The Improvement of Biomedical Named Entity Recognition with Semi-joint labeling presentation

Justin Liang-Te Chiu[1,2], Hong-Jie Dai[1,3], Richard Tzong-Han Tsai[4], Chi-Hsin Huang[1]
[1]*Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.*
[2]*Dept. of Computer Science & Information Engineering, National Taiwan Univ., Taipei, Taiwan, R.O.C.*
[3]*Dept. of Computer Science, National Tsing-Hua Univ., Hsinchu, Taiwan, R.O.C.*
[4]*Dept. of Computer Science & Engineering, Yuan Ze Univ., Taoyuan, Taiwan, R.O.C.*
Justin Liang-Te Chiu: thirddawn@iis.sinica.edu.tw
Hong-Jie Dai: hongjie@iis.sinica.edu.tw
Richard Tzong-Han Tsai: thtsai@saturn.yzu.edu.tw
Chi-Hsin Huang: sinyuhgs@iis.sinica.edu.tw

Abstract*-Named entity recognition (NER) is an important step for the information retrieval. In biomedical field, due to the fact that there is no community-wide agreement on how a particular name entities (NEs) should be named. To improve the performance of NER in biomedical field, this paper proposed to use a new tagging presentation, semi-joint labeling. This presentation works by adding chunking information on the tags which are not NEs. With the proposed presentation, we improve the F-score from 62.37% with IOB2 tagging presentation to 63.33%, and right and left boundary matches improves 0.88% and 0.56% respectively.*

Keywords: Named Entity Recognition, Machine Learning, Natural Language Processing

1. Introduction

In recent years, biomedical researches became a rising star in scientific studies. Hence, the numbers of biomedical literature in the large-scale databases, such as PubMed, are growing with an exponential speed. To retrieve the information from these huge numbers of biomedical literature, named entity recognition (NER) becomes an indispensable part in the information retrieval (IR) procedure. With the tools for assisting the biomedical researchers, they exploiting the stream of publications at a rate of 1,500 abstract a day. [1]

NER is a fundamental task which involves identifying some specific words or phrases in text and classifying them into different categories. NER was first defined in the general-language domain during the Message Understanding Conference [2]. The specialized domain NER started to get attentions in recent years. For natural language processing (NLP) researchers, a new domain of NER means new challenge. For example, in NER for the biomedical domain, set of entities is often restricted to biomedical name entities (NEs), such as IL-2 (protein), p53 (DNA/protein.) On the contrary, in general-language domain, each kind of entity is usually far from each other, such as persons, locations and organizations.

At the beginning of NER in biomedical field, handcrafted patterns [3] were proposed to recognize the different NE forms. Nevertheless, the main disadvantage is lacking of portability and scalability. Later, machine learning (ML) models were introduced to handle NER problems in biomedical field. ML base approaches are divided into two categories: classifier-based and sequence-model-based. The former is famed for naïve Bayes classifiers and Support Vector Machines (SVM) [4], while the latter is hidden Markov models (HMM) [5], Maximum Entropy Markov Models (MEMM) [6], and Conditional Random Fields (CRFs) [7].

In recent years, solving NER problems can split into few steps. First, the input sentence is broken into tokens. Then, each token will be assigned with a tag to identify its NE category and position. In biomedical NER tasks, many researches choose the IOB2 tag presentation [8, 9]. In this presentation, each word in the same sentence is treated as a token, and each token has a tag associated with the categories of NE. The tag depends on the position of the token within a NE, while "B" means beginning, "I" means inside and "O" means outside of NE. For example, the

sentence “SV40 early genes induce neoplastic properties in serous borderline ovarian tumor cells” has the following tags associated with each token:

SV40$_{\text{/B-DNA}}$ early$_{\text{/I-DNA}}$ genes$_{\text{/I-DNA}}$ induce$_{\text{/O}}$ neoplastic$_{\text{/O}}$ properties$_{\text{/O}}$ in$_{\text{/O}}$ serous$_{\text{/O}}$ borderline$_{\text{/O}}$ ovarian$_{\text{/O}}$ tumor$_{\text{/O}}$ cells$_{\text{/O}}$.$_{\text{/O}}$

In this paper, we propose a new presentation, the semi-joint labeling, for biomedical NER task. The presentation injects chunking information into the aforementioned IOB2 tag sets to improve the NER performance.

The remainders of this paper are organized as follows. In section 2, we describe the methods we use in our experiment. In section 3, we show the results of our experiment. Then, in section 4, we discuss the effectiveness of the semi-joint labeling formulation. Finally, section 5 presents our conclusions.

2. Methods

2.1. The Conditional Random Fields

CRFs are undirected graphical models. Each node of it represents a state trained to maximum a conditional probability [10], while each edge represents a dependency between two random variables.

The reasons why we choose CRF as our frameset instead of other sequence-model-based are as follow. Comparing with HMM, CRFs is conditional nature, which means that it allows us for the relaxation of independence assumptions, while HMM requires to ensure tractable inference. Furthermore, CRFs avoid the label bias problem due to it’s an undirected graphical model. Last but not least, CRFs perform better than both MEMM and HMM in other sequence label tasks [11-13]. Therefore, we choose CRFs as our ML model to examine our semi-joint labeling formulation.

2.2. Feature set

In order to avoid computation complex, the features chose in this work are the most significant features used by other works[7, 8, 14]. In the follow subsections, we describe them in detail.

2.2.1. Word features. Word feature is decided by the words proceed or follow the target word. It might be useful to determine whether the target word is NE or not. We set the content window size from -2 to 0, which is, the word before the previous word, the previous word and the current word.

2.2.2. Affix features. The Affix which included prefix and suffix are morphemes. They are mainly attached to some basic morphemes such as roots or stems to form word. In biomedical field, some of them can be used to determine named entity. For instance, words which are ended with “~ase” are usually a protein name. They length we used for prefix is 2 and 4, but for suffix is only 2.

2.2.3. Part-of-speech features. Part-of-speech (POS) is a category of word which is defined by morphological and syntactic behavior of the lexical item. It is useful in identifying NEs. We use GENIA tagger [15] to obtain our POS information

2.2.4. Conjunction features. We include a conjunction feature which will take four of our POS tagging as one feature in our experiment. This may help us find out some special format of particular set of POS tags, especially for a long NE.

2.3. The semi-joint labeling presentation

The semi-joint labeling (SJL) was first proposed in Chinese NER task [16]. In the following sections, we describe the details of semi-joint labeling.

2.3.1. Semi-joint labeling. In Chinese NER tasks, the Chinese NEs usually matches the word segmentation boundary. For example, the following sentence

俄羅斯$_{\text{/Location}}$ *總統* *普京*$_{\text{/Person}}$ *說*

The person name “普京” matches the Chinese word segmentation boundaries perfectly. Therefore, Wu et al. [16] injected the segmentation information to the original IOB2 tag sets to improve their Chinese NER performance. In English, however, no such segmentation information is available. Therefore we inject the chunking information as follows:

Our semi-joint labeling focus on the expected tag which was not be labeled as NE; the “O” tag. With the chunking information, we modify the “O” tag as shown in Table 1.

Table 1 Semi-joint labeling example

Chunking	Original Tag	Modified Tag
B-VP	O	B-O
B-NP	O	B-O
I-NP	O	I-O
I-NP	B-DNA	B-DNA
B-PP	I-DNA	I-DNA
B-NP	I-DNA	I-DNA
I-NP	O	B-O
I-NP	O	I-O
O	O	O

As you can see in Table 1, the “O” tag is modified according to the chunking information

while the NE part remains unchanged. The modification follows three rules. (1) If the original tag "O" matches chunking information with "B" tag, then the "O" will become "B-O". (2) If the original tag "O" matches chunking information with "I" tag, then check if it follows a "B-O" tag. If it does, then it'll be "I-O" or it'll be "B-O". (3) If the expected tag "O" matches chunking information with "O" tag, it means that the word has no chunk in the sentence. For example, it might be a symbol like "(" or ")", or a conjunction word "and". These words can't be classified into chunk types, so we remain the "O" unchanged.

The reasons why we proposed the new semi-joint labeling presentation is that IOB2 tag presentation only focuses on the tag of the NE part, while the non-NE part just labeled with "O". However, labeling the boundary of the non-NE part can also help deciding the boundary of NE. Instead of a bunches of "O", the "B-O" and "I-O" labels are more informative and useful in assisting NER because the "B-O", "I-O" tag provide extra boundary information, while the "O" tag only won't give these information. Hence, we proposed using the semi-joint labeling presentation to improve the accuracy of NER.

2.3.2. The chunking information. Our semi-joint labeling needs the chunking information. To acquire the chunking information, we constructed a biomedical full parser which is based on the Charniak parser [17]. This full parser is trained with the training data from GENIA Treebank [18], which contains 500 abstracts released from Tsujii laboratory on 2004 and 2005. It can generate parse trees for biomedical sentences automatically. After obtaining the parse tree, we use the similar algorithm described in [19] to derive the chunking information from our parse trees. Therefore, a parse tree illustrated in Figure 1 will be transformed into the following chunk:

$SV40_{/B\text{-}NP}$ $early_{/I\text{-}NP}$ $genes_{/I\text{-}NP}$ $induce_{/B\text{-}VP}$ $neoplastic_{/B\text{-}NP}$ $properties_{/I\text{-}NP}$ $in_{/B\text{-}PP}$ $serous_{/B\text{-}NP}$ $borderline_{/I\text{-}NP}$ $ovarian_{/I\text{-}NP}$ $tumor_{/I\text{-}NP}$ $cells_{/I\text{-}NP\text{-}/O}$

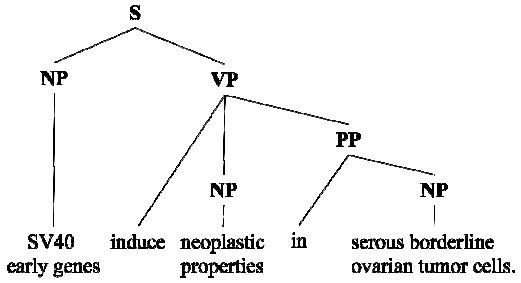

Figure 1. The parse tree

3. Results

3.1. Datasets

We use the dataset from JNLPBA 2004 shared task [10]. This dataset is converted from the GENIA corpus. The GENIA corpus consist of controlled search of MEDLINE using MeSH terms "human", "blood cells" and "transcription factors". There are also several biomedical NER systems using this GENIA corpus as their dataset. [20, 21] We make two experiments. For both of them, 10-fold cross validation (CV) was applied. The dataset from JNLPBA was divided into ten subsets. A single subset is keep as the test data, and the remaining subsets are used as training data for generating NER models. The CV process is then repeated 10 times, with each of the test sets being used for once.

3.2. Evaluation methods

The JNLPBA's evaluation script is used to evaluate the performance of our result. The script is a modified version of the script used in CoNLL-03 shared task [22]. The result is reported in the form of F-score, defined as F = (2PR) / (P + R), where P means the precision and R means the recall:

$$\text{Precision} = \frac{\text{the number of correctly found NE chunks}}{\text{the number of found NE chunks}}$$

$$\text{Recall} = \frac{\text{the number of correctly found NE chunks}}{\text{the number of true NE chunks}}$$

The evaluation script output three sets of F-score which are exact boundary matching, right boundary matching and left boundary matching [23]. Since the boundaries and categories of biomedical NE are usually ambiguous, matching boundaries is treated as another kind of evaluation for NER. In our experiment, we will take all three accuracy into consideration.

3.3. Experiment Design

We design two experiments to verify the effect using the semi-joint labeling presentation with the feature sets described in Section 2.2. The IOB2 tagging presentation is used as the baseline

3.4. Experiment Result

Table 2 The F-score based on different evaluation

	Complete Match	Left Boundary	Right Boundary
IOB2	62.37%	70.65%	65.73%
SJL	63.33%	71.53%	66.29%

Table 2 shows the average F-score of complete match, left and right boundary. The complete

match is the match for the whole NE, and the left and right boundary is for the match for one side of NE. As you can see, the proposed semi-joint labeling presentation outperform than the original IOB2 tag presentation.

4. Discussion

In our experiment, the proposed semi-joint labeling presentation improves the performance of NER. This presentation not only increases the F-score of the complete matching but also the left and right boundary matching. In the following section, we discuss some issues about our experiment.

4.1. The left and right boundary matching

Despite the fact that most people think only exact matching can be consider as a reliable evaluation, the left and right boundary matching provide additional useful information especially in the case where annotated corpora with the same adjectives annotated as part of some NE but not in others. In left boundary matching, if the left boundary matches exactly, the tagged NE is scored as a match. Under this rule, certain errors may be judged as correct. In these cases, the rightmost head words which represent the NE's category are skipped. This error may be acceptable in relation extraction and GO-ID assignment applications[24, 25] since the category matches, and the core term is successfully identified.

In right matching, if the right boundary matches exactly, the tagged NE is judged as correct. Applying this rule, errors due to missing or including preceding adjectives can be scored as correct. For example, in the sentence

Here, we report the identification of single-nucleotide polymorphisms (SNPs) in a region upstream of the minimal IL2 promoter.

The ambiguity occurs when some biologists label the “minimal IL2 promoter” as a NE, while others label the “IL2 promoter.”

As you can see in Table 2, the performance of the proposed semi-joint labeling is better than the IOB2 tag representation under complete, left, and right matching.

4.2. The Effects of Using Semi-joint Labeling

Here, we give an example to illustrate how semi-joint labeling significantly enhances the performance of our NER.

More interestingly, transfection experiments with CRE-CAT plasmide show that PGE2 activates the transcription of a CRE-containing promoter.

In this case, the “CRE-CAT plasmide” is recognized as non-NE in the NER model with IOB2 representation, but it is successfully recognized in our semi-joint labeling model. That is due to the words around the target NE (“with” and “show”) are labeled as “B-O” in our semi-joint labeling representation. It means that these two words are probably a new chunk. Hence, it can help our model to recognize the boundary of NEs.

5. Conclusion

We have presented a new presentation called semi-joint labeling presentation which can be used for improving the accuracy of NER in biomedical field. After applying this presentation, the complete matching accuracy increased from 62.37% to 63.33%, and the left and right boundary matching increase from 70.65% to 71.53% and from 65.73 to 66.29%.

According to the result generated by our experiment, we find out that modifying merely the part which is outside of NE instead of adding a new feature to every part of sentence still improves performance. This can be a new point of view in NER.

References

[1] K. B. Cohen and L. Hunter, "Natural Language Processing and Systems Biology," *Artificial Intelligence Methods and Tools for Systems Biology,* 2004.

[2] N. Chinchor, "MUC-7 named entity task definition," *Proceedings of the 7th Message Understanding Conference,* 1997.

[3] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: identifying protein names from biological papers," *Pacific Symposium on Biocomputing,* pp. 707-718, 1998.

[4] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, "Tuning support vector machines for biomedical named entity recognition," *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain,* pp. 1–8, 2002.

[5] S. Zhao, "Named Entity Recognition in Biomedical Texts using an HMM Model," *Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA),* 2004.

[6] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair,

"Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web," *Joint Workshop on Natural Language Processing in Biomedicine and Its Applications at Coling 2004,* 2004.

[7] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets," in *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)* Geneva, Switzerland, 2004.

[8] R. T.-H. Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu, "NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition," *BMC Bioinformatics,* vol. 7 Suppl 5, p. S11, 2006.

[9] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," *North American Chapter Of The Association For Computational Linguistics,* pp. 1-8, 2001.

[10] K. Jin-Dong, O. Tomoko, Y. T. Yoshimasa Tsuruoka, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in *Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04)*, 2004, pp. 70-75.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

[12] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval,* pp. 235-242, 2003.

[13] F. Sha and F. Pereira, "Shallow Parsing with Conditional Random Fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. vol. 1 Edmonton, Canada Association for Computational Linguistics, 2003, pp. 134-141.

[14] H.-J. Dai, H.-C. Hung, R. T.-H. Tsai, and W.-L. Hsu, "IASL Systems in the Gene Mention Tagging Task and Protein Interaction Article Sub-task," in *Proceedings of Second BioCreAtIvE Challenge Evaluation Workshop*, Madrid, Spain, 2007, pp. 69-76.

[15] Y. Tsuruoka, "Bidirectional inference with the easiest-first strategy for tagging sequence data," *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing,* pp. 467-474, 2005.

[16] R. T.-H. T. Chia-Wei Wu, and Wen-Lian Hsu, "Semi-joint Labeling for Chinese Named Entity Recognition," in *AIRS 2008* Berlin Heidelberg, 2008, pp. 107 - 116.

[17] E. Charniak, "A Maximum-Entropy-Inspired Parser," in *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics* Seattle, Washington: Morgan Kaufmann Publishers Inc., 2000, pp. 132-139.

[18] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus--a semantically annotated corpus for bio-textmining," *Bioinformatics,* vol. 19, pp. 180-182, 2003.

[19] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," *Proceedings of the Third ACL Workshop on Very Large Corpora,* pp. 82-94, 1995.

[20] K. J. Lee, Y. S. Hwang, and H. C. Rim, "Two-phase biomedical NE recognition based on SVMs," *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13,* pp. 33-40, 2003.

[21] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: a machine learning approach," *Bioinformatics,* vol. 20, pp. 1178-90, May 1 2004.

[22] E. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition," *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4,* pp. 142-147, 2003.

[23] R. T. H. Tsai, S. H. Wu, W. C. Chou, Y. C.

Lin, D. He, J. Hsiang, T. Y. Sung, and W. L. Hsu, "Various criteria in the evaluation of biomedical named entity recognition," *BMC Bioinformatics,* vol. 7, p. 14, 2006.

[24] J. H. Chiang and H. C. Yu, "MeKE: discovering the functions of gene products from biomedical literature via sentence alignment." vol. 19: Oxford Univ Press, 2003, pp. 1417-1422.

[25] C. Blaschke, E. A. Leon, M. Krallinger, and A. Valencia, "Evaluation of BioCreAtIvE assessment of task 2," *BMC Bioinformatics,* 2005.