Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 6 (2012) 379 – 386

Procedia Technology

2nd International Conference on Communication, Computing & Security [ICCCS-2012]

A Novel Technique for Name Identification from Homeopathy Diagnosis Discussion Forum

Mukta Majumder*, Utsav Barman, Rahul Prasad, Kumar Saurabh, Sujan Kumar Saha

Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi – 835215, India

Abstract

Named entities are the most informative element of a textual document and identification of the names is very much important for extracting further information from text. We have developed a conditional random field based system to identify the named entities from homeopathic diagnosis discussion forum text. We have manually annotated a training corpus for the task. As manual creation of a sufficiently large annotated corpus is costly and time consuming, we use an active learning based semi-supervised framework to increase the efficiency of the system with the help of un-annotated data. Our system achieves the highest f-value of 84.35.

Keywords: Named Entity Recognition; Homeopathic diagnosis; Automatic diagnosis system; Conditional Random Field; Semi-supervised Learning; Active Learning;

1. Introduction

Named Entity Recognition (NER) involves locating and classifying the names in text. As the names are the pivotal elements in text, NER is an important task having applications in information extraction, question answering, and machine translation and in most other Natural Language Processing (NLP) applications.

As an affordable diagnosis, homeopathy treatment is always very popular to common people. With the huge popularity of internet, online discussion forum in homeopathic domain also get lots of attention from those people. Now a day there are many homeopathic discussion forums available online where user can discuss about various diseases, symptoms and ask for their diagnosis. Based on these discussion posts and queries from user, experienced doctors and experts post their valuable view on a particular disease and suggest how to take appropriate medicine timely. Automatic diagnostic system or other text mining tools can be developed if these

* Corresponding author. Tel.: +91-7677632397; fax: +0-000-000-0000 .
E-mail address: mukta_jgec_it_4@yahoo.co.in

doi:10.1016/j.protcy.2012.10.045

discussions available in the web can be used effectively. To use these texts in text mining or Information Extraction (IE) systems, first we need to identify the named entities. In these corpora the primary named entities are drugs, diseases and symptoms. As these discussion texts are written by normal web users, these texts often contain a high amount of noises. Due to the noises, standard NLP tools often fail to perform properly on these corpora; and development of NLP systems on this type of corpora requires some special techniques. Development of NER system in homeopathic diagnosis discussion forum texts is more difficult compared to the NER task in general domain. The complicated and ambiguous naming convention of these medicine and disease names are a major difficulty of this task. In homeopathic domain Named Entities (NEs) are often long and include common words, conjunctions, prepositions and numeric value in between two words or at end. This makes the task of classification and boundary identification quite difficult. Spelling variation is another ambiguity to identify these types of NEs. A particular medicine name can be written differently by different users. For example, 'Nux Vomica' is written as 'NuxVom', 'Nux', 'NVom' and 'NV'. Further, the use of capitalization, parenthesis, hyphen and abbreviation in forum text does not follow a standard convention.

There are two main approaches to NER, namely rule based and Machine Learning (ML) based (R. Grishman et al., 1995-96; K. Fukuda et al., 1998; S.K. Saha et al., 2009). Rule based systems are difficult to develop for complex named entities and they require domain experts. Such systems are not portable to handle other NE types and domains. That is why ML based NER is a better choice for more complex domains. The success of a machine learning algorithm is crucially dependent on the features set used to train it. A supervised learning algorithm uses an annotated training corpus. The training set derived from an annotated corpus represents the NEs in terms of the feature values. There is another technique called hybrid systems which is a combination of these two for identifying NEs (N.V Sobhana et al., 2010). In this paper we present a NER system in the homeopathic discussion forum domain, where we have considered two name categories, drug/medicine names and disease names. We have used conditional random field (CRF) for the development of the baseline system. The machine learning approaches for NER system development require sufficient annotated data to train the system. As we could not find any openly available annotated data in this domain, first we have prepared a training corpus (~100K words) by manual annotation. Then we identify a set of features and train the system. The system achieves an f-score of 83.29. As the training data is not sufficient, the system suffers from poor recall (77.80%). We observed that a number of names are not recognized by the system. Then we have planned to adopt semi-supervised learning (SSL) approach to improve the accuracy. We have studied a semi-supervised technique approach namely active learning where we have used a large raw corpus to leverage the performance of the supervised classifier. There we have observed that use of SSL improves the accuracy upto 84.35. The rest of the paper is organized as follows. Section 2 discusses the related previous work. Section 3 presents the supervised learning technique using CRF. Section 4 briefly discusses semi supervised technique, active learning. And finally Section 5concludes the paper.

2. Related Previous Work

We found a number of NER systems in the literature. Most of the systems work in general or newswire domain where the major NEs are person, location and organization. Some domain specific NER systems are also there; these mainly work in biomedical (NEs are protein, DNA, RNA etc.), chemical or historical domains, but identifying NE like disease, drug and symptom is hardly available.

First we discuss a few works on NER task in general domain that primarily use supervised techniques. BBN's IdentiFinder (Daniel M. Bikel et al., 1997) is one of the most popular NER systems. This system is developed using Hidden Markov Model (HMM) and using word feature, capitalization and digit features. Mikheev et al. (1998) proposed a system, which worked on the MUC-7 data. Andrew Borthwick (1999) developed a Maximum Entropy based Named Entity (MENE) recognition system which was combined with a hand-coded system (namely, Proteus). GuoDong and Jian Su (2002) proposed a generative HMM-based chunk tagger. Apart from supervised learning, semi-supervised learning techniques are also used in NER system development. In late 90's Collins and Singer (1999) proposed a bootstrapping based technique to show that a

small amount of labeled data can be useful to develop a NE classifier. Mohit et al. (2005) proposed a syntactic features based semi supervised named entity (NE) tagger. A Maximum entropy (MaxEnt) based semi-supervised learning technique has been proposed by Saha et al. (2009) for NER in Hindi. Downey et al. (2007) introduced a novel approach to the first step to identify NE from online text. This system is capable of identifying complex named entities from Web corpus. The system is based on n-gram feature which is useful to recognize the entities, considered as a species of multiword units. Here we also discuss few works on biomedical NER. In biomedical domain several annotated corpora are openly available; for example, GENIA (J. Kim et al., 2003), JNLPBA (J. Kim et al., 2004), BioCreative (Yeh. A et al., 2005) and BioInfer (Pyysalo. S el al., 2007). Therefore most of the researchers used these corpora for the training purpose. Collier et al. (2000) described a Hidden Markov Model (HMM) based NER system to find NE from molecular-biological corpus using bigram feature. Ponomareva et al. (2007) developed a HMM based NER system to identify NE from biomedical text. The advantage of their work is that it only used POS information as domain knowledge. Shen D et al. (2003) proposed another HMM based model for biomedical NER. Also some good works using deep domain knowledge available in biomedical domain are "Exploring deep knowledge resources in biomedical name recognition" by Zhou and Su (2004). A maximum Entropy based hybrid system was proposed by Lin et al. (2004). This system is a combination of two steps process first uses machine learning algorithm and second post-processing uses rule based and dictionary-based technique. Kazama et al. (2002) introduced a Support Vector Machine (SVM) to find biomedical named entity in GENIA corpus. Burr Settles (2004) proposed a conditional random field (CRF) based machine learning system to recognized biological NE like PROTEIN, DNA, RNA, CELL-LINE, and CELL-TYPE by using Orthographic and Semantic Feature sets. We also find some related works on biomedical NER based on CRF in Tsai T et al. (2006). And a conditional random field (CRF) based open-source, executable survey, BANNER in biomedical named entity recognition has been presented by R. Leaman et al. (2008). Suakkaphong et al. (2011) proposed a system that is combination of semi supervised learning and conditional random fields, statistical machine learning method which has shown advantages over other statistical machine learning methods for NER to recognize disease named entities from biomedical texts.

3. Supervised Learning Using CRF

Here we discuss our baseline NER system which is based on Conditional Random Field (CRF) and in the domain of homeopathic forum texts. We have prepared a training data containing ~100K words collected from a homeopathy diagnosis related discussion forum. We have worked on various feature sets chosen from the set of candidate features mentioned in Section 3.3. The detail of the system is discussed below.

3.1. Conditional Random Field Model

CRF is a probabilistic framework for labeling and segmenting sequential data such as text (J. Lafferty et al., 2001). It is an undirected graphical models used to calculate the conditional probability of label sequences (S) given some observations sequences (O) (Wallach, H. M. et al., 2004). Applying CRF to an observation sequence which is the token sequence of text and state sequence is the corresponding label sequence in NER system. The conditional probability of a state sequence $S=<S_1, S_2...S_N>$; given an observation sequence $O=<O_1, O_2...O_N>$ is

$$P(s/o) = 1/(Z(o)) \exp \sum_{i=1}^{N} \sum_{j=1}^{M} \lambda_j f_j (S_{i-1}, S_i, o, i)$$

fj (s_{i-1}, s_i, o, i)is the feature function whose weight λ_j is to be learned via training and Z(o) is a normalization factor.

$$Z (o) = \sum_s \exp \sum_{i=1}^{N} \sum_{j=1}^{M} \lambda_j f_j (S_{i-1}, S_i, o, i)$$

3.2. Training and Test Data Set

The data set that we have used to train our system is taken from ABC homeopathic discussion forum. The corpora contain user discussion on various diseases and their diagnosis solutions. As the texts in the discussions are written by normal web-users, the corpora contain some amount of noise like spelling mistake, abbreviation, ungrammatical sentences etc. In this data set we are mainly interested on drugs and diseases names. From the corpora we have randomly selected about 6.5K sentences and annotated manually. The data contains ~100K words having ~2270 medicine and disease names. This manually annotated data is used to train the CRF model. Also we have annotated about 600 sentences (~12K words) which are used for testing the system. In our corpus we have considered two NE categories namely, Disease name (D) and Medicine name (M). The corpus is annotated using BIO format where 'B' represents the beginning word of a NE, 'I' represents the subsequent words of a NE that contains two or more words and 'O' denotes the not-name words. For example, a disease name "Juvenile Rheumatoid Arthritis" and a medicine name "Boiron Cina Mother Tincture" are annotated as

Juvenile # BD Rheumatoid # ID Arthritis # ID
Boiron # BM Cina # IM Mother # IM Tincture # IM

3.3. Feature Set Used to Train CRF Model

Features play an important role in ML based NER system development. In our baseline system we have worked on different types of candidate features and choose the best feature set. The selected candidate features do not require any deep domain knowledge.

3.3.1 Word Feature

To identify NE from bio-medical corpus the word feature is very much helpful. We have used the current word along with preceding and following words. That is word window of size three and five have been used in which target word is at the middle. If *W* is the target word then *W-2*, *W-1* are the two preceding words and *W+1*, *W+2* are the two following words in the word window.

3.3.2 Affix Feature

Especially in bio-medical domain the affix feature is highly important to identify the NEs. We have mainly used prefix and suffix of variable length for the training purpose of our system.

3.3.3 Capitalization Information

We have used different types of capitalization information as feature. The features we have used in our system are, *initial_capital* (the words starting with capital letter), *all_capital* (all the letters are capital) etc.

3.3.4 Numerical Feature

It is often found that medicine names are associated with some numeric values which represent their power, like Belladonna 30C, Arnica 10m, Gelsemium 6C etc. Therefore in our system we have used numerical information based features like *is_numerical* (feature value is true if the word contains any number).

3.3.5 Parts-of-speech Information

Part of Speech (POS) information is also an important feature for Named Entity Recognition System. Mainly the POS information of the target word and its surrounding words are used in our system.

3.4. Result

We have worked on various combination of the candidate features mentioned above in order to find out the best feature set. The identified best feature set is used to develop our baseline system. We have measured the performance of the system in terms of f-measure or f-value (F) which is defined as the harmonic mean of precision and recall.

$$F=\frac{(1+\beta^2)(\text{precision}\times\text{recall})}{(\beta^2\times\text{precision}+\text{recall})}$$

Recall is the ratio of number of NE words retrieved to the total number of NE words actually present in the corpus and precision is the ratio of number of correctly retrieved NE words to the total number of NE words retrieved by the system. β^2 represents the relative weight of recall to precision and normally its value is taken as 1. In Table I we have summarized the experimental results of our baseline NER system using the identified feature sets. The last row in the table presents the highest accuracy of the system. The highest accuracy obtained in the system is f-score 83.29 where we have used word, affix, POS, numeric and capitalization features. Here we have used suffixes and prefixes of length up to three and a word window of length up to five. Numerical, Capitalization and POS features are also effective. In the experiments we have observed that numerical features help to improve the accuracy of the medicine name class. In general domain it is reported by many researchers that the capitalization features are very much important in identifying the NEs. But in this domain we have seen that the capitalization features are not much helpful, the improvement in accuracy after adding the capitalization features is low. This is because of the noisy characteristics of the corpus; the NEs are not properly capitalized. From the Table I we have also observed that the recall values are low. For example, the recall value corresponding to the highest f-value is 77.80%, where the precision is 89.63%. That implies many of the NEs are not recognized by the system. This is due to the unavailability of sufficient annotated data. Now, to improve the system with the existing resource we plan to adopt the semi-supervised learning (SSL) technique where un-annotated data can be used along with the annotated data to train the system.

Table 1. Experimental Result Based on Feature Set

Feature Set	Precision	Recall	F-Measure
Word Window Three	90.69	66.20	76.53
Word Window Five	92.32	67.40	77.90
Words, Affix Length Two	89.41	76.00	82.16
Words, Affix Length Three	89.95	75.20	81.91
Word, Numeric, Capitalization	90.81	67.20	77.24
Word, Affix, Numeric, Capitalization	89.25	76.40	82.32
Word, Affix, POS	90.02	75.80	82.30
Word, Affix, POS, Numeric, Capitalization	**89.63**	**77.80**	**83.29**

4. Semi Supervised Learning

Semi-Supervised Learning (SSL) is much cheaper compared to the supervised learning in terms of leveled training data. Recently machine learning researchers have paid more attention to semi supervised learning to

reduce the need for large amounts of labeled training data. SSL is a machine learning technique that uses a small amount of labeled and a large amount of unlabeled data for training purpose. There are some well know semi supervised learning technique like Bootstrapping, Active learning etc. In our experiments we have used an active learning technique to improve the performance of our NER system.

4.1. Active Learning

Active learning is a technique by which we can overcome the draw-fall of using huge amount of manual annotated corpus for training purpose. Active learning is well-motivated in many modern machine learning problems where generating label training data is expensive. Some Active learning related works are found in literature like Shen et al. (2004) proposed multi-criteria based active learning approach to recognize name entity from biomedical text as well as in news wire domain. Yao et al. (2009) proposed an active learning technique, based on information density along with CRFs for Chinese Named Entity Recognition (NER).

In this learning process, a model is trained on a previously labeled data set, and then it classifies an unlabeled set to get self-labeled data. Then the confidence of labeling is measured and the low confidence portion is extracted. Next the low confident self-labeled data sets are given to human annotators (teacher) and corrected. Then this portion is added to the original training data set and a new classifier is trained. Based on different active learning settings, there are mainly three types active learning technique (B. Settles, 2009) (i) membership query synthesis, (ii) stream-based selective sampling, and (iii) pool-based active learning. We have used pool-based active learning technique in our system.

4.1.1 Active Learning in our System

In active learning we have fed the low confidence data to our CRF based trained model with some special handling. At first we have taken ~2000 raw sentences (D) from the ABC homeopathic discussion forum and label them using our baseline system M (trained using manually annotated training data T). Now classifier annotation confidence for each sentence in label data set (P) is extracted. The data set P is sorted in ascending ordered based on low confidence measure. Following this P is divided into equal size sub set p1, p2 ….pn of 200 sentences (~4K words) each and annotated using human supervision on the low-confident words of the top subsets (P_i). We add each sub set p_i to the original training data T in each iterations and calculate the performance of the resulting system. The active learning procedure used in our system is described below.

Algorithm 1: Active Learning Procedure
Begin
1. Select a raw data D.
2. Label D using baseline CRF model M.
3. Extract the confidence measure on system annotated data P.
4. Sort data set P in ascending order based on low confidence measure.
5. Divide P into equal sized subsets P_1, P_2, …, P_n.
6. Apply human supervision on the low-confident words of the top subsets (P_i).
7. Add P_i to T to generate T_i and train system using T_i.
8. Execute CRF on T_i and calculate F-Measure.
9. Repeat step 6 to 8 until no data is available or f-measure decreases.

End.

The active learning based experimental results are presented in Table 2. Here we have seen that in first few iterations addition of machine annotated data helps to improve the performance. Also we have observed that the recall value is increasing. The highest accuracy obtained in the system is a f-value of 84.35 (iteration 6 in the Table 2). The corresponding precision is 89.46% and the recall is 79.80%. In the next iteration (iteration 7)

the accuracy is degraded and we have stopped the SSL process. Comparing with the baseline accuracy we can observe that the recall value using the active learning technique is increased to 79.80% from baseline's recall value 77.80%. Therefore we can conclude that the active-learning based SSL technique is effective in our task.

Table 2. Experimental Result Based on Active Learning

Iteration	Precision	Recall	F-Measure
1.	89.65	78.00	83.42
2.	89.67	78.20	83.54
3.	89.16	79.00	83.77
4.	89.36	79.00	83.86
5.	89.63	79.60	84.32
6.	**89.46**	**79.80**	**84.35**
7.	89.43	79.60	84.23

5. Conclusion

In this paper we have presented a NER system in the homeopathic diagnosis discussion domain. First we have manually annotated a corpus containing ~112K words and used supervised learning technique to develop a system. There we have used CRF along with a set of identified features. As the system suffers from scarcity of annotated corpus, next we use semi-supervised learning technique in order to improve the system. We have used active-learning technique where we have used human supervision to correct the low confident annotations. In our experiments we have seen that use of active learning helps to improve the accuracy.

References

Grishman R., 1995. The New York University System MUC-6 or Where's the Syntax? In: Proceedings of the sixth message understanding conference; pp. 1-11

R. Grishman and B. Sundheim., 1996. Message understanding conference- 6: A brief history. In Proc. of COLING, pp. 466-471

Fukuda K, Tsunoda T, Tamura A, Takagi T., 1998. Toward information extraction: identifying protein names from biological papers. In: Proceedings of the pacific symposium on biocomputing; pp. 707–18

Saha S.K, Mitra P, Sarkar S., 2009. Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of Biomedical Informatics 42, pp. 905–911

Sobhana N.V, Pabitra Mitra, S.K. Ghosh., 2010. Conditional Random Field Based Named EntityRecognition in Geological Text. International Journal of Computer Applications (0975 – 8887) .Volume 1 – No. 3, pp. 119-122

Daniel M. Bikel, Scott Miller, Richard Schwartz,and Ralph Weischedel.Nymble., 1997. A highperformancelearning name-finder. In Proceedingsof the Fifth Conference on Applied NaturalLanguage Processing, pp. 194–201.

Andrei Mikheev, Claire Grover, and Marc Moens., 1998. Description of the LTG system used forMUC-7. In Proceedings of the Seventh MessageUnderstanding Conference.

Andrew Borthwick., 1999. A Maximum EntropyApproach to Named Entity Recognition. Ph.D.thesis, Computer Science Department, New YorkUniversity.

GuoDong Zhou. Jian Su., 2002. Named Entity Recognition using an HMM-based Chunk Tagger. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. (ACL), Philadelphia, pp. 473-480.

Collins, M., Singer, Y., 1999. Unsupervised models for named entity classification. In: Proceedingsf the Joint SIGDAT Conference on Empirical Methods in Natural LanguageProcessing and Very Large Corpora

Mohit, B., Hwa, R., 2005. Syntax-based semi-supervised named entity tagging. In: Proceedingsof the ACL Interactive Poster and Demonstration Sessions, pp. 57–60.

S.K. Saha, P. Mitra, and S. Sarkar.,2009. PReMI , LNCS 5909,. ©Springer-Verlag Berlin Heidelberg , pp. 225–230, 2009

Doug Downey, Matthew Broadhead, and Oren Etzioni., 2007. Locating Complex Named Entities inWeb Text. Proceedings of the 20th international joint conference on Artifical intelligence. pp. 2733-2739

Kim J, Ohta T, Tateisi Y, Tsujii J., 2003. Genia Corpus—a semantically annotated corpus for bio-textmining. Bioinformatics (Supplement: Eleventh International Conference on Intelligent Systems for Molecular Biology); pp.180–182.

Kim J, Ohta T, Tsuruoka Y, Tateisi Y, Collier N., 2004. Introduction to the bio-entity recognition task at JNLPBA. Nazarenko, editors, proceedings of the International Joint Workshop on Natural; pp. 70-75

Yeh A, Morgan A, Colosimo M, Hirschman L., 2005. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinfo; 6(Suppl. 1):S2.

Pyysalo S, Ginter F, Heimonen J, Bjorne J, Boberg J, Jarvinen J., 2007. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinfo 2007;8(50)

Collier N, Nobata C, Tsujii J., 200. Extracting the names of genes and gene products with a hidden Markov model. In: Proceedings of COLING: pp. 201-207

Ponomareva N, Pla F, Molina A, Rosso P., 2007. Biomedical named entity recognition: a poor knowledge HMM-based approach. LNCS 4592, pp:382–387.

Shen D, Zhang J, Zhou GD, Su J, Tan CL., 2003. Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In: Proceedings of ACL workshop on natural language processing in biomedicine; pp. 49–56.

Zhou GD, Su J., 2004. Exploring deep knowledge resources in biomedical name recognition. In: Proceedings of joint workshop on natural language processing in biomedicine and its applications (JNLPBA-2004); pp. 96–9.

Lin YF, Tsai TH, Chou WC, Wu KP, Sung TY, Hsu WL., 2004. A maximum entropy approach to biomedical named entity recognition. In: Proceedings of 4th workshop on data mining in bioinformatics; pp.56-61

Kazama J, Makino T, Ohta Y, Tsujii J., 2002. Tuning support vector machines for biomedical named entity recognition. In: Proceedings of the workshop on natural language processing in the bio-medical domain at ACL; pp. 1–8.

Settles B., 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of joint workshop on natural language processing in biomedicine and its applications (JNLPBA-2004); pp. 104-107

Tsai T, Chou WC, Wu SH, Sung TY, Hsiang J, Hsu WL., 2006. Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. Expert Syst Appl. vol 30(1), pp.117–28.

Leaman R, Gonzalez G., 2008. Banner: an executable survey of advances in biomedical named entity recognition. Pacific Symp Biocomput; pp.652–63

Nichalin Suakkaphong, Zhu Zhang, and Hsinchun Chen., 2011. Disease Named Entity Recognition Using SemisupervisedLearning and Conditional Random Fields. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY; pp. 727–737

J. Lafferty, A. McCallum, and F. Pereira., 2001. Conditional random fields: probabilisticmodels for segmenting and labeling sequence data. In InternationalConference on Machine Learning; pp. 282 - 289

Wallach, H. M., 2004. Conditional random fields: Anintroduction, Technical Report MS-CIS-04-21, University ofPennsylvania, Department of Computer and InformationScience, University of Pennsylvania.

Shen D, Jie Zhang, Jian Su, Guodong Zhou, Chew-Lim Tan., 2004. Multi-criteria-based active learning for named entity recognition. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics

Lin Yao, Chengjie Sun, Shaofeng Li, Xiaolong Wang, Xuan Wang., 2009. CRF-based Active Learning for Chinese NamedEntity Recognition. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. USA; pp. 1557-1561

B. Settles., 2009. Active Learning Literature Survey. In Proceedings of theConference on Empirical Methods in Natural Language Processing(EMNLP)