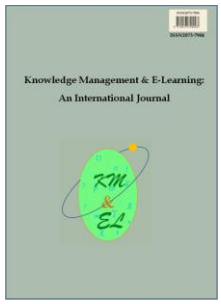
Knowledge Management & E-Learning



ISSN 2073-7904

Use of global context for handling noisy names in discussion texts of a homeopathy discussion forum

Mukta Majumder
Sujan Kumar Saha
Birla Institute of Technology, Mesra, India

Recommended citation:
Majumder, M., & Saha, S. K. (2014). Use of global context for handling noisy names in discussion texts of a homeopathy discussion forum. *Knowledge Management & E-Learning, 6*(1), 18–29.

Knowledge Management & E-Learning, 6(1), 18–29

Use of global context for handling noisy names in discussion texts of a homeopathy discussion forum

Mukta Majumder*

Department of Computer Science and Engineering
Birla Institute of Technology, Mesra, India
E-mail: mukta_jgec_it_4@yahoo.co.in

Sujan Kumar Saha

Department of Computer Science and Engineering
Birla Institute of Technology, Mesra, India
E-mail: sujan.kr.saha@gmail.com

*Corresponding author

Abstract: The task of identifying named entities from the discussion texts in Web forums faces the challenge of noisy names. As the names are often misspelled or abbreviated, the conventional techniques have failed to detect the noisy names properly. In this paper we propose a global context based framework for handling the noisy names. The framework is tested on a named entity recognition system designed to identify the names from the discussion texts in a homeopathy diagnosis discussion forum. The proposed global context-based framework is found to be effective in improving the accuracy of the named entity recognition system.

Keywords: Named entity recognition; Homeopathy; Discussion forum; Global context; Noisy text

Biographical notes: Mukta Majumder is a Ph.D research scholar in Computer Science and Engineering Department, Birla Institute of Technology, Mesra, Ranchi, India. He has completed his post graduation from National Institute of Technical Teachers Training and Research's, Kolkata, India and graduation from Jalpaiguri Government Engineering College, Jalpaiguri, India. His main research interests include Text Processing, Machine Learning, Microfluidic System, and Biochip etc.

Dr. Sujan Kumar Saha is an Assistant Professor in Computer Science and Engineering Department, Birla Institute of Technology, Mesra, Ranchi, India. He has completed his Ph.D from IIT Kharagpur, India, post graduation from IIT Delhi, India, and graduation from Kalyani Government Engineering College, Kalyani, India. His main research interests include Natural Language Processing and Machine learning etc.

1. Introduction

Named entities are the pivot elements of a textual document; therefore identifying named entities is one of the elementary tasks of information extraction and data mining. Named

Entity Recognition (NER) is the task of identifying and classifying the names in text. In this paper we present a NER system for identifying the names from the discussion text of a web discussion forum.

We chose an online homeopathy discussion forum namely http://www.abchomeopathy.com/ for the study. In this forum a patient can discuss about his or her diseases and symptoms and ask for the appropriate remedy to the doctor or expert members of the forum. As an affordable diagnosis, homeopathy treatment is always very popular to common people. With the huge popularity of the Internet, online discussion forums in homeopathic domain have received increased attention from those people. These disease-symptom-medicine related discussions carry a huge amount of valuable information which can be used effectively in various applications like developing automatic homeopathy clinical decision support systems or diagnostics systems and homeopathy remedy-disease related data bases. For developing such applications using this data, identification of medicine and disease names is obligatory. In this study, we attempted to develop a NER system in the domain of homeopathy web discussion forum text.

Designing 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 numeric values (especially with drug names) in between two words or at end. This makes the task of classification and boundary identification quite difficult.

Difficulty for identifying drug and disease NEs from online homeopathic diagnosis discussion forum corpus rather increases because of its noisy nature. Due to the informal setting, forum texts are highly error prone and contain various textual noises like misspellings, abbreviations, etc. Use of capitalization, parenthesis, hyphen and abbreviation in forum text does not follow a standard convention. The named entities in these texts are also noisy. As a result of these noises and informal nature of the texts, standard Natural Language Processing (NLP) tools, which are designed for general domain, often fail to produce moderate accuracy. Development of NLP tools or systems on this type of corpora requires some special techniques.

To develop a NER system primarily two approaches have been followed: rule based and machine learning based. Rule based approach (Grishman, 1995; Fukuda, Tsunoda, Tamura, & Takagi, 1998) requires domain expertise and a set of linguistic rules which are defined to identify the names. On the other hand machine learning based approaches (Borthwick, 1999; Kazama, Makino, Ohta, & Tsujii, 2002; Zhou & Su, 2002) require labeled training corpus where names are annotated manually. A machine learning algorithm uses this training data and a set of relevant features to extract required statistics in order to identify the names from a test data. For this NER system development we have used the machine learning based approach where we use Conditional Random Field (CRF) as the classification algorithm. For the task we have manually annotated a corpus containing ~150K words; ~135K of which is taken for training and 15K for testing purpose. We have considered two types of NEs, namely, drug names and disease names.

The performance of a machine learning classifier largely depends on the amount of its training data. As the training corpus is noisy and not sufficiently large, we observe that the system is unable to identify many names. We have analyzed the unidentified names and observed that a high portion of these are noisy. To improve the performance of the system we next decided to employ a framework for handling the noisy names.

In this paper we propose a technique for identifying the noisy names which are not recognized by the CRF based baseline system. The proposed technique is based on *Global Context* of the entities. Preparation of annotated data is costly and time consuming but a large amount of raw data is easily available. Therefore we make use of the raw forum text for extracting the global context. First we find the confidence measure of CRF, identify the tokens for which the classifier is less confident. Then for these tokens we extract their context (containing previous and next words) for their all occurrences in the discussion forum corpora. Next we check whether these contexts match the NE contexts extracted from the manually annotated training data. Accordingly we update the class specific probability value provided by the CRF classifier and run a Beam-search algorithm to re-annotate the data. In our experiments we observe that, this *Global Context* based re-annotation technique is able to identify a set of new NEs that improves the overall performance of the system.

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 represents the Conditional Random Field based baseline NER system. Section 4 describes a noisy named entity identification framework using global information. Section 5 presents the result of global context based framework and comparative discussion of the proposed system with other systems. Finally Section 6 discusses the conclusion and the future works.

2. Related work

In the literature a lot of NER systems are available which primarily work in general or newswire domain where the NEs are mainly person, location and organization names. A number of NER systems are also available that are targeted to identify domain specific NEs; for example, biomedical domain (NEs are protein, DNA, RNA etc.), chemical and historical domains. In the literature we are unable to find much work for identifying drug, disease and symptom names in Homeopathy domain.

At first we discuss a few works on the development of NER system that used a supervised classifier as the core module. BBN's IdentiFinder (Bikel, Miller, Schwartz, & Weischedel, 1997) is a popular one of these NER systems. This system is developed using Hidden Markov Model (HMM) along with word, capitalization and digit features. HMM was used in several other NER systems such as Collier, Nobata, and Tsujii (2000); Zhou and Su (2002); Shen, Zhang, Zhou, Su, and Tan (2003); Ponomareva, Pla, Molina, and Rosso (2007). Maximum Entropy classifier was used in the 'MENE' system developed by Borthwick (1999). Some other works which used Maximum Entropy classifier as machine learning algorithm are Lin et al. (2004) and Saha, Mitra, and Sarkar (2009). Support Vector Machine (SVM) is another machine learning classifier which is widely used for developing NER system (Kazama, Makino, Ohta, & Tsujii, 2002). A Conditional Random Field (CRF) based open-source, executable survey, 'BANNER' in biomedical named entity recognition has been presented by Leaman and Gonzalez (2008). Some other NER systems that used CRF are Settles (2004); Tsai et al. (2006).

Many of the systems used some external modules, post processing techniques, or domain knowledge to improve the performance. For example, MENE was combined with a hand-coded system Proteus (Borthwick, 1999); (Ponomareva, Pla, Molina, & Rosso, 2007) used some domain knowledge like POS information; the system developed by Zhou and Su (2004) used deep domain knowledge such as word information pattern, morphological pattern, out domain POS and semantic trigger to identify biomedical NEs. The Maximum Entropy based hybrid system by Lin et al. (2004) is a combination of two

stage process; first uses machine learning algorithm and second post processing uses rule based technique.

In recent times a substantial amount of research works have been carried out for extracting different kinds of information from informal web text (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010; Liu, Zhang, Wei, & Zhou, 2011; Majumder, Barman, Prasad, Saurabh, & Saha, 2012; Chan, Huang, Hui, Li, & Yu, 2013). Like other information extraction tasks identification of names from informal web text like blog, discussion forum, twitter etc. is more difficult than that of on a formal text. Here we present a few systems that worked on web text. Ritter, Clark, Mausam, and Etzioni (2011) proposed a T-NER system to identify NEs from Twitter data. By using LabeledLDA they have further increased the accuracy of their system. There are many other NER systems which work on Twitter text (Liu, Zhang, Wei, & Zhou, 2011; Li et al., 2012). Downey, Broadhead, and Etzioni (2007) introduced a novel approach to identify NE from online text. Their system is able to identify complex NEs 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. An automatic tagger for NER from online web corpus was presented by An, Lee, and Lee (2003). They have used an NE list and a web search engine to collect web documents which contain the NE in-stances. Then the data is refined through sentence separation and text refinement procedures and NE instances are finally tagged with the appropriate NE categories. There are some other NER systems which worked on online corpus to extract and classify NE (Ben Abdessalem Karaa, 2011). The similarity of all these system is that they all work on general domain NEs like, person, location, organization, date, time, title etc. Many of the researchers found difficulty in identifying NEs from online noisy text.

Identifying drug and disease NEs from diagnosis text is very rare. Only a few works are available in this domain. A CRF based NER system was developed by Suakkaphong, Zhang, and Chen (2011) to identify the disease names from biomedical literature (MEDLINE Abstract). This system also used two semi supervised techniques, bootstrapping and feature sampling. Majumder et al. (2012) proposed a CRF based NER system to indentify Drug and Disease NEs from an online discussion forum corpus. The performance of this system is further enhanced by the use of an active-learning based semi supervised framework. But none of these systems was focused on handling noisy NEs.

3. Proposed baseline NER system using CRF

This section describes our baseline NER system based on Conditional Random Field (CRF) which uses a homeopathy discussion forum corpus as train and test data. The size of our training data is ~135K words and test data is ~15K words. 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 (CRF) model

Conditional random field (CRF) is a probabilistic framework for labeling and segmenting sequential data such as natural language text (Lafferty, McCallum, & Pereira, 2001). In the last few years CRF is used widely in various NLP tasks like NER (Settles, 2004; Tsai et al., 2006), Multiple Choice Question (MCQ) generation (Goto, Kojiri, Watanabe, Iwata, & Yamada, 2010) etc. CRF is an undirected graphical models used to calculate the

conditional probability of values on desired output nodes given values assigned to other designated input nodes (Wallach, 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

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

Where f_j (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. Here Z (o) is calculated as

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

3.2. Training and testing data set

The data set that we have used to train our baseline NER system is taken from http://www.abchomeopathy.com/. In this data set we are mainly interested on drug and disease names. We have manually annotated ~135K words to train our baseline system and ~15K words for testing. The details about the data size are shown in Table 1. In the corpus we have considered only two NE categories, Disease name (SD-start of disease, CD- subsequence word of disease NE) and Medicine name (SM- start of medicine and CM subsequence word of medicine NE). The word other than NE category is tagged as '#O'.

For example, the data is annotated as follows:

High Blood Pressure (A Disease name): High #SD Blood #CD Pressure #CD

Arnica Montana 30C (A Medicine Name): Arnica #SM Montana #CM 30C #CM

Table 1
The data set

Total Amount of Data Selected For Annotation	~10K Sentences
Total Words in Annotated Data	~150K Words
Train Data Size	~135K Words
Test Data Size	~15K Words

3.3. Feature set used to train the CRF model

In the literature we observe that for the development of NER system a number of features have been used. In this work our primary objective is to test the performance of the global context, therefore we have used a simple and easily derivable feature set containing the surrounding words, affix, POS, numeric and capitalization information. Here we have experimented with word window and affixes of various length and chosen the best one.

3.3.1. Word feature

For building NER system word feature is widely used. We have used the current word along with preceding and following words. That is word window of size three; five and seven have been used in which target word is at the middle.

3.3.2. Affix feature

In bio-medical domain the affix feature is highly important to identify the NEs. We have mainly used prefix and suffix of variable length (two and three) for the training purpose of our baseline NER system.

3.3.3. Numeric feature

In homeopathy discussion forum corpus it is often found that medicine names are associated with some numeric values which represent the power of that particular drug, like Belladonna 30C, Arnica 10m, Gelsemium 6C etc. Therefore in our system we have used numerical features, like *is_numerical* (feature value is true if the NE contains any number).

3.3.4. Parts-of-speech (POS) information feature

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

3.3.5. Capitalization feature

It is found that Name Entity words are often capitalized. So we have used different types of capitalization information as feature. The features we have used in our system are, *initial_capital* (the word starting with capital letter) and *all_capital* (all the letters of the word are capital).

3.4. Performance of the baseline system

The performance of the system is measured in terms of f-measure or f-value which is defined as the harmonic mean of precision and recall.

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

Where 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. The experimental results of our Conditional Random Field based baseline NER system using the candidate feature set is summarized in Table 2.

Table 2
Experimental result of CRF based NER using the feature set

Feature	**Recall**	**Precision**	**F-Measure**
Word Window Three	66.97	91.25	77.25
Word Window Five	67.46	92.40	77.98
Word Window seven	66.29	90.77	76.62
Word and Affix of length Two	76.11	89.59	82.30
Word and Affix of length Three	75.63	89.74	82.08
Word, Affix, Capitalization	74.54	89.03	81.14
Word, Affix, Numeric	76.41	89.23	82.32
Word, Affix, POS	76.76	90.19	82.93
Word, Affix, POS, Numeric	**77.29**	**90.61**	**83.42**
Word, Affix, POS, Numeric, Capitalization	77.39	89.66	83.07

From the Table 2 we observe that the system achieves the highest f-mesure of 83.42 with precision 90.61 and recall 77.29 using the candidate feature set Word, Affix, POS and Numeric information. Suffix and prefix of variable length (two and three) and word window up to seven have been used. It is found that POS information for identifying drug and disease NEs is highly effective for discussion forum corpus. In experiment we have observed that numerical information is helpful to recognize medicine name, as drug NEs are often associated with some numerical value which specifies its power. But on overall accuracy it does not have much impact; as this feature is not ideal to identify disease NEs. In general domain it is reported by many researchers that the capitalization features are very much important in identifying the NEs. But in this homeopathy discussion forum domain we have seen that the capitalization features are not much helpful. As the text is noisy, Name Entities are not capitalized following standard grammatical rule and convention.

We observe that a number of NEs are not identified by the system as they are noisy. In order to identify these noisy names we use global context which is discussed in the next section.

4. Noisy named entity identification using global context

In the discussion forum text there is always a high probability of existing textual noise like misspelling and nonstandard abbreviations coined by the users. For example, in this homeopathy forum we have found that the actual disease names ‘Fistula’, ‘Fever’, ‘Abscess’ are often written in misspelled form like ‘Fistualla’, ‘Fiver’, ‘Absess’ respectively. Similarly the drug name ‘Nux Vomica’ is misspelled as ‘Nux Vom’ or ‘Nux Vomita’; ‘Silicea’ is misspelled as ‘Silecea’; ‘Nux Vomica’ is also sometimes abbreviated as ‘N-Vom’. In such cases our base line NER system is unable to identify these noisy NEs properly. Global context can be facilitative to recognize these misspelled

and abbreviated NEs. We use global context information to update the class specific probability value and re-annotate the test data. Our approach of using global context is summarized below.

4.1. Data used for global context

In this "ABC Homeopathy" discussion forum when a user initiates a discussion he/she introduces a new topic about that discussion. We track these topics and find those which contain maximum number of posts (topic with more than 40 posts). We have extracted ~30K posts on different topics available in the diagnosis discussion forum namely http://abchomeopathy.com/forum2.php as our global context reference set. Preparation of labeled data is costly and time consuming but these large amount of raw data is easily available. Therefore we make use of the raw forum text for extracting the global context.

4.2. Proposed global context based named entity recognition (GCBNER)

First we make a not-name word list from the training data. This list is not the complete list of not-name words but it will be used to reduce our re-annotation effort. We also make a class-specific NE context list by considering the previous 3 words and next 3 words of the NEs in the training data.

Next, for the test data we extract the probability of belongingness of the words into the classes (NE classes and the not name) computed by the CRF classifier. We find the words having close probability value (difference is less than 0.1) in the top two classes. Also we find the words that are identified as not-name by the CRF classifier but not occurring in the not-name list prepared from training data. These words will be re-annotated using global context.

GCBNER: is a global context based procedure to re-annotate test data to find Drug and Disease NE.

1. Make a not NE list (NNList) from training data.
2. Compile a class-specific NE context list (ContextList) with word window 7.
3. Find CRF probability distribution for each word in test data.
4. Select words that are not present in NNList but classified as not NEs by CRF.
5. Retrieve context information for these not NE words indentified by CRF at step 4 from global data.
6. Match these not NE's context with the ContextList.
   If more than one matches are occurred then:
   - Increase the class specific probability value of that word where match is found by a factor of 0.33 for corresponding class.
   - Reduce probability of other classes proportionally to keep the sum of probability as 1.
7. Run Beam-search algorithm for sequencing and re-annotation.

Fig. 1. The procedure of GCBNER

Find these identified words in the total forum data and for all occurrences of the word retrieve context information. Match these contexts with the NE context list extracted from the training data.

If more than one match is there (first match is obvious as the training data is also created from this discussion forum corpus) then increase the class specific probability value for that word by a factor of 0.33 (1/3 as, 3 classes are there – drug, disease and other or not-name) for that class. Reduce probability values for other classes proportionally to keep the sum of probabilities as one. Run Beam-search (Koehn et al., 2007; Dahlmeier & Ng, 2012; Wang & Ng, 2013) algorithm for sequencing and re-annotation.

The details of the proposed Global Context Based Named Entity Recognition (GCBNER) procedure are described in Fig. 1.

5. Result and discussion

This global context based procedure identifies a set of new entities that were not identified by the baseline system. Hence the accuracy of the system improves. With global context the system achieves an f-value of 86.09. Corresponding precision is 91.32 and recall is 81.43 (see Table 3). This improvement demonstrates that the proposed global context framework is useful for identifying the noisy names.

Table 3
Experimental result with global information

	Recall	**Precision**	**F-Measure**
Baseline NER's Accuracy On Drug	78.05	90.63	83.87
Baseline NER's Accuracy On Disease	77.12	89.79	82.97
Baseline NER's Over All Accuracy	**77.29**	**90.61**	**83.42**
GCBNER's Accuracy On Drug	81.67	91.82	86.45
GCBNER's Accuracy On Disease	81.23	90.76	85.73
GCBNER's Over All Accuracy	**81.43**	**91.32**	**86.09**

In literature we only find a very few works which deal with disease and drug NE identification. Suakkaphong, Zhang, and Chen (2011) developed a CRF based NER system to identify the disease NE from standard grammatical text (biomedical literature, "MEDLINE") which achieved an accuracy of f-measure of 73.94. This system also used two semi supervised techniques, bootstrapping and feature sampling to boost-up its performance. Their system is only cable of identifying disease NEs; it has no concern with medicine NEs. Another CRF based NER system has been proposed by Majumder et al. (2012) to indentify drug and disease NEs from an online discussion forum corpus. The performance of this system is further enhanced by the use of a semi supervised technique, namely active learning which achieved a highest accuracy of f-value 84.35. But the problem of handling noisy drug and disease NEs was not taken care in these works discussed above. Our proposed technique which achieves an accuracy of f-measure 86.09 works in online discussion forum corpus and efficiently identifies noisy drug and disease NEs.

To identify the noisy names we have used global information extracted from raw forum data. As the forum data is noisy in nature, misspellings and abbreviations are often occurred in this corpus. Therefore it may be happened in some cases that using this

discussion forum corpus for extracting global context, increases difficulty or ambiguity to identify names. For example 'Nux' (medicine NE) and 'Not' (not NE, other class) can be misspelled as 'Nut' or 'Nox'. Now the system can incorrectly predict it (misspelled 'Not') as medicine NE. In that case use of global context may decrease the performance or accuracy of the system. This is a limitation of using global context to identify noisy NEs.

6. Conclusion and future work

In this paper we have presented a NER system in homeopathy diagnosis discussion forum domain using Conditional Random Field as machine learning algorithm. Now this NER system can be helpful to develop clinical decision support system or automatic diagnosis system in homeopathy domain. As the discussion forum corpus is noisy in nature; lots of misspelled and abbreviated named entities are there in the corpus. The baseline CRF classifier fails to identify several of these noisy NEs. In order to identify noisy names we have proposed a global context based framework with the help of global information collected from the huge online homeopathy discussion forum corpus. In our experiments we observe that the proposed framework is able to improve the accuracy of the system.

We have shown that our proposed framework perfectly works on homeopathy discussion forum data and identifies noisy drug and disease NEs. In future we like to extend our work by applying the global context based concept in other web data like product reviews, blogs, Twitter data etc.

References

An, J., Lee, S., & Lee, G. G. (2003). Automatic acquisition of named entity tagged corpus from world wide web. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2* (pp. 165–168).

Ben Abdessalem Karaa, W. (2011). Named entity recognition using web document corpus. arXiv preprint arXiv:1102.5728. Retrieved from http://arxiv.org/ftp/arxiv/papers/1102/1102.5728.pdf

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language processing* (pp. 194–201).

Borthwick, A. (1999). *A maximum entropy approach to named entity recognition*. Ph.D.thesis, Computer Science Department, New York University.

Chan, R. Y. Y., Huang, J., Hui, D., Li, S., & Yu, P. (2013). Gender differences in collaborative learning over online social networks: Epistemological beliefs and behaviors. *Knowledge Management & E-Learning (KM&EL), 5*(3), 234–250.

Collier, N., Nobata, C., & Tsujii, J. (2000). Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th conference on Computational linguistics* (pp. 201–207).

Dahlmeier, D., & Ng, H. T. (2012). A beam-search decoder for grammatical error correction. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 568–578).

Downey, D., Broadhead, M., & Etzioni, O. (2007). Locating complex named entities in web text. In *Proceedings of the 20th international joint conference on Artifical intelligence* (pp. 2733–2739).

Fukuda, K., Tsunoda, T., Tamura, A., & Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. *Pacific Symposium on*

Biocomputing, 707–718.

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning (KM&EL), 2*(3), 210–224.

Grishman, R. (1995). The New York University system MUC-6 or Where's the syntax? In *Proceedings of the 6th conference on Message understanding* (pp. 167–175).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177–180). Association for Computational Linguistics.

Kazama, J., Makino, T., Ohta, Y., & Tsujii, J. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3* (pp. 1–8).

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 282–289).

Leaman, R., & Gonzalez, G. (2008). Banner: An executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing, 13*, 652–663.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., & Lee, B. S. (2012). TwiNER: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 721–730). ACM.

Lin, Y. F., Tsai, T. H., Chou, W. C., Wu, K. P., Sung, T. Y., & Hsu, W. L. (2004). A maximum entropy approach to biomedical named entity recognition. In *Proceedings of the 4th workshop on data mining in bioinformatics* (pp. 56–61).

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 359–367).

Majumder, M., Barman, U., Prasad, R., Saurabh, K., & Saha, S. K. (2012). A novel technique for name identification from homeopathy diagnosis discussion forum. *Procedia Technology, 6*, 379–386.

Ponomareva, N., Pla, F., Molina, A., & Rosso, P. (2007). Biomedical named entity recognition: A poor knowledge HMM-based approach. *Lecture Notes in Computer Science, 4592*, 382–387.

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524–1534). Association for Computational Linguistics.

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

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

Shen, D., Zhang, J., Zhou, G., Su, J., & Tan, C. L. (2003). Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL workshop on natural language processing in biomedicine* (pp. 49–56).

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short

text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 841–842). ACM.

Suakkaphong, N., Zhang, Z., & Chen, H. (2011). Disease named entity recognition using semi supervised learning and conditional random fields. *Journal of the American Society for Information Science and Technology, 62*(4), 727–737.

Tsai, T., Chou, W. C., Wu, S. H., Sung, T. Y., Hsiang, J., & Hsu, W. L. (2006). Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. *Expert Systems with Applications, 30*(1), 117–128.

Wallach, H. M. (2004). *Conditional random fields: An introduction.* Technical Report MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania.

Wang, P., & Ng, H. T. (2013). A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the NAACL-HLT* (pp. 471–481).

Zhou, G., & Su, J. (2002). Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 473–480).

Zhou, G., & 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) (96–99).