognizing named entities from Vietnamese text. The important part of labeled propagation is generating the matrix T (see Fig. 1) in which each element is a weight of the similarity among the data. In NER using labeled propagation, each node of the graph corresponds to a word. Each row in the matrix Y (see Fig. 2) corresponds to a word and each column is an entity label of that word. In our method, matrix T is built by measuring the similarity between words.

Since entities are usually nouns or noun phrases, to reduce the number of unnecessary nodes or rows in the matrix, only nouns and noun phrases are chosen by us as the graph's node. The word contexts are not lost by this treatment since it is already used when we measure the similarity between two words in the matrix T. When we do like this we can reduce computation time.

II. NAMED ENTITY RECOGNITION USING LABELED PROPAGATION

A. Labeled Propagation

In the labeled propagation method, the labeled data and unlabeled data are presented as vertices in a connected graph. A node's labels propagate to neighboring nodes according to their proximity, while we clamp the labels on the labeled data. Suppose that we have a graph G = (V, E); the set of node V represented the training data; the set of edges E represented the similarity between data. The similarity between these data is given by a matrix T, in which t_{ij} is the weight of the edge (i, j) between two neighbor nodes x_i and x_j.

$$\begin{bmatrix} t_{1,1} & t_{1,2} & t_{1,3} & \cdots & t_{1,n} \\ t_{2,1} & t_{2,2} & t_{2,3} & \cdots & t_{2,n} \\ t_{3,1} & t_{3,2} & t_{3,3} & \cdots & t_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{n,1} & t_{n,2} & t_{n,3} & \cdots & t_{n,n} \end{bmatrix}$$

Figure 1. Matrix T: $t_{i,j}$ is the similarity weight between two data. n is the number of data (unlabeled and labeled data)

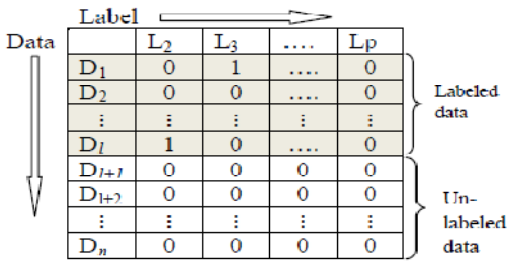

Figure 2. Matrix Y

Starting from the labeled nodes from 1 to l and unlabeled nodes from l+1 to n, each node propagates its label information to its neighbors through weighted edges, this process continues until a global stable stage is achieved. The higher the weight is, the easier the label goes through. Therefore, the more similar label has the higher weight. The label propagation algorithm proposed by the authors in [12] is presented in Fig. 3. In this algorithm, both labeled data and unlabeled data are nodes of a graph. Edges of the graph represent the similarity weight between two data. Matrix Y presents the relation between data and label. Matrix T presents the similarity between two nodes.

In the label propagation process, the initial labeled data (hand-annotated corpus) is retained in each iteration to provide the source label. It means that in each iteration, l rows of the matrix Y have the same values as l rows of the initial matrix Y_0.

Step 1: Initialization
- Set the iteration index t=0
- Let Y^0 be the initial matrix. Y^0 =1 if y_i has the label r_j, otherwise Y^0 =0.
- Let Y_L^0 be the top l rows of Y^0 corresponding to l hand annotated data and Y_U^0 be the remained rows corresponding to u un-annotated data

Step 2: Propagate labels of any node to their neighbor by $Y^{t+1} = \overline{T}Y^t$, where $\overline{T}$ is the normalize matrix of T.

Step 3: Retain the initial hand annotated data. Replace the top l rows of Y^{t+1} with Y_L^0.

Step 4: Repeat steps 2 and 3 until Y converges

Step 5: Assign x_h ($l+1 \leq h \leq n$) with a label: $y_h = \text{argmax}_j Y_{hj}$

Figure 3. Labeled propagation algorithm in [12]

B. Measuring the Word Similarity

As the purpose of measuring word similarity in this research is for named entity recognition, the context surrounding the entity has a very important role.

To measure the similarity between two words A_i and B_i, words in the context window size of 7 (three words on the left and three words on the right of the current token) are examined. Therefore, the similarity between two words A_i and B_i become the similarity between two sets $\{A_{-3}, A_{-2}, A_{-1}, A_i, A_{+1}, A_{+2}, A_{+3}\}$ and $\{B_{-3}, B_{-2}, B_{-1}, B_i, B_{+1}, B_{+2}, B_{+3}\}$. We noticed that if this method is applied to measure the similarity between two words in the same sentences, the difference between them are too small. As a result, the effectiveness of the proposed measures is not high. Therefore, we only measure the similarity between two words in different sentences.

The idea of this algorithm is: first, the input data (unlabeled and labeled data) are preprocessed to get its part of speech and orthographic labels. Then, by identifying the pair of words A_i and B_i and the context surrounding them, we get two context sets. Each word in the first context set will be matched with all words in the second context set based on the following process.

(1) Initial Sim_w=0
(2) Check whether the two words have the same part of speech. If it is true $Sim_w = Sim_w+1$
(3) Check whether the two words have the same spelling. If it is true $Sim_w=Sim_w+1$
(4) Check whether the two words have the same orthographic. If it is true $Sim_w = Sim_w+1$
(5) Calculate the similarity between two words based on semantic tree, SimS, and add it to Sim_w

Our proposed algorithm is shown in Fig. 4 below.

Input: Two words W, W' and the context around two words
Output: The similarity between two words Sim_W (W, W')
Method:
1: Initial
- Sim_w=0
- T = number of word in the first context set
- T'= number of word in the second context set

2: Preprocessing
- Identify the part of speech of all word
- Identify the orthographic of all word

3: Calculate the similarity based on the context around two words
For i=1 to N1 do
For j=1 to N2 do
{
1. If a[i] and a[j] have the same part of speech then $Sim_w = Sim_w + 1$
2. If the word a[i] is the same to the word a[j] then $Sim_w = Sim_w+1$
3. if a[i] and a[j] have the same orthographic then $Sim_w = Sim_w+1$
4. Calculate the similarity between two words based on semantic tree SimS
5. $Sim_w = Sim_w + SimS$
}

Figure 4. Our proposed algorithm to measure the similarity between two words

Calculating the similarity between two words based on a semantic tree

Wordnet is an useful resource to calculate the similarity between words. However, there is no Wordnet for Vietnamese. Instead, we only have a semantic tree that represents the semantic relation between concepts and a Vietnamese word dictionary that describes words and the concepts they belong to. Both of these resources are created by Vietnam Lexicography Centre (Vietlex - http://www.vietlex.com). To calculate the similarity between words, we combine information taken from these two resources.

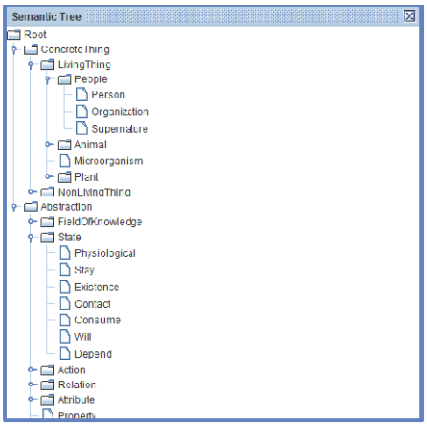

Figure 5. Structure of the semantic tree

To calculate the similarity between two words c1 and c2, the system allocates these words in the word dictionary and then calculates their semantic similarity by the following formula:

$$SimS(c1, c2) = 1/dist\ (c1, c2),$$

in which, dist (c1,c2) is the distance between the conceptual classes of the words c1 and c2 on the semantic tree.

For example, the word "con trai" (sun) belongs to Person class; the word "con mèo" (cat) belongs to Animal class in the dictionary. And dist(Person, Animal), the distance between Person class and Animal class in the Vietnamese semantic tree. is 2. Therefore, the semantic similarity between these words is ½. The word "nông dân" (farmer) and the word "công dân" (citizen) belong to the same class Person in the dictionary. Therefore, the distance between these two words is 1.

C. The Proposed Labeled Propagation algorithm

One disadvantage of the original labeled propagation is that the output of labeled propagation method is influenced by the label that have high frequency. That is, when carrying out the matrix multiplication in the original labeled propagation algorithm, many entities are assigned with the high frequency labels although they should be assigned with other labels.

Our proposed solution to this problem is as follow. Instead of assigning 1 to elements on the top of matrix Y_0, these elements are assigned by 1- n_X/l, in which n_X is the number of words that have label X and we want to assign value to element at column that have label X. l is the number of labeled data.

For example:

The number of labeled data is 1276 words, in which 12 words are B-per, 5 words are I-per, 15 words are B-loc, 12 words are I-loc, 20 words are B-org, 12 words are I-org, and 1200 words are Other. So in matrix Y_0, the other labeled words are assigned by 1-1200/1276, the I-per labeled words are assigned by 1-5/1276, the B-per labeled words are assign by 1-12/1276, .etc.

In other words, the weight of a label is inversely proportion with their frequency. By carrying our this proposed solution, the problem of dominant labels in labeled propagation algorithm is solved.

III. EXPERIMENTS AND DISCUSSION

The data set used in our experiments consists of 950 documents (850 unlabeled documents, and 100 documents labeled manually). All of these documents were taken from newspaper websites on economics, politics, cultures and education.

These documents were first processed by a phrase chuncker and then converted to BIO form (B is the begining of the class, I is the inside of the class, O is the outside of the class). By using BIO form, 7 labels are used in our data, which are: B_Per, I_Per, B_Org, I_Org, B_Loc, I_Org, and Other. Because of that, matrix Y has 7 columns and n rows (n is the number of word in the data after choosing noun phrase). Matrix T has n columns and n rows. When calculating the similarity between two words, the window size of 7 (the current word, three words on the left, and three words on the right of the current word) was used.

Three experiments were carried out: (i) using supervised CRF, (ii) using the original labeled propagation, and (iii) using our improved labeled propagation. These experiments were evaluated based on Precision, Recall, and F-measure, in which:

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

In order to compare the original labeled propagation with our improved labeled propagation, 100 labeled documents and 850 unlabelled documents were used. The results are shown in Table I.

Table I shows that the F-measure values of Person, Location and Organization when using our labeled propagation (81.25%, 81.69%, and 70.16%, respectively) are higher than those when using the original one (73.69%, 80.88%, and 49.20% respectively).

TABLE I. COMPARE THE ORIGINAL LABELED PROPAGATION TO OUR LABELED PROPAGATION

Entity Type	Original labeled propagation			Our labeled propagation		
	P	***R***	***F***	***P***	***R***	***F***
Person	100	58.33	73.69	92.85	72.22	81.25
Location	90.16	73.33	80.88	86.56	77.33	81.69
Organization	88.57	34.06	49.20	55.44	95.53	70.16

To compare supervised CRF with our improved labeled propagation, 100 labeled documents were divided into 4 packages (each package has 25 documents). Four experiments were carried out: using one package as labeled documents; using two packages as labeled documents; using three packages as labeled documents and using four packages as labeled documents. All these experiments used the same unlabeled documents (850 unlabeled documents). The results are shown in Table 2 below.

TABLE II. COMPARE OUR LABELED PROPAGATION TO SUPERVISED CRF

Method	Data	Person	Location	Organization
Labeled Propagation	25	46.15	63.86	63.91
	50	77.27	73.28	68.29
	75	79.36	74.24	70.12
	100	81.25	81.69	70.16
CRF	25	43.64	52.94	41.67
	50	71.65	55.74	49.16
	75	73.97	79.52	80.00
	100	86.57	89.16	89.72

Table II shows that when the number of labeled document is from 25 to 50, F-measures of supervised CRF are smaller than F-measures of our labeled propagation. However, when the number of labeled documents increases (from 70 to 100), F-measures of CRF improve gradually compared to the labeled propagation method. This is consistent with the nature of these two methods: the labeled propagation method does not learn rules from data, it only bases on the similarity among data. Meanwhile, the CRF learns principles of the annotated data to label the un-annotated data. Because of that reason, when the size of labeled data is too small, the CRF system does not have enough information to identify the rule of data. As a result, the system accuracy in this case is low. When the size of labeled data increases, the CRF system identifies the rule of data more accurately. As a result, the system accuracy in this case is increased. With the labeled propagation method, when the labeled data increases, the system can find labeled data closer to the new data, so the accuracy of system also increases. However, since this method bases only on the similarity among data instead of based on the rules of data, it can not provide a good result like CRF. Our experiment results indicate that the labeled propagation method is good when the labeled data is small; otherwise CRF is a better choice.

IV. CONCLUSION

In this paper, we have proposed an approach to named entity recognition by semi-supervised learning on Label Propagation. The contributions of this paper are: (i) proposing a method to measure the similarity between words; (ii) reducing computation time by choosing only nouns or noun phrases as candidate named entities; and (iii) proposing a method to reduce the influence of high frequency labels to the label propagation process. Our experiments were carried out with three NER algorithms: CRFs; the original labeled propagation; and our proposed labeled propagation. The experimental results show that our labeled propagation method outperforms the original one in [12]. In addition, the proposed labeled propagation method provides better results comparing to CRF when the data size is small. When the data size is large, the CRF is better. Our future work includes: (i) carrying out experiments with a larger corpus to get a better evaluation of the proposed method; and (ii) experimenting the system with other entity types.

ACKNOWLEDGMENTS

This work was supported by the Vietnam Ministry project, under Grant B2012 – 01 - 24.

REFERENCES

[1] A. Borthwick, "Maximum Entropy Approach to Named Entity Recognition," Ph.D. thesis, New York University, 1999.

[2] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, H. Cunningham, "Shallow methods for named entity conference resolution," in Proc. of TALN 2002 Workshop, Nancy, France, 2002.

[3] J. Chen, D. Ji, L.C. Tan, & Z. Niu, "Relation Extraction Using Label Propagation Based Semi-supervised Learning," in proceeding of 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistic, 129-136, 2006.

[4] M.B. David, Y.N. Andrew and I.J. Michael, "Latent Dirichlet allocation," Journal of Machine Learning Research 3: pp. 993-1022, January 2003.

[5] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, Vol. 42, No. 1-2, pp. 177-196, 2001.

[6] W. Liao and S. Veeramachaneni, "A Simple Semi-supervised Algorithm for Named Entity Recognition," in Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pages 28–36, 2009.

[7] A. McCallum and W. Li, "Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons," in Proceedings of CoNLL, pages 188.191, Canada, 2003.

[8] B. Mohit and R. Hwa, "Syntax-based semi-supervised Named Entity Tagging," in Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 57–60, Michigan, 2005.

[9] T.H. Nguyen and T.H. Cao, VN-KIM IE: Automatic extraction of Vietnamese named-entities on the Web. *Journal of new Generation Computing*, 25(3):277-292.

[10] C. Niu, W. Li, J. Ding and K.S. Rohini, "A Bootstrapping Approach to Named Entity Classification Using Successive Learner," in Proceedings of the 41st AnnualMeeting of the ACL, pp. 335–342 (2003).

[11] R.C. Sam, H.T. Le, T.T. Nguyen and T.H. Nguyen, "Combining Proper Name-Coreference with Conditional Random Fields for Semi-supervised Named Entity Recognition in Vietnamese Text," The *15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2011, Shenchen, China, pp. 512-525.

[12] Z. Xiaojin and Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation," CMU CALD tech report CMU-CALD-02-107.

[13] Q.T. Tran, T.X.T. Pham, Q.H. Ngo, D. Dinh, and N. COLLIER, "Named entity recognition in Vietnamese using classifier voting," in *ACM Transactions on Asian Language Information Processing* (TALIP).

[14] Y. Wong and T. Ng.Hwee, "One Class per Named Entity: Exploiting Unlabelled Text for Named Entity Recognition," in Proceedings of IJCAI, pp. 1763–1768 (2007).