Disease Named Entity Recognition Using Conditional Random Fields

Hidayat Ur Rahman
Lahore Leads University
Near Garden Town Kalma Chowk
Lahore Pakistan
Hidayat.R@gmail.com

Thomas Hahn
University of Arkansas
South University Avenue
Little Rock,AR,72204
Thomas.F@gmail.com

Richard Segall
Arkansas State University
Computer Inform Tech Department
State University,AR 72404-0130
rsegall@astate.edu

Abstract

Named Entity Recognition is a crucial component in bio-medical text mining.In this paper a method for disease Named Entity Recognition is proposed which utilizes sentence and token level features based on Conditional Random Field's using NCBI disease corpus. The feature set used for the experiment includes orthographic,contextual,affixes,n-grams,part of speech tags and word normalization.Using these features,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,F-score of 88% and 85% were reported.

1 Introduction

The increasing amount of bio-medical literature requires more robust approaches for information retrieval and knowledge discovery because every single day more information is published than humans can read. Unique challenges specifically to bio-medical Named Entity Recognition (NER) are caused due to its structure,since biomedical Named Entities(NEs) consist of symbols and abbreviations to infer relationships, thus the length of Bio-medical NEs are not consistent,which is the primary reason why Bio-NER have low performance compared to general purpose NER (Lishuang, L.and W. Fan, 2013).Bio-NER is the most important step in the extraction of knowledge, which has the overall aim of identifying specific concepts or categories, such as gene, protein, disease,drug, etc. Current trend in NER is based on machine learning (ML) approaches, ML based approach provides the flexibility of statistical and rule-based techniques. However, the performance of machine learning techniques highly depends on the availability of sufficient training data in order to adequately train the machine learning classifiers (M. Krallinger et al, 2011).In this article Bio-NER for disease names has been carried out to handle the challenges of boundary detection and entity classification using Conditional Random Fields(CRF).The model consists of an enriched set of features including boundary detection features, such as word normalization, affixes, orthographic and part of speech(POS) features. For the semantic labeling features, such as n-grams and contextual features have been used.

2 Methodology

For disease NER our methodology follows the traditional machine leaning approach. Figure. I depicts the work-flow of our methodology. Firstly, raw text is obtained from training, testing and development set,then pre-processing is carried out to remove characters and symbols such as underscore character, full stop etc. After pre-processing various features are extracted as described in section 2.1. The features are fed into a sequential CRF as described in section 2.2. Thus, structured output in the form of annotated named entities is obtained. This section provides details about feature extraction and classification.

2.1 Feature Set

Feature extraction plays a vital role in the classification accuracy of machine learning classifier as well as the NER system. The selection of relevant feature set improves the classification performance of Bio-NER. Table.1 shows the list of features used for Bio-NER and their short descriptions is listed below

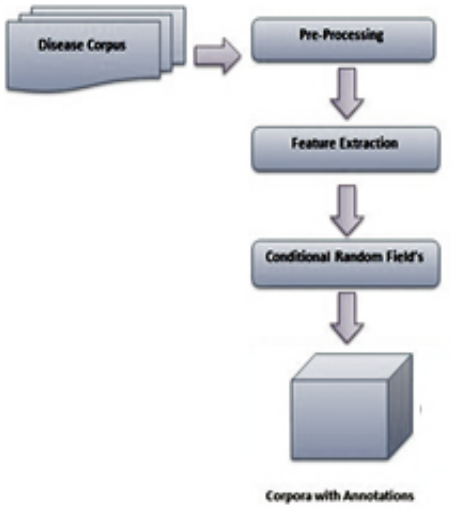

Figure 1: Flow Chart of Proposed System)

Word Normalization

Word normalization attempts to reduce different forms of words, such as nouns, adjectives,verbs, etc. to their root form . For word normalization, Porter stemmer has been used to reduce disease names to its root form. Below are few examples of disease names for word root reductions obtained with the Porter stemmer algorithm.

- Colorectal cancer – colorect cancer
- Endometrial cancer – endometri cancer
- Alzheimer disease – alzheim diseas
- Neurological disease – neurolog diseas
- Arthritis – arthriti

Orthographic Features

Orthographic features are related to the orthography of the text, such as Capitalization, Digits, Numeric, Single Caps, All Caps,numerics and punctuation. Such features are very effective in boundary detection (Collier, Nigel and K. Takeuchi, 2004). The eleven orthographic features below have been used in our model:

- IDASH: Whether a token/word contains an inner dash such as A-T, G6PD-deficient, Palizaeus-Merzbacher disease.
- 2IDAH: If the number of IDASH counts equals to 2 e.g. X-linked Emery-Dreifuss muscular dystrophy, Borjeson-Forssman-Lehmann,

Feature	Description
word normalization	Stemmed form of the Named Entity
Contextual features	$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}$
POS	$Pos_{w-2}, Pos_{w-1}, Pos_w$ Pos_{w+1}, Pos_{w+2}
Orthographic features	Uppercase,lowercase,title, hyphen,Alphanumeric etc
Word N-grams	$w_i -_2 /w_{i-1}, w_i -_1 /w_i, w_i/w_{i+1}, w_i +_1 /w_{i+2}$
POS	$POS_{w-2}, POS_{w-1}, POS_{wi},$
N-grams	POS_{w+1}, POS_{w+2}
Prefix	PREFIX(w_i)
Suffix	SUFFIX(w_i)

Table 1: Feature set for named entity recognition

- ALLCAPS: Is set to true if all the alphabets in a given token are capital examples includes DMD, BMD, FD, APC, FAP and HDD etc.
- TITLE: If the first alphabet in a token is capitalized such as Alzheimer disease, Huntington disease, Combined genetic deficiency of C6 and C7.
- LOW: All the alphabets in a given word are in lower case e.g. myotonic dystrophy, idiopathic dilated cardiomyopathy, and facial lesions.
- MIXED: If a given sequence of words contains both upper and lower case such as DMD defects, hypo myelination of the PNS, deficiency of active AVP.
- ALPNUM: If a given words contains both numeric and text like abnormality of CYP27, C6 deficiency, achondrogensis 1B, abnormality of CYP27.
- PARN: If a multi-word contains parenthesis such as Arginine vasopressin (AVP) deficiency, palmoplanter keratoderma (PPK) conditions, sporadic (nonhereditary) ovarian cancers

- BRACKS: Bracket is contained within a token,example includes hypoxanthine phosphoribosyl transferees [HPRT] deficiency.

- GREEKS: Greek letters such as I,II,III,IV etc,is contained within a token e.g. type IIA vWD,Type II ALD,type II Gaucher disease,type II GD and type III GD

- SLASH: Character / is contained within the multi-word token such as cleft lip/palate, CL/P, breast and/ovarian cancers, glucose/galactose malabsorption.

Ngrams

N-grams are defined by a sequence of n-tokens or words. The most common n-gram is uni-gram which,contains a single token.Other n-grams examples are bi-grams and tri-grams containing 2 and 3 tokens respectively. In this experiment unigram and bi-gram have been used,in this method all the digits within a word are replaced with d e.g,the uni-gram of 33 is dd, uni-gram of nt943 is ntddd.Bigram examples are ALD/Eighteen, skin tumor/caused, APC/protein, breast or ovarian cancer/novel, etc.

Part of Speech(POS) tags

POS tags are helpful in defining boundary of a phrase,inclusion of POS has been advocated by (J. Kazama and T. Makino, 2002).Our experiment includes POS tags of contextual features and bigrams. Adding POS tags to our feature set, the performance of the classifier is boosted as shown in Table.3

Affixes

Prefix and suffix feature has shown better performance in the recognition of NEs in this experiment.In (J. Kazama and T. Makino, 2002) the authors collected most frequent suffixes and prefixes from the training data. Prefix and suffix are n character in length at the beginning and end of a token respectively (Zhou, G. Dong, and J. Sui, 2002). In our model all the combinations n=1 through 4 have been used to boost performance.The prefix for the word "tumour" are t, tu, tum and tumo, the suffix for the same word are r,ur,our and mour.Beside contextual features,affixes yielded improvements in the overall performance as shown in Table.3.

Contextual features

Contextual features refer to the word preceding and following the NEs. Contextual features are the most important features in this experiment for semantic labeling of disease names. In our experiment four contextual features are selected. Two words preceded and two followed the named entities. E.g. for the term bactracin in colon carcinoma loss cells, colon carcinoma represents the named entities while bactracin in and loss cells represent the two preceding word and two following word.

2.2 Conditional Random Field's (CRF)

CRF is a probabilistic model used for labeling sequential data. It is widely used for POS tagging and NER. (Huang H-S and Lin Y-S, 2007). CRF has several advantages over the Hidden Markov Model (HMM) and Support Vector Machine (SVM). CRF includes rich feature sets,i.e. overlapping features using conditional probability. For example, given a sequence $X = x_1, x_2, x_3, x_4.....x_n$ and its labels $Y = y_1, y_2, y_3, y_4.....y_n$, the conditional probability $P(Y \mid X)$ is defined by CRF as follows $P(Y \mid X) \propto exp(w^T f(y_n, Y_{n-1}, x))$ (Sutton, C. and McCallum, 2011).W is a weight vector defined by $w = (w_1, w_2, w_3.w_M)^T$. Theses weight are associated with features having length equal to $M.f(y_n, y_{(n-1)}), x) = f(y_n, y_{(n-1)}), x), f2(y_n, y_{(n-1)}), x)$, $f3(y_n, y_{(n-1)}), x)...f_M = (y_n, y_(n-1)), x))^T$ is a feature function.The weight vector is obtained using L-BFGS method.In our experiment CRFSUITE has been used which is the python Application programming interface(API) of CRF++.

3 Experimental Setup

3.1 Dataset

Our experiment is based on National Center for Biotechnology Information (NCBI) disease corpus, which is freely available at NCBI website(`http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/`) NCBI corpus includes 793 abstracts, which consist of 2783 sentences and a total of 6900 disease names (Dogan, R. and Islamaj, 2012). Annotations of disease names are based on the criteria that,disease mentions which describes a family of specific diseases are annotated as

disease class e.g. autosomal recessive disease, whereas text referring to specific disease are annotated as specific disease, such as Diastrophic dysplasia.Strings referring to more than one disease names are annotated as composite mention, e.g. Duchene and Becker muscular dystrophy are two disease mentions and hence it is categorized as composite mention. Certain disease mentions are used as modifier for other concepts, e.g. a string may denote a disease name but it is not a noun phrase and hence it is annotated as modifier, e.g. colorectal cancer.Table 2 shows the distribution of disease names in training,testing and development set.

Classes	**Train set**	**Test set**	**Dev set**
Modifiers	1292	264	218
Specific Disease	2959	556	409
Composite Mention	116	20	37
Disease Class	781	121	127

Table 2: Dataset used in experiment

3.2 Classification and Feature Selection

Table.3 shows contribution of features and its effect on performance of CRF. The feature set is mainly divided into Contextual(Cc),Normalized(Nm),Ngrams,Affixes(Ax),Part of speech (POS) and Orthographic(O). Performance evaluation has been carried out using the metrics precision, recall and F-score. Results obtained in Table.3 is based on applying 10 Fold cross validation on the training set. Orthographic features were taken as a benchmarks, which results in F-score of 0.53. This is considered as the lowest F-score reported in this experiment. Addition of normalized features resulted in an increasing the F-score by 21%.Further addition of POS tags increased the performance by 12%.With the addition of N-gram features the overall F-score achieved is 0.91.Finally,with the addition of affixes,the final F-score obtained is 0.94.Compared to other state of the art Bio-NER systems,such as BANNER,our system has a higher level of F-score using 10 fold cross validation on training set due to the selection of good features for disease NER.

Features	**p**	**r**	**f**
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+ Ngram	0.92	0.92	0.91
O+Nm+POS+ Ngram+Cc	0.92	0.92	0.92
O+Nm+POS+ Ngram+Cc+Affixes	0.94	0.94.	0.94

Table 3: Combination of different features

4 Result and Discussion

For result visualization we have plotted the f-score of individual classes. In Figure.2 the F-score of individual dataset has been plotted. In Figure.2 DC denotes Disease Class,CM denotes Composite Mention,SD denotes Specific Disease and MD denotes Modifier.Figure II shows that the highest F-score have been reported by Modifier for Training,Testing and Development set respectively,followed by Specific Disease. The lowest F-score has been shown by Composite Mentions followed by Disease Class.One reason for the relatively poor performance of the Composite Mention is the inadequate training samples compared to the training samples of Specific disease and Modifiers,which exceed 1000.The relatively poor performance of the Disease Class is because it has been based on the second smallest training sample since, the performance of machine learning based techniques heavily depends on the number of training samples.

5 Conclusion

This paper presents a machine learning approach for human disease NER using NCBI disease corpus. The system takes the advantage of rich feature set which,helps in representation and distinguishing of related concepts and categories.Simple features including orthographic, contextual, affixes, bigrams, part of speech and normalized tokens without exploiting features such as head nouns, dictionaries etc.The model has achieved state of the art performance for semantic labeling of named entities using the NCBI disease

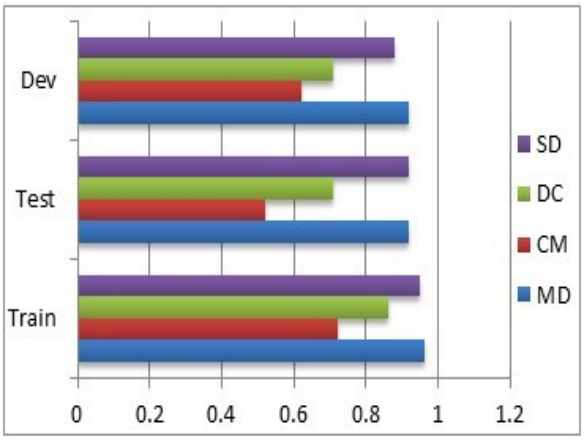

Figure 2: f-score comparison of training,testing and development set

corpus.Each feature set represent some knowledge about the named entity and hence, in order to evaluate the overall benefit for each feature, all possible combinations of feature additions need to be considered.

References

Collier, Nigel, and K. Takeuchi. 2004. *Comparison of character-level and part of speech features for name recognition in biomedical texts*, volume 36. Journal of Biomedical Informatics.

Ratinov, Lev, and D. Roth. 2009. *Design challenges and misconceptions in named entity recognition*. Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics.

J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. 2002. *Tuning Support Vector Machines for Biomedical Named Entity Recognition*. Proceedings of Workshop on NLP in the Biomedical Domain. 1–8

Zhou, G. Dong, and J. Sui. 2002. *Named entity recognition using an HMM-based chunk tagger* proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. 473–480

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

Klinger R, Friedrich CM, Fluck J, Hofmann-Apitius. 2007. *Named entity recognition with combinations of conditional random fields*. Proceedings of the 2nd BioCreative Challenge Evaluation Workshop.

Sutton, C. and McCallum. 2011. *An introduction to conditional random fields*. Foundations and Trends in Machine Learning.267–373

Dogan, R. Islamaj, and Z. Lu. 2012. *An improved corpus of disease mentions in PubMed citations*. Proceedings of the 2012 workshop on biomedical natural language processing. Association for Computational Linguistics.91–99

L. Lishuang, W. Fan, and D. Huang. 2013. *A Two-Phase Bio-NER System Based on Integrated Classifiers and Multiagent Strategy.*. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 10.4.897-904

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