Chinese Named Entity Recognition with Graph-based Semi-supervised Learning Model

Aaron Li-Feng Han* Xiaodong Zeng+ Derek F. Wong+ Lidia S. Chao+

* Institute for Logic, Language and Computation, University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands
+ NLP2CT Laboratory/Department of Computer and Information Science
University of Macau, Macau S.A.R., China
l.han@uva.nl nlp2ct.samuel@gmail.com derekfw@umac.mo lidiasc@umac.mo

Abstract

Named entity recognition (NER) plays an important role in the NLP literature. The traditional methods tend to employ large annotated corpus to achieve a high performance. Different with many semi-supervised learning models for NER task, in this paper, we employ the graph-based semi-supervised learning (GBSSL) method to utilize the freely available unlabeled data. The experiment shows that the unlabeled corpus can enhance the state-of-the-art conditional random field (CRF) learning model and has potential to improve the tagging accuracy even though the margin is a little weak and not satisfying in current experiments.

1. Introduction

Named entity recognition (NER) can be regarded as a sub-task of the information extraction, and plays an important role in the natural language processing literature. The NER challenge has attracted a lot of researchers from NLP, and some successful NER tasks have been held in the past years. The annotations in MUC-7[1] Named Entity tasks (Marsh and Perzanowski, 1998) consist of entities (organization, person, and location), times and quantities such as monetary values and percentages, etc. among the languages of English, Chinese and Japanese.

The entity categories in CONLL-02 (Tjong Kim Sang, 2002) and CONLL-03 (Tjong Kim Sang and De Meulder, 2003) NER shared tasks consist of persons, locations, organizations and names of miscellaneous entities, and the languages span from Spanish, Dutch, English, to German.

The SIGHAN bakeoff-3 (Levow, 2006) and bakeoff-4 (Jin and Chen, 2008) tasks offer standard Chinese NER (CNER) corpora for training and testing, which contain the three commonly used entities, i.e., personal names, location names, and organization names. The CNER task is generally more difficult than the western languages due to the lack of word boundary information in Chinese expression.

Traditional methods used for the entity recognition tend to employ external annotated corpora to enhance the machine learning stage, and improve the testing scores using the enhanced models (Zhang et al., 2006; Mao et al., 2008; Yu et al., 2008). The conditional random filed (CRF) models have shown advantages and good performances in CNER tasks as compared with other machine learning algorithms (Zhou et al., 2006; Zhao and Kit, 2008), such as ME, HMM, etc. However, the annotated corpora are generally very expensive and time consuming.

On the other hand, there are a lot of freely available unlabeled data in the internet that can be used for our researches. Due to this reason, some researchers begin to explore the usage of the unlabeled data and the semi-supervised learning methods based on labeled training data and unlabeled external data have shown their advantages (Blum and Chawla, 2001; Shin et al., 2006; Zha et al., 2008; Zhang et al., 2013).

[1] http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 15–20,
Beijing, China, July 30-31, 2015.

2. Semi-supervised Learning

In the semi-supervised learning model, a sample $\{Z_i = (X_i, Y_i)\}_{i=1}^{n_l}$ is usually observed with labeling $Y_i \in \{-1,1\}$, in addition to independent unlabeled samples $\{X_j\}_{j=n_l+1}^{n}$ with the $n = n_l + n_u$. The $X_k = (X_{k1}, X_{k2}, \dots, X_{kp})$ $k \in (1, n)$ is a p-dimentional input (Wang and Shen, 2007). The labeled samples are independently and identically distributed according to an unknown joint distribution $P(x, y)$, and the unlabeled samples are independently and identically distributed from distribution $P(x)$. Many semi-supervised learning models are designed through some assumptions relating $P(x)$ to the conditional distribution, which cover EM method, Bayesian network, etc. (Zhu, 2008).

The graph-based semi-supervised learning (GBSSL) methods have been successfully employed by many researchers. For instance, Goldberg and Zhu (2006) design the GBSSL model for sentiment categorization; Celikyilmaz et al. (2009) propose a GBSSL model for question-answering; Talukdar and Pereira (2010) use the GBSSL methods for class-Instance acquisition; Subramanya et al. (2010) utilize the GBSSL model for structured tagging models; Zeng et al., (2013) use the GBSSL method for the joint Chinese word segmentation and part of speech (POS) tagging and result in higher performances as compared with previous works. However, as far as we know, the GBSSL method has not been employed into the CNER task. To testify the effectiveness of the GBSSL model in the traditional CNER task, this paper utilizes some unlabeled data to enhance the CRF learning through GBSSL method.

3. Designed Models

To briefly introduce the GBSSL method, we assume $D_l = \{(x_j, r_j)\}_{j=1}^{l}$ denote l annotated data and the empirical label distribution of x_j is r_j. Assume the unlabeled data types are denoted as $D_u = \{x_i\}_{i=l+1}^{m}$. Then, the entire dataset can be represented as $D = D_u \cup D_l$. Let $G = (V, E)$ corresponds to an undirected graph with V as the vertices and E as the edges. Let V_l and V_u represent the labeled and unlabeled vertices respectively. One important thing is to select a proper similarity measure to calculate the similarity between a pair of vertices (Das and Smith, 2012). According to the smoothness assumption, if two instances are similar according to the graph, then the output labels should also be similar (Zhu, 2005).

There are mainly three stages in the designed models, i.e., graph construction, label propagation and CRF learning. Graph construction is performed on both labeled and unlabeled data, and the unlabeled data is automatically tagged through the label propagation stage. Then, the tagged external data will be added into the manually annotated training corpus to enhance the CRF learning model.

3.1 Graph Construction & Label Propagation

We follow the research of Subramanya et al. (2010) to represent the vertices using character trigrams in labeled and unlabeled sentences for graph construction.

A symmetric k-NN graph is utilized with the edge weights calculated by a symmetric similarity function designed by Zeng et al. (2013).

The feature set we employed to measure the similarity of two vertices based on the co-occurrence statistics is the optimized one by Han et al. (2013) for CNER tasks, as denoted in Table 1.

Feature	Meaning
$U_n, n \in (-4,2)$	Unigram, from previous 4th to following 2nd character
$B_{n,n+1}, n \in (-2,1)$	Bigram, 4 pairs of features, from previous 2nd to following 2nd character

Table 1: Feature set for measuring vertices similarity in graph construction and training CRF model.

After the graph construction on both labeled and unlabeled data, we use the sparsity inducing penalty (Das and Smith, 2012) label propagation algorithm to induce trigram level label distributions from the constructed graph, which is based on the Junto toolkit (Talukdar and Pereira, 2010).

3.2 CRF Training

In the CRF model, assume a graph $G = (V, E)$ comprising a set V of vertices or nodes together with a set E of edges or lines and $Y = \{Y_v | v \in V\}$ so Y is indexed by the vertices of G. The joint distribution over the label sequence Y given X is presented as the form:

$$P_\theta(y|x) \propto exp\left(\sum_{e\in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v\in V,k} \mu_k g_k(v, y|_v, x)\right)$$

The f_k and g_k are the feature functions and μ_k and λ_k are the parameters that are trained from specific dataset (Lafferty et al., 2001). The feature set employed in the CRF learning is also the optimized one as shown in Table 1. The training method utilized for the CRF model is a quasi-newton algorithm[2]. The automatically annotated corpus by the graph based label propagation will affect the trained parameters μ_k and λ_k.

4. Experiments

4.1 Data

We employ the SIGHAN bakeoff-3 (Levow, 2006) MSRA (Microsoft research of Asia) training and testing data as standard setting. To testify the effectiveness of the GBSSL method for CRF model in CNER tasks, we utilize some plain (unannotated) text from SIGHAN bakeoff-2 (Emerson, 2005) and bakeoff-4 (Jin and Chen, 2008) as external unlabeled data. The data set is introduced in Table 2 from the aspect of sentence number.

	Bakeoff-3 Corpus		External
Sentence Number	Training	Testing	Unlabeled
	50,425	4,365	31,640

Table 2: Corpus Information.

4.2 Result Analysis

We set two baseline scores for the evaluation. One baseline is the simple left-to-right maximum matching model (MaxMatch) based on the training data, another baseline is the closed CRF model (Closed-CRF) without using unlabeled data. The employment of GBSSL model into semi-supervised CRF learning is denoted as GBSSL-CRF.

The training costs of the CRF learning stage are detailed in Table 3. The comparison shows that the extracted features grow from 8,729,098 to 11,336,486 (29.87%) due to the external dataset, and the corresponding iterations and training hours also grow by 12.86% and 77.04% respectively.

	Training Costs		
	Feature	Iteration	Time (h)
Closed-CRF	8,729,098	350	4.53
GBSSL-CRF	11,336,486	395	8.02

Table 3: Training Cost for CRF Learning.

The evaluation results are shown in Table 4, from the aspects of recall, precision and the harmonic mean of recall and precision (F1-score). The evaluation shows that both the Closed-CRF and GBSSL-CRF models have largely outperformed baseline-1 model (MaxMatch). As compared with the Closed-CRF model, the GBSSL-CRF model yielded a higher performance in precision score, a lower performance in recall score, and finally resulted in a faint improvement in F1 score. Both the GBSSL-CRF and Closed-CRF show higher performance in precision and lower performance in recall value.

	Evaluation Scores		
Total-score	Total-R	Total-P	Total-F
MaxMatch	48.8	59.0	53.4
Closed-CRF	**77.95**	90.27	83.66
GBSSL-CRF	77.84	**90.62**	**83.74**

Table 4: Evaluation Results.

To look inside the GBSSL performance on each kind of entity, we denote the detailed evaluation results from the aspect of F1-score in Table 5. The detailed evaluation from three kinds of entities shows that both the GBSSL-CRF and Closed-CRF show higher performance in LOC entity type, and lower performance in PER and ORG entities.

	Detailed Evaluation		
Sub-F-score	PER-F	LOC-F	ORG-F
MaxMatch	61.4	53.1	46.9
Closed-CRF	77.95	**88.56**	80.88
GBSSL-CRF	**78.17**	88.39	**81.35**

Table 5: Detailed Evaluation Results.

Fortunately, the GBSSL model can enhance the CRF learning on the two kinds of difficult entities PER and ORG with the better performances of 0.28% and 0.58% respectively. However, the GBSSL model decreases the F1 score

[2] http://www.nag.com/numeric/fl/nagdoc_fl23/html/E04/e04conts.html

on LOC entity by 0.19%. The lower performance of GBSSL model on LOC entity may be due to that the unlabeled data is only as much as 62.75% of the training corpus, which is not large enough to cover the Out-of-Vocabulary (OOV) testing words of LOC entity; on the other hand, the unlabeled data also bring some noise into the model.

5. Related Work

Nadeau (2007) employs the semi-supervised learning method to recognize 100 entity types on English documents with little supervision. Similarly, Liao and Veeramachaneni (2009) propose a simple semi-supervised algorithm for English entity recognition. Liu et al. (2011) design an interesting application of the semi-supervised learning model for online tweets document for English NER.

Pham et al. (2012) use semi-supervised learning method of CRFs into the Vietnamese NER task with generalized expectation criteria. Similarly, Vo and Ock (2012) utilize a hybrid approach semi-supervised learning approach into the NER task for Vietnamese document.

Wang et al. (2013) and Che et al. (2013) recently propose the usage of bilingual constraints to enhance the NER accuracy.

Some advanced technologies of GBSSL methods are introduced in the papers Zhu and Lafferty (2005), Culp and Michailidis (2008), and Zhang and Wang (2011), etc.

6. Conclusion and Future Work

This paper makes an effort to see the effectiveness of the GBSSL model for the traditional CNER task. The experiments verify that the GBSSL can enhance the state-of-the-art CRF learning models. The improvement score is a little weak because the unlabeled data is not large enough. In the future work, we decide to use larger unlabeled dataset to enhance the CRF learning model.

The feature set optimized for CRF learning may be not the best one for the similarity calculation in graph construction stage. So we will make efforts to select the best feature set for the measuring of vertices similarity in graph construction on CNER documents.

In this paper, we utilized the Microsoft research of Asia corpus for experiments. We will use more kinds of Chinese corpora for testing, such as CITYU and LDC corpus, etc.

The GBSSL model generally improves the tagging accuracy of the Out-of-Vocabulary (OOV) words in the test data, which are unseen in the training corpora. In the future work, we plan to give a detailed analysis of the GBSSL model performance on the OOV words for CNER tasks.

Acknowledgements

This work was supported by the Research Committee of the University of Macau (Grant No. MYRG2015-00175-FST and MYRG2015-00188-FST) and the Science and Technology Development Fund of Macau (Grant No. 057/2014/A). The first author was supported by NWO VICI under Grant No. 277-89-002.

References

A. Blum, & Chawla, S. 2001. Learning from labeled and unlabeled data using graph mincuts. In *ICML*-2001.

Asli Celikyilmaz, Marcus Thint, and Zhiheng Huang. 2009. A graph-based semi-supervised learning for question-answering. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2 (ACL '09),* Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 719-727.

Wanxiang Che, Mengqiu Wang and Christopher D. Manning. 2013. Named Entity Recognition with Bilingual Constraints. In *NAACL* 2013.

Mark Culp and George Michailidis. 2008. Graph-Based Semisupervised Learning. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 30, NO. 1.

Dipanjan Das and Noah A. Smith. 2012. Graph-based lexicon expansion with sparsity-inducing penalties. In *Proceedings of NAACL*, pages 677-687.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. Pp. 123-133.

Andrew B. Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing (TextGraphs-1)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 45-52.

Aaron Li-Feng Han, Derek F. Wong, and Lidia S. Chao. 2013. Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics. In *Language Processing* *and Intelligent Information Systems. Lecture Notes in Computer Science*, Volume 7912, pp 57-68.

J. Lafferty, McCallum, A., Pereira, F.C.N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceeding of 18th International Conference on Machine Learning*, pp. 282–289. Massachusetts.

Gina-Anne Levow. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney.

Wenhui Liao and Sriharsha Veeramachaneni. 2009. A Simple Semi-supervised Algorithm For Named Entity Recognition. *Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*, pages 58–65.

Xiaohua Liu , Shaodian Zhang , Furu Wei , Ming Zhou. 2011. Recognizing Named Entities in Tweets. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

G. Jin, and X. Chen. 2008. The fourth international CLP bakeoff: Chinese word segmentation, named entity recognition and Chinese pos tagging. In: *Sixth SIGHAN Workshop on CLP*, pp. 83–95.

Elaine Marsh, and Dennis Perzanowski. 1998. MUC-7 Evaluation of IE Technology: Overview of Results. Technical report.

Xinnian Mao; Yuan Dong; Saike He; Sencheng Bao; Haila Wang. 2008. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*. pp. 90-93.

David Nadeau. 2007. Semi-Supervised Named Entity Recognition-Learning to Recognize 100 Entity Types with Little Supervision. PHD thesis. University of Ottawa

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 *Proceedings of the 2012 International Conference on Asian Language Processing*. IEEE Computer Society, Washington, DC, USA, 85-88.

Hyunjung Shin, N. Jeremy Hill, and Gunnar R¨atsch. 2006. Graph Based Semi-Supervised Learning with Sharper Edges. *ECML* 2006, LNAI 4212, pp. 402–413.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 167-176.

Partha Pratim Talukdar and Fernando Pereira. 2010. Experiments in Graph-based Semi-Supervised Learning Methods for Class-Instance Acquisition. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1473–1481.

E. F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *CONLL*-02.

E. F. Tjong Kim Sang; De Meulder, Fien. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *CoNLL*-2003.

Duc-Thuan Vo, Cheol-Young Ock. 2012. A Hybrid Approach of Pattern Extraction and Semi-supervised Learning for Vietnamese Named Entity Recognition. *Lecture Notes in Computer Science* Volume 7653, 2012, pp 83-93.

Junhui Wang and Xiaotong Shen. 2007. Large Margin Semi-supervised Learning. *Journal of Machine Learning Research*.

Mengqiu Wang, Wanxiang Che, Christopher D. Manningy. 2013. Effective Bilingual Constraints for Semi-supervised Learning of Named Entity Recognizers. In *AAAI*-2013.

Xiaofeng Yu; Wai Lam; Shing-Kit Chan; Yiu Kei Wu; Bo Chen. 2008. Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*. pp. 102-105.

Xiaodong Zeng; Derek F. Wong; Lidia S. Chao; Isabel Trancoso. 2013. Graph-based Semi-Supervised Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. *ACL* 2013.

Zheng-Jun Zha; Tao Mei; Jingdong Wang; Zengfu Wang; Xian-Sheng Hua. 2008. Graph-based semi-supervised learning with multi-label. *Multimedia and Expo, 2008 IEEE International Conference on*, pp.1321-1324.

Changshui Zhang, Fei Wang. 2011. Graph-based semi-supervised learning. *Frontiers of Electrical and Electronic Engineering in China*. Volume 6, Issue 1, pp 17-26.

Suxiang Zhang; Ying Qin; Juan Wen; Xiaojie Wang. 2006. Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 158–161.

Tongtao Zhang, Rongrong Ji, Wei Liu, Dacheng Tao, and Gang Hua. 2013. Semi-supervised learning with manifold fitted graphs. In *Proceedings of the*

Twenty-Third international joint conference on Artificial Intelligence (IJCAI'13), Francesca Rossi (Ed.). AAAI Press 1896-1902.

Hai Zhao; Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*. pp. 106-111.

Junsheng Zhou; Liang He; Xinyu Dai; Jiajun Chen. 2006. Chinese Named Entity Recognition with a Multi-Phase Model. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 213–216.

Xiaojin Zhu. 2005. Semi-Supervised Learning with Graphs. PHD thesis. CMU-LTI-05-192.

X. Zhu, & Lafferty, J. 2005. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *ICML*-2005.

Xiaojin Zhu. 2008. Semi-Supervised Learning Literature Survey. University of Wisconsin.