Towards a Unified Named Entity Recognition System
Disease Mention Identification

Tsendsuren Munkhdalai, Meijing Li, Khuyagbaatar Batsuren and Keun Ho Ryu
Database/Bioinformatics Laboratory, School of Electrical & Computer Engineering, Chungbuk National University, Cheongju, South Korea

Keywords: Feature Learning, Semi-Supervised Learning, Named Entity Recognition, Conditional Random Fields.

Abstract: Named Entity Recognition (NER) is an essential prerequisite task before effective text mining can begin for biomedical text data. Exploiting unlabeled text data to leverage system performance has been an active and challenging research topic in text mining due to the recent growth in the amount of biomedical literature. In this study, we take a step towards a unified NER system in biomedical, chemical and medical domain. We evaluate word representation features automatically learnt by a large unlabeled corpus for disease NER. The word representation features include brown cluster labels and Word Vector Classes (WVC) built by applying k-means clustering to continuous valued word vectors of Neural Language Model (NLM). The experimental evaluation using Arizona Disease Corpus (AZDC) showed that these word representation features boost system performance significantly as a manually tuned domain dictionary does. BANNER-CHEMDNER, a chemical and biomedical NER system has been extended with a disease mention recognition model that achieves a 77.84% F-measure on AZDC when evaluating with 10-fold cross validation method. BANNER-CHEMDNER is freely available at: https://bitbucket.org/tsendeemts/banner-chemdner.

1 INTRODUCTION

One essential task in developing an information extraction system is the Named Entity Recognition (NER) process, which basically defines the boundaries between typical words and biomedical terminology in a particular text, and assigns the terminology to specific categories based on domain knowledge.

Gene and protein mention recognition in biomedical text has been a main focus of the bio-text mining community and many systems have been developed (Leaman 2008, Munkhdalai 2013). In contrast, recognition of disease has received much less attention. Proposed solutions include rule-based, dictionary-based, and Machine Learning (ML) approaches.

In the dictionary-based approach, a prepared terminology list is matched through a given text to retrieve chunks containing the location of the terminology words (Karopka 2006, Jimeno 2008, Gurulingappa 2010). However, medical and chemical text can contain new terminology that has yet to be included in the dictionary.

The rule-based approach defines particular rules by observing the general features of the entities in a text. In order to identify any named entity in text data, a rule-generation process has to process a huge amount of text to collect accurate rules. In addition, the rules are usually collected by domain experts, requiring a lot of effort.

Since the Machine Learning (ML) approach was adopted, significant progress in disease NER has been achieved (Leaman 2009, Chowdhury 2010, Neveol 2009). Robert et al. introduced the AZDC corpus and adapted Conditional Random Fields (CRF)-based gene mention recognition system, BANNER for disease NER. They also combined the CRF model with a dictionary-based method and showed a significant improvement. Chowdhury et al. studied different combination of feature sets, including dictionary lookups and tags extracted by a syntactic dependency parser. Those special feature combinations in conjunction with carefully designed postprocessing rules were observed to boost the performance at a higher rate. However, incorporation of the domain dependent dictionary into a ML system makes it non-trivial to adapt such a system for another domain. This leads to an individual system that is only applicable to a particular NER task (such system might only address to gene mention recognition problem) rather a unified system that

Munkhdalai T., Li M., Batsuren K. and Ryu K..
Towards a Unified Named Entity Recognition System - Disease Mention Identification.
DOI: 10.5220/0005287802510255
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 251-255
ISBN: 978-989-758-070-3

could be applied to multiple NER tasks, such as gene, chemical and disease mention recognition.

Recently, Semi-Supervised Learning (SSL) techniques have been applied to NER. SSL a ML approach that typically uses a large amount of unlabeled and a small amount of labeled data to build a more accurate classification model than that which would be built using only labeled data. SSL has received significant attention for two reasons. First, preparing a large amount of data for training requires a lot of time and effort. Second, since SSL exploits unlabeled data, the accuracy of classifiers is generally improved. There have been two different directions of SSL methods, semi-supervised model induction approaches which are the traditional methods and incorporate the domain knowledge from unlabeled data into the classification model during the training phase (Munkhdalai 2012, Munkhdalai 2010), and supervised model induction with unsupervised, possibly semi-supervised feature learning. The approaches in the second research direction induce a better feature representation by learning over a large unlabeled corpus. Recently, the studies that apply the word representation features induced on the large text corpus have reported improvement over baseline systems in many natural language processing (NLProc) tasks (Turian 2010, Huang 2012, Socher 2011).

In this study, we take a step towards a unified NER system in biomedical, chemical and medical domain. We evaluate generally applicable word representation features automatically learnt by a large unlabeled corpus for disease NER. The word representation features include brown cluster labels and Word Vector Classes (WVC) built by applying k-means clustering to continuous valued word vectors of Neural Language Model (NLM). The experimental evaluation using Arizona Disease Corpus (AZDC) showed that these word representation features boost system performance significantly as a manually tuned domain dictionary does. BANNER-CHEMDNER (Munkhdalai 2013), a chemical and biomedical NER system has been extended with a disease mention recognition model that achieves a 77.84% F-measure on AZDC when evaluating with 10-fold cross validation method.

The rest of this paper is organized as follows. Section 2 introduces the proposed methodology and the stages in the disease NER pipeline. Section 3 reports the performance evaluation of the system based on combination of word representation features, and a comparison against the existing systems. Finally, we summarize the main conclusions achieved and present our future work direction.

2 DISEASE NAMED ENTITY RECOGNITION

This section introduces a detail of the proposed disease NER pipeline. First, we per-form preprocessing on MEDLINE and PMC document collection and then extract two different feature sets, a base feature set and a word representation feature set, in the feature processing phase. The unlabeled set of the collection is fed to unsupervised learning of the feature processing phase to build word classes. Finally, we apply the CRF sequence-labeling method to the extracted feature vectors to train the NER model. These steps will be described in subsequent sections.

2.1 Preprocessing

First, the text data is cleansed by removing non-informative characters and replacing special characters with corresponding spellings. The text is then tokenized with BANNER simple tokenizer. The BANNER tokenizer breaks tokens into either a contiguous block of letters and/or digits or a single punctuation mark. Finally, the lemma and the part-of-speech (POS) information were obtained for a further usage in the feature extraction phase. In BANNER-CHEMDNER, BioLemmatizer (Liu 2012) was used for lemma extraction, which resulted in a significant improvement in overall system performance for biomedical and chemical NER.

In addition to these preprocessing steps, special care is taken to parse the PMC XML documents to get the full text for the unlabeled data collection.

2.2 Feature Processing

We extract features from the preprocessed text to represent each token as a feature vector, and then an ML algorithm is employed to build a model for NER.

The proposed method includes extraction of the baseline and the word representation feature sets. The word representation features can be extracted by learning on a large amount of text and may be capable of introducing domain background to the NER model.

The entire feature set for a token is expanded to include features for the surroundings with a two-length sliding window. The word, the word n-gram, the character n-gram, and the traditional orthographic information are extracted as the baseline feature set. The regular expressions that reveal orthographic information are matched to the tokens to give orthographic information.

For word representation features, we train Brown clustering models (Brown 1992) and Word Vector (WV) models (Collobert 2008) on a large PubMed and PMC document collection. Brown clustering is a hierarchical word clustering method, grouping words in an input corpus to maximize the mutual information of bigrams. The VW model is induced via a neural language model and consists of n-dimensional continuous valued vectors, each of which represents a word in the training corpus. Further, the word vectors are clustered using a K-means algorithm to drive a Word Vector Class (WVC) model. Since Brown clustering is a bigram model, this model may not be able to carry wide context information of a word, whereas the WVC model is an n-gram model (usually n=5) and learns broad context information from the domain corpus. We drive the cluster label prefixes with 4, 6, 10 and 20 lengths in the Brown model, 50 and 100 dimensions of the WVs, and the WVCs as word representation features.

2.3 Supervised Learning

CRF - a probabilistic undirected graphical model has been used successfully in a large number of studies on NER, because it takes advantage of sequence labelling by treating each sentence as a sequence of tokens. We apply a second-order CRF model, where the current label is conditioned on the previous two using a Begin, Inside, Outside (BIO) tagging format of the tokens. In the BIO tagging format, each token is classified either at the beginning, inside or outside of a named entity, and a postprocessing task forms the named entity mentions by merging the tagged tokens.

We use a Machine Learning for Language Toolkit (MALLET) library for training the CRF model, because the BANNER system provides a convenient interface to work with it. The BANNER system also includes two types of general post-processing that could be useful for any NER tasks in bio-text data. The first type is based on the symmetry of parenthesis, brackets or double quotation marks. Since these punctuation marks are always paired, BANNER drops any named entity mention containing mismatched parentheses, brackets or double quotation marks. The second type of post-processing is dedicated to resolving abbreviations of named entities.

3 RESULTS

First, we evaluated combination of word representation features using 10-fold cross validation. We then compared the result against existing systems

3.1 Dataset

We evaluated the system using AZDC corpus (Leaman 2009) for disease mention identification. The dataset consists of 793 annotated abstracts containing 2,873 sentences, 3,093 unique disease mentions.

For the unlabeled data, we collected around 1.4 million PubMed abstracts and full text articles from the whole PMC database available at the time (over 2 million documents). After preprocessing, we derived two different text corpora: a PubMed abstract corpus consisting of a vocabulary of 1,136,085 entries for induction of Brown clustering models, and a merged corpus of both resources with a vocabulary of 4,359,932 entries for training WV models. Given the limited resources and time, we were able to induce the Brown clustering models only with the PubMed abstract corpus.

3.2 Performance Evaluation

We followed an experimental setting similar to the one in Robert et al. in order to compare our results with that of the BANNER system. We performed 10-fold cross validation on AZDC in such a way that all sentences of the same abstract are included in the same fold. The results of the ten folds are averaged to obtain the final outcome.

Table 1 shows the performance comparison of the different runs with varied feature settings. We started conducting a run with a basic feature setting, and gradually increased the complexity of the feature space for further runs. A Brown model with a larger number of clusters tended to obtain a higher F-measure. Unlike Brown clustering, a large or a lower number of WVCs degraded the performance. We found the WVC model with 300 different classes the best performing one on this task. Further, the combination of the different WVC models significantly improved the F-measure. We achieved the best performance, a 77.84% F-measure with the model based on the baseline feature set, the 1000-Brown clustering, and 300, 500 and 1000 WVCs (the baseline + Brown 1000 + WVC 300 + WVC 500 + WVC 1000 setup).

Table 1: Disease NER evaluation results of different runs with varied features. Feature groups are separated by (+) and followed by the corresponding parameters.

Features	Precision (%)	Recall (%)	F-score (%)
Baseline + Brown 300	78.96	72.41	75.53
Baseline + Brown 1000	78.93	73.55	76.1
Baseline + Brown 1000 + WVC 1000	79.6	73.59	76.45
Baseline + Brown 1000 + WVC 300	79.63	75	77.21
Baseline + Brown 1000 + WVC 500	78.88	74.39	76.54
Baseline + Brown 1000 + WVC 500 + WVC 300	80.06	74.66	77.25
Baseline + Brown 1000 + WVC 500 + WVC 1000	79.29	74.04	76.54
Baseline + Brown 1000 + WVC 500 + WVC 300 + WVC 1000	80.44	75.45	77.84

Table 2: Comparison of BANNER and our system results.

Systems	Precision (%)	Recall (%)	F-score (%)
BANNER	78.5	69.9	74
BANNER (with dictionary)	80.9	75.1	77.9
BANNER-CHEMDNER	80.44	75.45	77.84

3.3 Performance Comparison

Table 2 reports the comparison of BANNER and our system results. Our system outperforms the basic BANNER setup by a 3.84% F-measure. BANNER combined with dictionary matching performs slightly better than our system. Our system achieves a higher recall, since it is based on ML. In contrast, the precision of BANNER with dictionary is better. In fact, this is the main advantage of dictionary-based methods.

In our system, we do not rely on any lexicon nor any dictionary other than the free text in the domain in order to keep the system applicable to other NER tasks in bio-text data, even though the usage of such resources is reported to considerably boost system performance.

4 CONCLUSIONS

We took a step towards a unified named entity recognition system in biomedical, chemical and medical domain. We evaluated word representation features automatically learnt by a large unlabeled corpus for disease named entity recognition system. The word representation features include brown cluster labels and word vector classes built by applying k-means clustering to continuous valued word vectors of neural language model.

The experimental evaluation using Arizona disease corpus showed that these word representation features boost system performance significantly as a manually tuned domain dictionary does. BANNER-CHEMDNER, a chemical and biomedical named entity recognition system has been extended with a disease mention recognition model that achieves a 77.84% F-measure on Arizona disease corpus when evaluating with 10-fold cross validation method.

ACKNOWLEDGEMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No-2013R1A2A2A01068923) and by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2008-0062611).

REFERENCES

Leaman, R., Gonzalez, G., 2008. Banner: An Executable Survey of Advances in Biomedical Named Entity Recognition. In *Pacific Symposium on Biocomputing.*

Munkhdalai, T., Li, M., Batsuren, K., Ryu, K. H., 2013. Banner-Chemdner: Incorporating Domain Knowledge

in Chemical and Drug Named Entity Recognition. In *Fourth BioCreative.*

Karopka, T., Fluck, J., Mevissen, H., Glass, A., 2006. The autoimmune Disease Database: a dynamically compiled literature-derived database. *BMC Bioinformatics.*

Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., Rebholz-Schuhmann, D., 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics.*

Gurulingappa, H., Klinger, R., Hofmann-Apitius, M., Fluck, J., 2010. An Empirical Evaluation of Resources for the Identification of Disease and Adverse Effects in Biomedical Literature. In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining.*

Leaman, R., Miller, C., 2009. Enabling Recognition of Disease in Biomedical Text with Machine Learning: Corpus and Benchmark. In *Symposium on Languages in Biology and Medicine.*

Chowdhury, M. F. M., Lavelli, A., 2010. Disease Mention Recognition with Specific Features. In *Biomedical Natural Language Processing.*

Neveol, A., Kim, W., Wlbur, W. J., Lu, Z., 2009. Exploring Two Biomedical Text Genres for Disease Recognition. In *Biomedical Natural Language Processing.*

Munkhdalai, T., Li, M., Kim, T., Namsrai, O., Seon-phil, J., Jungpil, S., Ryu, K. H., 2012. Bio Named Entity Recognition based on Co-training Algorithm. In *AINA 2012.*

Munkhdalai, T., Li, M., Unil, Y., Namsrai, O., Ryu, K. H., 2012. An Active Co-Training Algorithm for Biomedical Named-Entity Recognition. *KIPS.*

Turian, J., Ratinov, L., Bengio, Y., 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL.*

Huang, E. H., Socher, R., Manning, C. D., Ng, A. Y., 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *ACL.*

Socher, R., Lin, C. C, Ng, A. Y., Manning, C. D., 2011. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *ICML.*

Liu, H., Christiansen, T., Baumgartner, W. A., Verspoor, K., 2012. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *J. Bio. Sem.*

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., Lai, J. C., 1992. Class-Based n-gram Models of Natural Language. In *ACL.*

Collobert, R., Weston, J., 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *ICML.*