A Method to Generate a Machine-Labeled Data for Biomedical Named Entity Recognition with Various Sub-Domains

Juae Kim[1], Sunjae Kwon[1], Youngjoong Ko[2], and Jungyun Seo[1]
Computer Science, Sogang University, Sinsu-dong 1, Mapo-gu, Seoul, Korea[1]
Computer Engineering, Dong-A University, 840 Hadan 2-dong, Saha-gu, Busan, Korea[2]
{juaekim, soon91jae , seojy}@sogang.ac.kr[1],
yungjoong.ko@gmail.com[2]

Abstract

Biomedical Named Entity (NE) recognition is a core technique for various works in the biomedical domain. In previous studies, using machine learning algorithm shows better performance than dictionary-based and rule based approaches because there are too many terminological variations of biomedical NEs and new biomedical NEs are constantly generated. To achieve the high performance with a machine-learning algorithm, good-quality corpora are required. However, it is difficult to obtain the good-quality corpora because annotating a biomedical corpus for machine-learning is extremely time-consuming and costly. In addition, most previous corpora are insufficient for high-level tasks because they cannot cover various domains. Therefore, we propose a method for generating a large amount of machine-labeled data that covers various domains. To generate a large amount of machine-labeled data, firstly we generate an initial machine-labeled data by using a chunker and MetaMap. The chunker is developed to extract only biomedical NEs with manually annotated data. MetaMap is used to annotate the category of biomedical NE. Then we apply the self-training approach to bootstrap the performance of initial machine-labeled data. In our experiments, the biomedical NE recognition system that is trained with our proposed machine-labeled data achieves much high performance. As a result, our system outperforms biomedical NE recognition system that using MetaMap only with 26.03%p improvements on F1-score.

1 Introduction

As biomedical research has been actively studied, the attention of bioinformatics with natural language processing is rapidly increasing. According to generate an amount of data in the biomedical domain, to extract and retrieval the high-quality information is increasingly important (Zeng et al., 2015). Then, extracting the Biomedical Named Entity (biomedical NE) is also important to comprehend the biomedical text. There are two steps to extract biomedical NEs. The first step is an identification of biomedical entities from text. Second, biomedical entities that identified are classified into some several categories such as protein, drug, cell-line, and disease. These categories provide useful information to high-level applications. To extract the biomedical NEs are considered challenging task because there are too many terminological variations of biomedical NEs and new biomedical NEs are constantly generated with the course of time. That is why studies using machine learning show higher performance that dictionary or rule based approaches. There are many efforts that extract a high-quality biomedical NEs. (Robert et al., 2015) suggests a chemical named entity recognizer implemented by combining two independent machine learning models. (Li et al., 2015), (Li et al., 2016) apply the latest technology which is deep learning. The named entity recognizer of (Li et al., 2015), (Li et al., 2016) are based on Recurrent Neural Network and LSTM. However, annotating a biomedical corpus for machine-learning is extremely time-consuming and costly because of the requirement of medical experts. In addition, most previous corpora are insufficient for high-level application likes question answering (QA) system that requires various biomedical information. Because biomedical NE cat-

Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017), pages 47–51,
Taipei, Taiwan, November 27, 2017.

Category (abbreviation)	Category (UMLS semantic groups)
ACTI	Activities & Behaviors
ANAT	Anatomy
CHEM	Chemicals & Drugs
CONC	Concepts & Ideas
DEVI	Devices
DISO	Disorders
GENE	Genes & Molecular Sequences
GEOG	Geographic Areas
LIVB	Living Beings
OBJC	Objects
OCCU	Occupations
ORGA	Organizations
PHEN	Phenomena
PHYS	Physiology
PROC	Procedures

Table 1: UNLS semantic groups

egories were limited to specific sub-domains of biomedicine in each corpus. For example, the biomedical NE recognition system that trained with GENIA corpus (Kim et al., 2003), only cover the gene and protein subdomain.

Therefore, we propose a method for creating the automatically labeled corpus that covers various domains. The biomedical NE recognition system that is trained with our proposed machine-labeled corpus can extract biomedical NEs in various domains.

We utilize the open source biomedical NE recognition tool, MetaMap (Aronson, 2001) to provide various biomedical information as categories. MetaMap extracts biomedical NEs from raw texts and matches them into semantic types of UMLS. Unified Medical Language System (UMLS) (Lindberg et al., 1993) is a thesaurus that provides the biomedical NEs and their semantic categories as an annotation. UMLS includes an amount of concept categories and a semantic network likes UMLS semantic type. Similar UMLS semantic types are grouped into UMLS semantic group. In this paper, we regard the annotation of biomedical NEs as UMLS semantic group. Table 1 shows UMLS semantic groups. UMLS semantic groups cover not only gene and protein but various sub-domains such as anatomy, procedure, and medical device. Biomedical NEs and their semantic types from MetaMap are usefully used in biomedical QA systems and biomedical topic modeling systems that require similarity of biomedical terms. MetaMap is a useful tool to analyze the biomedical text, however, there are several limitations (Zhang and Elhadad., 2013) because UMLS covers the huge knowledge and there are a lot of newly generated biomedical NEs in a real world. First, MetaMap extracts not only biomedical NEs but also common entities or even verbs that are not biomedical NEs clearly. Second, MetaMap does not resolve the ambiguity of UMLS semantic group types without context. That is, one entity can have several annotations. This limitation is 'ambiguity problem' in this paper. Finally, if some biomedical NE is not recorded in UMLS thesaurus, MetaMap cannot assign UMLS semantic group. We named this limitation 'out–of–vocabulary problem'. Thus, we present how to develop an effective biomedical NE recognition system by applying the advantages and overcoming the limitations of MetaMap.

We had been devised NE recognition system find the similarity of biomedical NEs in a query and biomedical NEs in candidate answer in biomedical QA system based on information retrieval (Lee et al., 2016). This QA system is entered for 2016 BioASQ challenge (Tsatsaronis et al., 2012).

2 Proposed Method

To overcome limitations of MetaMap, we propose a method to automatically construct a biomedical NE recognition data using a small amount of labeled data, a large amount of unlabeled data and MetaMap. Firstly, we develop a chunker learned by a small amount of biomedical training data annotated by the people. This chunker is used to generate the labeled training data for the self-training approach. The chunking results of unlabeled biomedical corpus become the inputs of MetaMap. Then the outputs of MetaMap, UMLS

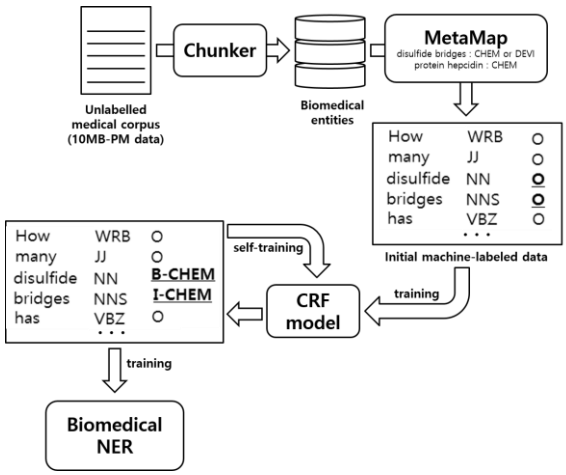

Figure 1: Overview of proposed method

semantic types, become match the UMLS semantic groups are used to make an initial machine-labeled data. We develop a biomedical NE recognition system with initial machine-labeled data and conditional random fields (CRF) classifier. The initial machine-labeled data has a form of semantic role labelling with an IOB2 format. This classifier is an initial model for self-training. Our proposed self-training process is described in Section 2.2. During self-training, this process is repeated, and then overcome the ambiguity problem and out–of–vocabulary problem.

Figure 1 shows the example of overcoming the ambiguity problem. In initial machine-labeled data, 'disulfide bridges' denoted by underline has a 'O' tag, but after the self-training process the entity, disulfide bridges, has a 'CHEM' tag.

2.1 Initial Machine-Labeled Data

In this section, we present the method for generating the initial machine-labeled data for self-training. We propose a semi-supervised approach using MetaMap, a small amount of labeled training data and a large amount of unlabeled data to generate a large amount of machine-labeled training data. The unlabeled data is randomly selected titles and abstracts from the PubMed biomedical articles. In this paper, we call the unlabeled data as a 10MB-PM data.

In order to generate training data, we first develop a chunker to overcome the problem that MetaMap extract several entities even if they are not biomedical NEs. The chunker is developed with a small amount of labeled data. The NE candidates of the 10MB-PM data are chunked by the chunker and then only the NE chunks are exploited as the input of MetaMap to label biomedical NEs.

We now explain how to create a large amount of initial machine-labeled data with the analysis results of MetaMap. NE chunks from the chunker are used as the input of MetaMap and then MetaMap analyses their semantic types. We mapped 133 UMLS semantic types to 15 UMLS semantic groups depanding on semilarity. The semantic groups are regarded as annotation of biomedical NEs. If MetaMap outputs only one semantic group type for a biomedical NE chunk, we assume that this semantic group type is a correct annotation of the NE chunk. In this case, each word of the NE chunk is labeled by B or I tag with UMLS group as annotation. For example, there is a NE chunk 'protein hepcidin' from 10MB-PM data. The output of MetaMap as semantic group type of 'protein hepcidin' is 'CHEM'. To generate automatically labeled data, we tag the 'B-CHEM' to 'protein' and 'I-CHEM' to 'hepcidin'. However, in some cases, MetaMap outputs several semantic groups or no group type for an input NE chunk. In these cases, the NE chunks are not considered as a biomedical NEs and their words are initially labeled by O tags. We try to recover them in order to generate a robust machine-labeled data through the self-training.

2.2 Self-training

The self-training approach is one of the semi-supervised learning. In Self-training, a classifier is trained with the manually annotated data. Then the unlabeled data is input the classifier. All or part of a result of the classifier is used as a training data.

Through a self-training procedure, we can overcome the ambiguity and out-of-vocabulary problems in machine-labeled data. The proposed self-training procedure performs the following steps in an iterative procedure. First, CRFs is trained using initial machine-labeled data, and then the sequential labelling results of the trained CRFs are compared the ones of the initial machine-labeled data as an answer. If the entity with O tag in the initial machine-labeled data are changed B or I tag by the trained CRFs model, the tags of this entity is replaced with the new B or I tag in new training data. This new training data is used as the input training data for the next self-training iteration. This iterative procedure is con-

ducted until the performance of the classifier is converged. In each repeated step, new labeled biomedical NEs are added in the training data.

In case of biomedical NE chunk that has the ambiguity problem or out-of-vocabulary problem, we cannot make a decision to the categories of that NE chunks. Thus we annotate the O tag to that NE chunk. In other words, entities that have ambiguity UMLS semantic group or no UMLS semantic group cannot exist in automatically labeled data. However, their categories are recovered appropriate tag by proposed self-training. It is the main reason that the proposed NE recognition system showed high recall scores in our experiments.

For example, the word, 'drisapersen,' is biomedical NE of the drugs but this entity is not recorded in the UMLS thesaurus. After executing our self-training procedure, the word 'drisapersen' and its semantic group type, 'CHEM,' are analyzed as a correct biomedical NE; 'CHEM' is a chemical category as a semantic group type that includes drugs, protein, steroid, vitamin, and others.

3 Experiments

3.1 Experimental Settings

We constructed a manually annotated dataset that consists of 1,249 biomedical question/answer pairs from BioASQ 2015 and 2016. A half of the annotated data was used as a training data and the other data was used as a test data. The test dataset is composed of 624 question/answer pairs and 1,492 biomedical NEs. In test data, there are 1,492 biomedical NEs. The 10MB-PM data is organized by 7,629 PubMed articles that are arbitrary selected. To evaluate the quality of machine-labeled data, we evaluated biomedical NE recognition systems that are trained with the machine-labeled data of each self-training iteration with precision, recall, and f1-score.

3.2 Experimental Result and Evaluation

'MetaMap' in the first row is a model that is used MetaMap as biomedical NE recognition system. We input the raw text into MetaMap, and then evaluated the performance. An 'Initial model' is the model that is trained with initial machine-labeled data. The initial machine-labeled data is generated by applying chunker and MetaMap. A 'Second-iteration model' in the third row means self-training model with initial model. Row 4-6 are result of models that each iteration of self-training.

In Table 2, the initial model shows much-improved performance. The performance of MetaMap is much lower than the initial model. In particular, the precision score of MetaMap is much worse than recall score because it outputs common nouns and verbs as biomedical NEs. On the other hand, the chunker used in initial model is trained to extract only biomedical NEs with a small amount of manually annotated data. That is why the precision score of initial model is more improved than MetaMap.

To recover NE entities with ambiguity and out-of-vocabulary, the self-training procedure is developed in our biomedical NE recognition system. The performance changes of self-training according to the number of iteration times are shown in row 2 to row 6 in Table 2. The best performance was obtained in the third iteration. The performance of third ST model and fourth ST model is a lower than second ST model. That is conducting the too much self-training makes some decrease of performance because noises can be generated by recovering the wrong biomedical NEs. With the self-training method, we can increase the F1-score from 68.04% to 69.91%.

Model	Precision	Recall	F1-score
MetaMap	34.98%	58.8%	43.88%
Initial model	83.01%	57.64%	68.04%
Second iteration model	79.93%	60.32%	68.75%
Third iteration model	79.91%	62.13%	69.91%
Fourth iteration model	77.85%	59.58%	67.50%
Fifth iteration model	71.88%	49.87%	58.88%

Table 2: The performance changes according to our proposed method

4 Conclusion and Future Work

In this paper, we proposed the method for generating machine-labeled biomedical NE recognition data with the self-training method. Through various experiments, we verified the performances of the biomedical NE recognition system that trained with our proposed machine-labeled data. The final system outperformed MetaMap with 26.03%. In addition, the proposed method has more strong points. To generate a large amount of data, we only used a small amount of training data. Therefore

the cost to generate a data for the biomedical NE recognition systems can be reduced. Since MetaMap as an open toolkit is used, developers can build up the biomedical NE recognition systems without expert's help in biomedical domains.

As a future work, we plan to apply deep neural network techniques to construct the biomedical NE recognition systems.

Acknowledgement

This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW)(2015-0-00910) supervised by the IITP(Institute for Information & communications Technology Promotion)

References

Alan R. Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. "GENIA corpus—a semantically annotated corpus for bio-textmining." *Bioinformatics*, 19(1), 2003, i180-i182.

Hyeon-gu Lee, Minkyoung Kim, Harksoo Kim, Juae Kim, Sunjae Kwon, Jungyun Seo, Jungkyu Cho, and Yi-reun Kim. "KSAnswer: Question-answering System of Kangwon National University and Sogang University in the 2016 BioASQ Challenge." *ACL 2016*, 2016, pp.45-49.

Lishuang Li, Liuke Jin, Zhenchao Jiang, Dingxin Song, and Degen Huang. "Biomedical named entity recognition based on extended recurrent neural networks." *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015.

Lishuang Li, Liuke Jin, Yuxin Jiang, and Degen Huang. "Recognizing Biomedical Named Entities Based on the Sentence Vector/Twin Word Embeddings Conditioned Bidirectional LSTM." *China National Conference on Chinese Computational Linguistics*. Springer International Publishing, 2016.

Donald A. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. "The Unified Medical Language System." *IMIA Yearbook*, 1993, pp.281-291.

Leaman Robert, Chih-Hsuan Wei, and Zhiyong Lu. "tmChem: a high performance approach for chemical named entity recognition and normalization." *Journal of cheminformatics*, 7(1), 2015.

George Tsatsaronis Michael Schroeder, Georgios Paliouras Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari Thierry Artieres, Michael R. Alvers Matthias Zschunke, and Axel-Cyrille Ngonga Ngomo. "BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering." *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*, 2012.

Zhiqiang Zeng, Hua Shi, Yun Wu, and Zhiling Hong. "Survey of natural language processing techniques in bioinformatics." *Comp Math Methods Med 2015*, article 674296.

Shaodian Zhang, and Noémie Elhadad. "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts." *Journal of biomedical informatics*, 46(6), 2013, pp.1088-1098.