Biomedical Disease Name Entity Recognition Using NCBI Corpus

Hidayat Ur Rahman
Lahore Leads University
5Tipu Block Near Garden Town Near
Kalma Chowk, Lahore 54000 Pakistan
+92-3329702722
Hidayat.Rhman@gmail.com

Thomas Hahn
University of Arkansas at Little Rock
2801 South University Avenue
Little Rock, AR, 72204
+ 1 (501) 301 4890
Thomas.F.Hahn3@gmail.com

Dr. Richard Segall
Arkansas State University
Computer Inform Tech Department
State University, AR 72404-0130
+ 1 (870) 972-3989
rsegall@astate.edu

***Abstract*— Named Entity Recognition (NER) in biomedical literature is a very active research area. NER is a crucial component of biomedical text mining because it allows for information retrieval, reasoning and knowledge discovery. Much research has been carried out in this area using semantic type categories, such as “DNA”, “RNA”, “proteins” and “genes”. However, disease NER has not received its needed attention yet, specifically human disease NER. Traditional machine learning approaches lack the precision for disease NER, due to their dependence on token level features, sentence level features and the integration of features, such as orthographic, contextual and linguistic features. In this paper a method for disease NER is proposed which utilizes sentence and token level features based on Conditional Random Fields using the NCBI disease corpus. Our system utilizes rich features including orthographic, contextual, affixes, bigrams, part of speech and stem based features. Using these feature sets our approach has achieved a maximum F-score of 94% for the training set by applying 10 fold cross validation for semantic labeling of the NCBI disease corpus. For testing and development corpus the model has achieved an F-score of 88% and 85% respectively.**

***Keywords*— NCBI disease corpus, naïve Bayesian, Bayesian networks, Non nested generalized exemplars;**

I. Introduction

Biomedical Named Entity Recognition (NER) is based on dictionary-based, rule-based and machine learning approaches [1] and [2]. In the dictionary based approach all the terms are not defined in dictionary. This is the major limitation of this approach [3]. Rule-based approaches make decisions based on certain rules, which are learned from the data in form of text terms. But these rules are not applicable in all cases [3]. On the other hand, machine learning approaches require enormous annotated data to train the algorithm [4]. Nowadays machine learning approaches are commonly used for NER, e.g., Support Vector Machines (SVM) [5], Maximum Entropy (ME) [6], Hidden Markov Models (HMM) [7] and Conditional Random Fields (CRF) [8]. In [9] an HMM model has been proposed to distinguish between DNA, RNA, protein, cell-type and cell-line. Kazema et al. proposed an SVM based approach to identify DNA, cell-type, cell-line, protein and lipid achieving an f-score of 73.6% [10]. In [11] CRFs based NER system was developed to recognize protein mentions achieving an F-score of 78.4%. Beside CRFs in [12], the author used ME to distinguish between 23 different biological categories achieving an F-score of 72%.

Performance of biomedical NER as compared to general purpose NER is not satisfactory [13]. Many approaches have been used to enhance the performance of biomedical NER systems, e.g. adding biomedical domain knowledge [14] [15], applying post-processing [14] and combining different machine learning classifiers to perform a hybrid classification scheme [16]. Some of the above mentioned applications are discussed below.

The exact biomedical term could be referred to by abbreviations or synonyms. Therefore, abbreviation and synonym recognition are used to unify and normalize biomedical entities for biomedical NER. For example, in [17] the authors have used logistic regression for abbreviation scoring based on the Medstract corpus thus achieving a recall of 83% and precision of 80%. In [18] an abbreviation recognition system has been developed using the AB3P corpus. Thus, a recall of 95.86% and precision of 86.64% could be achieved. In [19] pattern-matching rules were developed for matching abbreviations with their respective full term. Thus, a recall of 70% and a precision of 95% could be obtained. In [20] a system was developed based on collocations yielding a recall of 88.5% and precision of 96.3%. In [21] a rule-based synonym recognition system was developed, in [22] a pattern matching system was developed to match abbreviations with their corresponding full names.

A lot of current research is interested in entity recognition and normalization [23]. In the BioCreative III competition, one task was focused on gene normalization, i.e. to identify and link genes to the standard database [24]. Such system has also been developed in [25]. Relationships between biomedical entities, e.g. protein-protein interactions, gene-disease interactions are investigated in [26].

Much work has been done in the field of relationship mining. For example, in [27] a relationship mining system was developed using MetaMap to identify biomedical entities [28] while using linguistic rules to determine the semantic relationships between them. In [29] a gene-disease relationship extraction system was developed from Medline abstracts using machine learning approach. It performed better than dictionary- and rule-based approaches.

The research in this work focuses on biomedical disease classification using the National Center for biotechnology (NCBI) corpus and applying combinations of machine learning approaches. We found that selecting rich features and combining classifiers contribute to a better performance.

II. DATASET DETAILS

Our dataset is the National Center for Biotechnology Information (NCBI) Disease Corpus. It is available at http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/. It consists of 793 abstracts containing 2783 sentences, 3224 unique disease names [30] and about 6,900 disease names in total. NCBI corpus annotators have annotated every sentence of the PubMed abstracts excluding organism names (e.g. human, virus and bacteria), gender (male and female), general terms (deficiencies and syndromes), biological references and nested disease. Annotations were done using a web base tool called PubTator [31]. The corpus annotations were assigned four categories based on the nature of the disease which consist of 3922 specific disease annotation, 1029 disease class annotations, 1774 modifiers and 173 composite mentions. The dataset is further divided into training, testing and development set as shown in the table below

Classes	Training set	Testing set	Development set
Modifiers	1292	264	218
Specific Disease	2959	556	409
Composite Mention	116	20	37
Disease Class	781	121	127

Table-1: Description of Train, test and Development

III. FEATURE SET

To improve classification accuracy, selecting and defining the features is very important. Enriching the feature set can improve the performance of a particular machine learning algorithm. To train our algorithm we used the following features:

1. Word Normalization
2. Orthographic
3. Part of Speech (POS) Tags
4. N-grams
5. Affixes
6. Contextual

Each of these 6 features is explained in more detail below:

A. *Word Normalization*

Word normalization attempts to reduce different form of words such as noun, adjective, verb etc. to its reduced/stemmed or root form . Common technique used for word normalization is the use of stemmer or lemmatizer, which stems word to its base form. Following are the various patterns analyzed which are reduced to its root form.

- Colorectal cancer → colorect cancer
- Endometrial cancer → endometri cancer
- Alzheimer disease → alzheim diseas
- Neurological disease → neurolog diseas
- Arthritis → arthriti
- Deficiency of DPD → defici of DPD
- Premenopausal ovarian cancer →premenopaus ovarian cancer
- Neurodegeneration → neurodegener
- Familial deficiency of the seventh component of complement → famili defici of the seventh compon of complement

B. *Orthographic Features*

Orthographic features are related to the geometry and indentation of the text such as capitalization, digits, numbers, numerics, single caps, all caps, two caps, punctuation, symbols etc. Such features are very effective in NER. Use of orthographic feature has been advocated in [32-34].

C. *Part Of Speech (POS) Tags*

Usually POS tags help define the boundaries of phrases. In some scenarios POS tags have improved NER performance [34-35]. Since POS tagging is a challenging and computationally demanding process some researchers have not used it in NER [36]. We have improved performance by including POS tags.

D. *N-grams*

N-grams are defined by a sequence of n tokens or words. The most common n-gram is unigram because it contains a single token. Other n-grams are bigrams and tri-grams containing 2 and 3 tokens respectively. Generally, N-grams are represented by the equation $P(W) = \prod_{i=1}^{|W+1|} P(w_i|w_0 \dots \dots .. w_{i-1})$ ------- (1).

From equation (1) $P(w_i|w_0 \dots \dots ... w_{i-1}) \approx P(w_i)$ which represents unigrams, while bigrams add one more word and can be represented as $P(w_i|w_0 \dots \dots ... w_{i-1}) \approx P(w_i|w_{i-1})$ and hence tri-grams adds two more words $P(w_i|w_0 \dots \dots ... w_{i-1}) \approx P(w_i|w_{i-2}w_{i-1})$ and hence other N-gram models can be found so on. In our experiment we only used bigrams and unigrams.

E. Affixes

Prefix and suffix features have significantly improved performance in the recognition of named entities. In [37] the authors have collected most frequent suffixes and prefixes from the training data, while in [38] the authors have grouped the prefixes and suffixes into 23 categories. In our experiment beside contextual features affixes has shown significant improvement.

F. Contextual features

Contextual features refer to the word preceding and following the named entities. Let w_0 be the current token i.e. named entity, so for each feature we use two token instances around it i.e. $c = (w_{-2}, w_{-1}, w_0, w_1, w_2)$. Now for each token w_0 which appears in the text at location $w_i, w_{i+1}, w_{i+2} \dots \dots \dots w_n$ the same features are calculated or more specifically c= $\prod_{i=-2}^{2} w_i$ …….. (2) Is the contextual window. In our experiment contextual features are the most important features in the recognition of NEs combined with affixes. Initially two contextual features followed by the current word were selected for the experiment. However, when realizing their importance four contextual features were selected. See equation 2, i.e. the two words preceding and the two words following the NE.

IV. CLASSIFICATION SCHEME

In this research Conditional Random Fields (CRF) was applied to the NCBI disease corpus. CRF is a probabilistic model for labeling sequential data; it's widely used for part of speech tagging and named entity recognition [39, 40]. CRF has several advantages over the HMM and SVM. CRF is based on a discriminative model. Hence, it includes a rich feature set containing overlapping features using conditional probability. Given a sequence $X = \{x_1, x_2, x_2, x_3, x_4 \dots . x_n\}$ and its labels $Y = \{x_1, x_2, x_2, x_3, x_4 \dots . x_n\}$, the conditional probability $P(Y|X)$ is defined by CRF as follows [41]:

$$P(Y|X) \propto \exp(\vec{w}^T \vec{f}\,(y_n, y_{n-1}, \vec{x})) \qquad (2)$$

$\vec{w}$ Is a weight vector defined by $\vec{w} = (w_1, w_2, w_3 \dots \dots . w_M)^T$ These weights are associated with features having length equal to M. $\vec{f}\,(y_n, y_{n-1}, \vec{x}) = \overrightarrow{f1}\,(y_n, y_{n-1}, \vec{x}), \overrightarrow{f2}\,(y_n, y_{n-1}, \vec{x}), \overrightarrow{f3}\,(y_n, y_{n-1}, \vec{x}) \dots \dots \dots \overrightarrow{fM}\,(y_n, y_{n-1}, \vec{x}))^T$
f is a feature function. Weight vectors (denoted by w) are obtained using the L-BFGS method [42]. In our experiment CRFSUITE has been used, which is the Python implementation of CRF [43].

V. RESULT AND DISCUSSION

Table-2 shows the contributions of features and their effects on the performance of CRF. The feature set is divided into Contextual (Cc), Normalized (Nm), Unigrams (Ug), bigrams (bg), Affixes (Ax), Part of speech (POS) and Orthographic (O). Performance evaluation was carried out using standard metrics such as precision, recall and F-score.

$$\text{Precision}= \frac{\text{Number of correctly classified Named entities}}{\textit{Total found Named entities}}$$

$$\text{Recall} = \frac{\text{Number of correctly classified Named entities}}{\textit{Total number of true Named entities}}$$

$$\text{F-score} = \frac{2(Precision*recall)}{(Precision+recall)}$$

Results obtained in Table-2 is based on applying 10 Fold cross validation on the training set.

Feature combination	***precision***	***recall***	***F-score***
O	0.54	0.62	0.53
O+ Nm	0.77	0.76	0.74
O+ Nm+ POS	0.87	0.87	0.86
O+ Nm + POS +Un	0.91	0.91	0.91
O+ Nm + POS + Un + Bg	0.92	0.92	0.91
O+ Nm+ POS +Un + Bg + Cc	0.92	0.92	0.92
O+ Nm +POS +Un + Bg +Cc + Affixes	0.94	0.94.	0.94

Table-2: Performance evaluation of Feature set.

Table-2 shows combinations of different features for improving CRF performance. Oorthographic features were taken as a benchmark. The benchmark performance was an F-score of 0.53, a precision of 0.54 and a recall of 0.62. Adding stemmed or normalized features improved the F-score to 0.74, the precision to 0.77 and the recall to 0.76. Adding part of speech tags further improved the F-score by 12 percent. Nevertheless, the part of speech tags were recently removed from the NER system. Unigram-based models have been the primary models in NER and hence we included them in our system. Adding the unigram features improved the F-score by 5%. Adding bigram-features did not raise the overall F-score but improved precision and recall by 1%. Adding contextual features only improved the F-score slightly by 1% but had no effect on precision and recall. Combining all features, i.e. orthographic, normalized, part of speech, unigram, bigram, contextual features and affixes yielded 94% for precision, recall and F-score. This performance was achieved with a 10-fold cross-validation on the training set due to the rich feature selection.

Figure 1 shows the F-scores for each of the 4 classes. In our experiment the following four classes were defined:

- Disease Class = DC
- Composite Mention = CM
- Specific Disease = SD
- Modifier = MD

The F-scores of the training, development and testing sets are plotted in figure 1. The best F-scores could be achieved for the Modifier class. For this class an F-score of 0.96 could be reached for the training dataset and for the development and testing dataset an F-score of 0.92 was obtained. The second highest F-scores could be achieved for the Specific Disease class. For this class the F-score of the training dataset was 0.95, for the testing set it was 0.92 and for the development set it was 0.88. The third highest F-scores were achieved for the Disease Class. For this class the F-score for the training set was 0.86 and the F-scores for the testing and development set were both 0.71. The F-scores were lowest for the Composite Mention class. For this class the F-score for the training set was 0.72, for the testing set it was 0.52 and for the development set it was 0.62. We observed a positive correlation between the size of the training sample sets and F-score. The largest training sample comprising of over 1,000 was available for the Modifier class, followed by the Special Disease class, followed by the Disease Class having the second smallest training sample followed by the Composite Mention class, which had the smallest training sample. The performance of machine learning algorithms depends on the size of the training sample. Too small training samples increase the risk of under fitting while too large training samples increase the risk for over fitting.

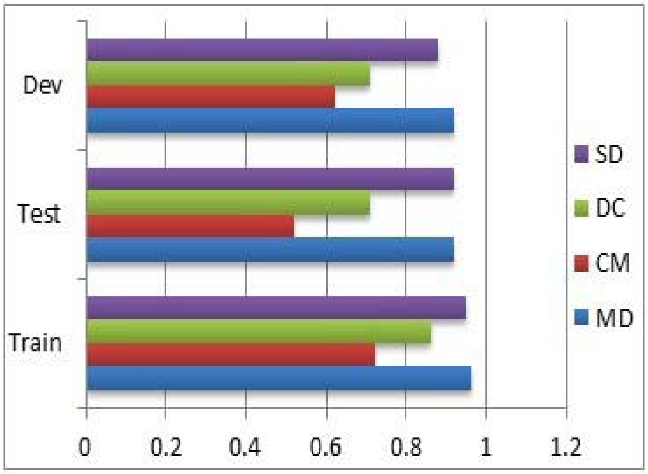

Figure-1: F-score Comparision of Training, Testing and Development Data sets.

We compared the performance of our approach, which is based on combining features with that of BANNER using the same dataset and classes. The results of this comparison are shown in table 3. Details about BANNER results can be found in [30]. The data in table 3 indicates that our approach yielded much higher F-scores than BANNAR for the training, testing and development set. The F-score obtained with our approach is 10% higher for the training set, 7% higher for the testing set and 4% higher for the development set. Hence, we clearly succeeded in outperforming BANNER.

System	Dataset	Precision	Recall	F-Measure
CRF Result	Training	0.94	0.94	0.94
	Testing	0.88	0.89	0.88
	Development	0.86	0.86	0.85
BANNER Result	Training	0.86	0.82	0.84
	Testing	0.83	0.80	0.81
	Development	0.82	0.81	0.81

Table-3: Comparison of BANNER and CRF results: For both Classifiers Precision, Recall and F-score are reported.

Figure 2 also shows that our F-scores (depicted in blue) are much higher than those of BANNER (depicted in red)

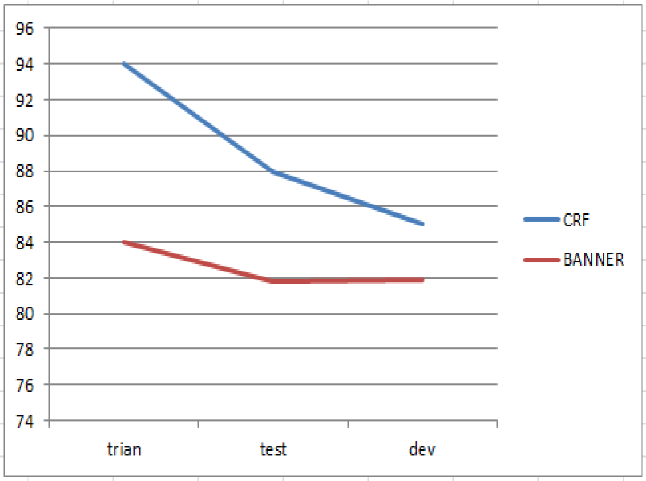

Figure-2: Plot of BANNER Vs Proposed Model

In summary it can be concluded that CRF based on 6 features clearly outperformed BANNER. This clearly shows that the sequential classifier CRF is well suited for classifying biomedical literature based on rich features.

VI. Conclusion

This paper presents a machine learning approach for human disease named entity recognition using the NCBI disease corpus. The system takes the advantage of background knowledge obtained from the selected features to better distinguish between the four classes. Improvements due to feature additions have been demonstrated. The highest improvement could be obtained when adding a second feature to the first. However, in order to evaluate the overall benefit for each feature, all possible combinations of feature additions need to be considered.

References

[1]. A.M. Cohen, W.R. Hersh A survey of current work in biomedical text mining Brief Bioinform, 6 (2005), pp. 57–71

[2]. L. Li, R. Zhou, D. Huang,Two-phase biomedical named entity recognition using CRFs. Comput Biol Chem, 33 (2009), pp. 334–338

[3]. D. Rebholz-Schuhmann, A.J. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, *et al.* Assessment of NER solutions against the first and second CALBC Silver Standard Corpus.J Biomed Semantics, 2 (Suppl. 5) (2011)

[4]. M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-Aryamontri, A. Winter, *et al.*The Protein–Protein Interaction tasks of BioCreative III: Classification/ranking of articles and linking bio-ontology concepts to full text.BMC Bioinformatics, 12 (Suppl. 8) (2011)

[5]. M.S. Habib, J. Kalita, Scalable biomedical Named Entity Recognition: investigation of a database-supported SVM approach.Int J Bioinform Res Appl, 6 (2010), pp. 191–208

[6]. S.K. Saha, S. Sarkar, P. Mitra. Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inform, 42 (2009), pp. 905–911

[7]. Y.M.N. Ephraim.Hidden Markov processes.IEEE Trans Inform Theory, 48 (2002), pp. 1518–1569

[8]. He Y, Kayaalp M. Biological entity recognition with conditional random fields. In: AMIA annu symp proc; 2008. p. 293–7.

[9]. Zhou GD, Su J. Exploring deep knowledge resources in biomedical name recognition. In: JNLPBA; 2004. p. 96–99

[10]. Kazama J, Makino T, Ohta Y, Tsujii J. Tuning support vector machines for biomedical named entity recognition. In: Association for computational linguistics Morristown, NJ, USA; 2002. p. 1–8.

[11]. T. Tsai, W.C. Chou, S.H. Wu, T.Y. Sung, J. Hsiang, W.L. Hsu,Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. Expert Syst Appl, 30 (2006), pp. 117–128

[12]. Lin YF, Tsai TH, Chou WC, Wu KP, Sung TY, Hsu WL. A maximum entropy approach to biomedical named entity recognition. In: The 4th ACM SIGKDD workshop on data mining in bioinformatics; 2004. p. 56–61.

[13]. C.R. Yen-Ching, Tsai Tzong-Han, Hsu Wen-Lian. New challenges for biological text-mining in the next decade. J Comput Sci Technol, 25 (2010), pp. 169–179

[14]. Y. Sasaki, Y. Tsuruoka, J. McNaught, S. Ananiadou. How to make the most of NE dictionaries in statistical NER.BMC Bioinformatics, 9 (Suppl. 11) (2008), p. S5

[15]. Zhou GDaJS. Exploring deep knowledge resources in biomedical name recognition. In: JNLPBA; 2004.

[16]. B.S. Fei Zhu. Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing. PLoS One, 7 (6) (2012), p. e39230

[17]. J.T. Chang, H. Schutze, R.B. Altman. Creating an online dictionary of abbreviations from MEDLINE.J Am Med Inform Assoc, 9 (2002), pp. 612–620

[18]. C.J. Kuo, M.H. Ling, K.T. Lin, C.N. Hsu. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. BMC Bioinformatics, 10 (Suppl. 15) (2009), p. S7

[19]. H. Yu, G. Hripcsak, C. Friedman.Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc, 9 (2002), pp. 262–272

[20]. H. Liu, C. Friedman. Mining terminological knowledge in large biomedical corpora. Pac Symp Biocomput (2003), pp. 415–426

[21]. J. McCrae, N. Collier. Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC Bioinformatics, 9 (2008), p. 159

[22]. A.M. Cohen, W.R. Hersh, C. Dubay, K. Spackman. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC Bioinformatics, 6 (2005), p. 103

[23]. H.-Y.K. Zhiyong Lu, Wei Chih-Hsuan, Huang Minlie, Liu Jingchen, Kuo Cheng-Ju, Hsu Chun-Nan, *et al.*The gene normalization task in BioCreative III.BMC Bioinformatics, 12 (2011)

[24]. C.N. Arighi, P.M. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, A. Chatr-Aryamontri, *et al.*

[25]. BioCreative III interactive task: an overview. BMC Bioinformatics, 12 (Suppl. 8) (2011),

[26]. M. Huang, J. Liu, X. Zhu. GeneTUKit: a software for document-level gene normalization. Bioinformatics, 27 (2011), pp. 1032–1033

[27]. C.N. Arighi, Z. Lu, M. Krallinger, K.B. Cohen, W.J. Wilbur, A. Valencia, *et al.*Overview of the BioCreative III workshop. BMC Bioinformatics, 12 (Suppl. 8) (2011), p. S1

[28]. Ben Abacha, P. Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomed Semantics, 2 (Suppl. 5) (2011), p. S4

[29]. A.R. Aronson, F.M. Lang. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc, 17 (2010), pp. 229–236

[30]. Rezarta Islamaj, Dogan Zhiyong Lu. An improved corpus for disease mentioned in Pubmed citatations Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 91–99, Montr´eal, Canada, June 8, 2012

[31]. Leaman, R.,Miller,C.Gonzalez. enabling recognition of disease in biomedical text with machine learning: corpus and Benchmarks. Symposium on languages in biology and medicine 2009. Pg 82-89.

[32]. Wei.C, Kao.H,Lu.Z. 'Pubtator: A Pubmed-like interactive curation system for document triage and literature curation. In procedings of BioCreative workshop 2012 pg145-150.

[33]. N. Collier, K. Takeuchi. Comparison of character-level and part of speech features for name recognition in biomedical texts. J Biom. Inform. 37. pp423-435. 2004.

[34]. D. Shen, J. Zhang, G. Zhou, S. Jian and L. Tan, Effective Adaptation of a Hidden Markov Modelbased Named Entity Recognizer for Biomedical Domain, In: Proceedings of ACL 2003 Workshop on NLP in Biomedicine, Sapporo, Japan, pp4956, 2003.

[35]. Tsai, T.-H., Wu, S.-H., & Hsu, W.-L. (2005). Exploitation of linguistic features using a CRFbased biomedical named entity recognizer. to appear in ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Detroit

[36]. L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In CoNLL, 6.

[37]. J. Kazama, T. Makino, Y. Ohta, J. Tsujii. Tuning Support Vector Machines for Biomedical Named Entity Recognition. In: Proceedings of Workshop on NLP in the Biomedical Domain, ACL 2002. pp1-8. 2002.

[38]. G. Zhou and J. Su. Named Entity Recognition using an HMM-based Chunk Tagger. In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 473-480 2002.

[39]. Huang H-S, Lin Y-S, Lin K-T, Kuo C-J, Chang Y-M, Yang B-H, Chung I-F, Hsu C-N: High-recall gene mention recognition by unification of multiple background parsing models. Proceedings of the 2nd BioCreative Challenge Evaluation Workshop 2007, 23:109-111.

[40]. Klinger R, Friedrich CM, Fluck J, Hofmann-Apitius M: Named entity recognition with combinations of conditional random fields. In Proceedings of the 2nd BioCreative Challenge Evaluation Workshop

[41]. Porter M.F. "Snowball: A language for stemming algorithms". 2001.

[42]. Ms. Anjali Ganesh Jivani "A Comparative Study of Stemming Algorithms" Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938

[43]. Dekang Lin and Xiaoyun Wu. 2009. Phrase Clustering for Discriminative Learning. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1030–1038, Suntec, Singapore, August. Association for Computational Linguistics.