Named Entity Recognition in Biomedical Domain: A Survey

T. M. Thiyagu
Research Scholar
Department of Computer Science & Engineering
College of Engineering, Guindy
Anna University, Chennai

D. Manjula, PhD
Professor
Department of Computer Science & Engineering
College of Engineering, Guindy
Anna University, Chennai

Shruthi Shridhar
Student
Department of Computer Science & Engineering
College of Engineering, Guindy
Anna University, Chennai

ABSTRACT
Named Entity Recognition plays an important role in locating and classifying atomic elements into predefined categories such as person names, locations, organizations, expression of times, temporal expressions etc. Named entity recognition is also called as entity chunking, entity identification and entity extraction. It is a subtask of information extraction, where the structured text is extracted from unstructured text. Named Entity Recognition (NER) is one of the major tasks in Natural Language Processing (NLP). NER has been an active area of research for the past twenty years. An ability to automatically perform NER i.e., identify occurrences of NE in Web contents can have multiple benefits, such as improving the expressiveness of queries and also improving the quality of the search results. A number of factors make building highly accurate NER a challenging task. Though a lot of progress has been made in detecting named entities, NER still remains a big problem at large. In this paper, we explore various methods that are applied to solve NER in the biomedical domain.

Keywords
Arabic NER, Named Entity Recognition, Information Extraction, NER tools.

1. INTRODUCTION
In the world of exponentially growing web data, information is rapidly increasing. When user requests for any information, the response to the request is in an unstructured form. The conversion of unstructured information to structured information is called Information Extraction. The goal of the information extraction system is to find and understand limited relevant parts text; normally we are going to run it into a context where system gathers information from across many pieces of text. The goal of doing this information gathering is to produce some sort of structural representation to the relevant information. Subtasks of IE are named entity recognition, noun phrase co-reference resolution, semantic role recognition, entity relation recognition, and date and timeline recognition. Named Entity Recognition is a process where all the named entities are identified the proper nouns and classified into their pre-defined appropriate class. Very important applications of Named Entity Recognition (NER) are Information Retrieval, Information Extraction, Question Answering, Machine Translation, POS Tagging, Text Mining and Automatic Summarization etc. Various approaches are imposed for NER in various languages. These approaches mainly involve rule-based, list lookup and machine learning.This paper is organized into 4 sections. Section 1 presents the introduction. Section 2 gives the introduction about NER; section 3 presents the various NER approaches in biomedical domain; Section 4 concludes the paper and provides overall discussion.

2. NAMED ENTITY RECOGNITION
Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Named entity recognition (NER) or tagging is the task of finding names such as organizations, persons, locations, etc. in the text. Since whether or not a word is a name and the entity type of a name are determined mostly by the context of the word as well as by the entity type of its neighbours, NER is often posed as a sequence classification problem and solved by methods such as hidden Markov models (HMM) and conditional random fields (CRF). In NER the aim is to distinguish between character sequence that represents noun phrases and character sequence that represents normal text. Named entities are the pronouns which are the names of users for the referring persons, organizations, locations, etc. these named entities are classified into different categories as shown in Fig 1.

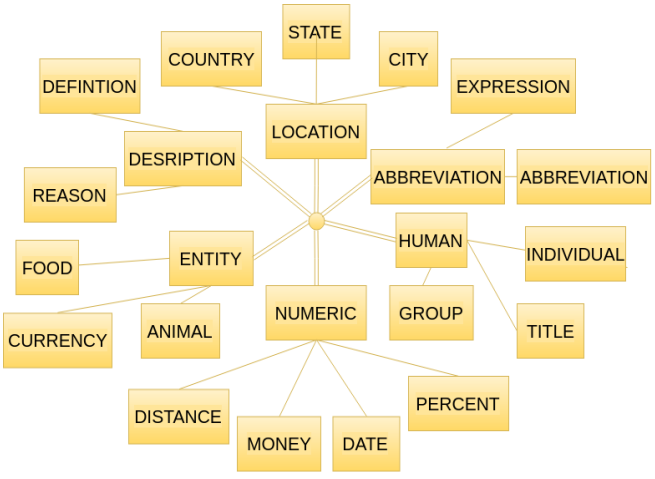

Fig. 1. A single Named Entity split into more specific Named Entities

In Biomedical domain, as one of the fundamental biomedical text mining tasks, Named Entity Recognition (NER), aims at identifying chunks of text referring to specific entities of interest. It plays a key role in disease-treatment relation extraction gene function identification and semantic relation extraction between concepts in molecular biology. Recently, several attempts have been performed to transform existing named entity recognition systems in general domain into the biomedical area. However, due to the non-standard nomenclature in biomedical research, few of them achieved satisfactory performance, thus biomedical named entity

recognition (Bio-NER) continues to be a challenging task.

3. OVERVIEW OF NAMED ENTITY RECOGNITION APPROACHES

There are a number of NER approaches that are already in use today in the biomedical domain. We list approaches here to compare them and conduct a literature survey.

3.1 Named Entity Recognition: Fallacies, Challenges and Opportunities

In this paper Marrero et al. argue that NER is in fact not a solved problem, and show how the lack of agreement around the concept of NE has important implications for NER tools and, especially, for their evaluation. Current evaluation forums related to NER do not solve this problem basically because they deal with very different tasks. The NER task has to define an adequate and meaningful set of the semantics of interest, according to the user requirements, and a collection of documents representative as much as possible of the ones expected in real settings. General purpose NER tools reflect this situation, moving from recognizing a few types of entities from the journalistic (people, organizations, organizations, dates) or military domains (vehicles, weapons), to new categories such as food products and names of events like hurricanes as in the tool YooName.

The approach traditionally followed consists in the use of very large corpora and the creation of different query sets from year to year, allowing researchers to compare their systems over time with different collections rather than just one. In the traditional NER task there is no such thing as a ranking of entities or a perfect result, so using these measures leads us to a problem of construct validity because the results of the evaluation experiments do not measure what we intend to. Therefore, the new NER-related evaluation forums cannot be considered just a modern version of MUC, CoNLL and ACE, because they deal with very different tasks.

NER has been considered a solved problem when the techniques achieved a minimum performance with a handful of NE types, document genre and usually in the journalistic domain. Instead, evaluation forums should focus on specific applications that allow researchers to focus on specific user needs and progressively improve NER techniques. This would allow us to measure the actual domain-specific state of the art and, at some point, have at our disposal a sufficiently large and generic evaluation framework to assess whether the generic NER task is a solved problem or not. Finally, it is necessary to reconsider the effort required to adopt a tool to a new type of entity or collection, as it usually implies the annotation of a new document collection. The recurrent use of supervised machine learning techniques during the last decade contributed to making these tools portable but at the expense of significant annotation efforts on behalf of the end user. The real applicability of the former largely depends on the resources available to the end user, and therefore its evaluation should contemplate how much effort is required on her behalf to adapt the tool to her needs.

3.2 Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts

Biomedical named-entity recognition (BM-NER) is used for biomedical language processing. We can obtain clinical information from clinical reports and drug names from the discharge summaries. Named-entity recognition (NER) is used to identify the names and locations from news and tweets. An unsupervised approach to the whole problem is presented by Zhang et al. The results from the experiments are used for demonstrating the results obtained from datasets of different genres. Sequence labelling models such as Hidden Markov Model (HMM) and Conditional Random Fields (CRF) are used to model transitions between the specified labels. The similarity-based method is primarily used in word sense disambiguation (WSD), taking into account that the meaning of a given word is closely related to the varied distribution of each and every word around it.

Zhang et al suggest that the exploitation of IDF can be further improved. In the previous approaches, the IDF value is obtained by taking the average values of all the words in the phrase. A longer sequence with many common words is a huge disadvantage. The above experiment is based on the Biomedical Named Entity Recognition (BM-NER) using seed term extractor, an NP chunker and an IDF filter. The best accuracy obtained is above 50%. The classification of entities shows very good results for all the entity classes provided in the datasets. The paper concludes that including nested NPs and better chunker and improved IDF values would improve the accuracy of the algorithm.

3.3 NCBI disease corpus: A resource for disease name recognition and concept normalization

Automatic disease recognition is a highly difficult task because there may be a dis-ambiguity in the name of a disease and the same abbreviation may stand for two different diseases. The current idea for disease recognition makes use of disease mention recognition and disease concept recognition. A PubMed article is used as a reference for the name of the diseases. Dogan et al. have developed a disease corpus, the NCBI disease corpus. The NCBI disease corpus is manually annotated for the mention of every disease name as in the abstracts of the PubMed articles.

The inference method is the first work in the process of disease normalization. This method was initially based on the combination of several string matching rules that mapped the annotated string to the names of the diseases. On an average, the NCBI disease corpus contains 5.08 disease mentions and 3.28 disease concepts. The number of articles having a single mention of the article is 5%. An error analysis was performed to help illustrate the relative strengths and weaknesses of the different methods employed. The methods experienced both false negatives and false positives due to term variations not present in the lexicon. This error was subsequently ignored by MetaMap.

The inference method produced an accuracy of 79% and showed that this method linked disease mentions to their corresponding medical vocabulary entry with high precision. The NCBI disease corpus, a richly annotated corpus with disease names and their corresponding MeSH and/or OMIM identifiers was introduced in this paper. Experiments demonstrated the feasibility of using the corpus as the basis for training learning models in both named entity identification and concept recognition.

3.4 Supervised Methods for Symptom Name Recognition in Free-Text Clinical Records of Traditional Chinese medicine: An empirical study

A pragmatic approach to the Chinese word segmentation was proposed by Zhonghua Yu et al. in this paper. A hybrid

Chinese NER model based on multiple features was proposed and the model was evaluated on the general text dataset. NER in the TCM community was attempted initially based on the method used for English NER in discharge summaries of Western medicine.

The highest FM_{rec} (95.12%) of SNR in chief complaints reported in this paper is much higher than the best Fmrec of NER in English discharge summaries which are reported to be 85.20%. These characteristics facilitate the SNR task as performed by sequence classifiers. SNR is one of the chide task for exploiting the content of FCRs of Traditional Chinese Medicine (TCM) and constructing clinical expert systems for TCM. It provides an opportunity to effectively and efficiently make abundant use of the already available knowledge. The general sequence labelling strategy is appropriately adapted for the SNR task for several empirical reasons.

3.5 Tmchem: a High-Performance Approach for Chemical Named Entity Recognition and Normalization

Chemical names have systematic and semi-systematic methods for describing the chemical structure. Leaman et al. use a model combination approach where the models are very different. They use different tokenization's, feature-sets, CRF implementation, CRF parameters and variations of post-processing.

Model 1 had a slight drop in performance between the evaluation and test sets. On the other hand, model 2 had slightly better performance. Model 2 was found to have the highest f-measure reported on the CEM task. The model combination provided a high recall and hence identified the approach of preparing and combining the multiple CRF models along with the post processing to increase the overall effectiveness. But it was also found that combining the results for various models was impractical and the large size of model 2 hinders the use of it.

The future work lies in the improvement in the performance which can be concluded from our error analysis. Like gene and other concept recognition tasks, it is important to investigate how to normalize detected chemical mentions to standard terminologies or ontologies in future studies.

3.6 Drug Name Recognition: Approaches and Resources

In this paper, Liu et al. concentrates on the review of various DNR studies which includes the aspects like challenges in DNR, approaches used for DNR and the future improvements that could be introduced in this field. One of the main elements that constitute the medical information are drugs and it is crucial that we perform a drug name recognition (DNR). In DNR, we recognize the drug names present in the unstructured medical text and classify them into predefined categories.

The few challenges faced in DNR arises due to the fact the DNR is a named entity recognition. Firstly, due to the various naming methodologies used to name the drug, such as using abbreviations and acronyms, mixed symbols and common words. Secondly, due to the rapid rise of new drugs every day. And finally due to the misspelling of the drug names in various medical records.

The typical procedure for the DNR system consists of three steps, namely - preprocessing, DNR and post-processing. In the preprocessing stage, the aim is to transform the original input texts into representations required by the DNR approaches, enriching the lexical and syntactic information by using NLP tools such as openNLP, GENIA tagger etc., Fine-grained tokenization was used as it was better compared to coarse-grained tokenization as the fine-grained tokenization applied some extra processing over the coarse-grained tokenization. The second stage includes the DNR. In this stage, the knowledge of the drugs is essential to classify the unstructured data that is present. Then finally is the post-processing stage where the DNR results are refined using the heuristic rules and knowledge resources as well.

There are various approaches to the drug name recognition which include dictionary-based approaches, rule-based approaches, machine learning-based approaches as well as hybrid approaches. The dictionary-based approach consists of a dictionary of drug names that consist mainly of publicly available knowledge resources and or synonyms of various drug names. In this approach, the drug names are identified by matching the drug dictionaries against the texts. The drawback about this approach is the yield of low precision due to poor quality in drug dictionaries.

The rule-based approach uses rules to describe the composition patterns or context of drug names. The drawback with this approach lies in the generation of rules which is very time-consuming. Also, the rules designed for one class of drugs are not applicable to the other classes.

The machine learning-based approach uses classification and labelling of data. The most popular tag is BIO for DNR. BIO stands for beginning, inside and outside of the drug. The selection of the machine learning model is crucial to this approach. This approach achieves high results given a large and high quality annotated dataset for training. The drawback is due to the high cost and time-consumption of the annotated dataset and hence domain experts are needed in the process of creating these datasets.

Finally, the hybrid-approaches consists of the advantages of the approaches and avoid their limitations. At the end of the processing, a post-processing step is required to eliminate the conflicts of the various approaches used to obtain the result. And hence this approach is the best approach when compared to the others.

The primary future works consist of performance improvement and training of the datasets to remove the imbalance. Imbalance can be removed by automated text generation techniques based on formal grammar.

3.7 A study of active learning methods for named entity recognition in clinical text

In this study simulated active learning experiments are conducted by Chen et al. using an existing NER corpus with annotated medical problems, treatments and lab tests in the clinical notes. The data set used contains 349 clinical documents with 20,423 unique sentences. The data set is divided into a pool of data to be queried and as independent test sets .5-fold cross validation is used in the experiment. The annotations are converted to BIO format where B-beginning of an entity, I- inside the entity and O-outside the entity. The active learning (AL) experiment consists of the following phases - initial model generation, querying, training, and iteration. Model quality, number of words and number of entities in the annotated set are stored.

The querying algorithms used are uncertainty based, diversity based, and baseline algorithms. The uncertainty based query

algorithms are based on the assumption that the most uncertain sentences are the most informative. The algorithm implements methods such as least confidence (LC), margin, N-best sequence entropy, Dynamic N- best sequence entropy, word entropy, and entity entropy. It, however, suffers from the drawback that it is highly dependent on the quality of the model. Diversity-based algorithms consider information outside the model. It is advantageous as it is independent of the model and it is more efficient in querying as the pairwise similarity score between sentences can be computed. The following are the new features that come into play here: word similarity, Syntax similarity, Semantic similarity and combined similarity. Baseline algorithms consider the length of the words in a sentence. They include querying methods such as length words and length concepts. A random function can also be added.

We evaluated our AL enabled clinical NER using the same type of learning curve that plots the F measure versus the number of annotated strings assuming annotation cost is the same for each sentence. The new method assumes that annotation cost is proportional to the length of the sentence and is better suited to compute the real annotation cost. Area under the learning curve (ALC) score is computed for both methods. ALC1 score is the F-measure versus the number of sentences and ALC2 is the F-measure versus the number of words. Additional curves like the entity count curve that plots the number of entities versus the number of sentences and sentence length curve that plots the number of words versus the number of annotated strings is drawn.

Our results are based on a 5 fold cross-validation. ALC1 scores of uncertainty based sampling methods (0.83) out-performed the two baselines (0.82) which in turn outperformed diversity based methods. All algorithms except syntax similarity were better than random sampling. For AC2 all uncertainty based methods outperformed random sampling. In diversity sampling only semantic sampling outperformed random sampling, ALC2 of baseline methods did not exceed random sampling. The entity count curve reports the total entity count at each iteration while the sentence curve reports the total number of words at each iteration. A simulated study to compare different active algorithms for clinical NER was conducted. The effectiveness of each algorithm, however, has to be further evaluated.

3.8 Boosting drug named entity recognition using an aggregate classifier

Korkontzelos et al propose Active learning approach for Drug Named Entity Recognition in this paper. Active learning framework can be used for reducing the amount of human effort required to create a training corpus. To classify the tokens, two classifiers are used, MaxEnt (Maximum Entropy) which is also known as multinomial logistic regression and a perceptron classifier. The model parameters which yields the result which matches with the empirical expectations and classifies instances so that conditional likelihood is maximized. MaxEnt maximizes entropy while conforming to the probability distribution drawn by the training set. Perceptron is a linear classifier that tunes the weights in a network during the training phase, so as to produce the desired output. The aggregate classifier is classifier is compatible with any dictionaries and recognition systems and could be applied in other domains and sequence recognition tasks. A simple voting-system assuming that the predictions of dictionaries are more reliable than predictions of machine learners. The heterogeneous models have been collaborated to give one model. The model is mixed up with many heterogeneous. The sub-tree mutations and one point crossover between parents of different sizes and shapes. In the pattern augmentation step, 11 patterns are augmented. Genetic programming paradigm nature in that it is a never-ending process. The evolution process is stopped after a certain number of iterations and controlled by setting a predefined maximum tree depth. The patterns were evolved assuming that each will span a single word term. The precision of using the gold standard annotations is comparably high. The usage of the drug dictionary helped a lot in automating the annotation. In the proposed algorithm, the assumptions do not work well. The proposed model is always willing to go with knowledge-based engineering, not with leaning. The classifier trained on gold standard annotations achieved comparable precision but much higher recall. Future work is based on the PK Corpus that involving that highly tagged corpus data, the recall value can be increased further. The proposed model would be extended to probability-based model so that all the three measures can be increased further and all the drug names and drug targets can be identified and annotated without ambiguity.

3.9 Named Entity Recognition over electronic health records through a combined dictionary-based approach

This paper focuses on the challenge of extracting information from the narrative text contained within EHR through combined named entity recognition. In this paper Quimbaya et al. propose an approach for dictionary-based, combined NER aimed at improving entity recall dealing with the aforementioned challenges associated with clinical narrates. The data set contains several files from which we use the Complete Set for the Risk Factors Task combining the training and test sets to have 1304 texts in total, to build the dictionary of terms we use the UMLT meta-thesaurus to obtain names of the diseases and the medications. To study the impact of the fuzzy and stemmed versions of the NER, they compare four different combinations of the results against the gold standard: e, ef, es, efs where e means it contains the exact, the fuzzy and the stemmed annotations. Also, the measure F2, which favours recall over precision, improves in all the models. This paper proposes an approach for named entity recognition in narrative texts of EHR. This task is difficult because of the particular and unique characteristics of texts in health records (codified, condensed, jerky language). This approach combines a direct match technique with frizzy matching and stemmed matching. The proposed method was tested using the i2b2 NLP Data Set which includes a gold standard with annotations. Experimental results on this data set show an improvement of the recall while having a limited impact on precision. Nevertheless, when analyzing the text, the proposed methods do not take into account the surrounding words (the context) appearing near a named entity candidate. In future works, taking this context into account may be of great interests for improving both precision and recall.

3.10 A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set

Al-Hegami et al. perform NER on bio-medical data using machine learning approach. This approach involves steps such as data preprocessing, feature extraction, classification and evaluation. The preprocessing phase makes use of chunking or instance pruning so that the training data is in a suitable format. The feature extraction process is designed in a

domain-independent fashion in such a way that the features are binary. Here we extract morphological, orthographic, text-based, linguistic-based and other domain-dependent knowledge. The extracted morphological features include capitalization, numeric, punctuation, uppercase, lowercase, single character, symbol, includes hyphen, includes slash, letters and digits, capital and digits, includes caps. The machine learning phase makes use of SVM, Naive Bayes and Artificial Neural Networks.

The performance measures to evaluate NER used here are precision, recall and weighted mean. A number of experiments have been conducted and evaluated by using ten-fold cross-validation. The following results were obtained while empirically comparing 10 features and three classification approaches the classification approaches used were KNN, Naive Bayes and decision tree classification. The obtained proved that KNN had the highest effect on the quality of biomedical NER. The results obtained using Nave Bayes is less than that obtained using KNN. Similarly, the results obtained using Naive Bayes is less than that obtained using KNN. A model for NER is produced based on machine learning techniques. Based on the result analysis the model is found to effective for biomedical named entity recognition.

3.11 Entity recognition in the biomedical domain using a hybrid approach

In the biomedical domain, there is a large number of scientific findings day by day. This leads to increased difficulty for the scholars in using the data. Through this paper Basaldella et al. propose a method to solve this problem using named entity recognition in which we detect terms belonging to a set of predefined entities. This paper presents a very efficient NER as well as concept recognition and evaluation. The method consists of two stages. The first stages comprise of a dictionary based annotator and the second is a machine learning classifier.

Due to the use of different type of algorithm, different features are used to train neural networks and conditional-random-fields. The main difference between the two is that NN works in both n-ary features and continuous-valued features while CRF works only in continuous-valued features. Hence these two features are used as per the context.

The future work of this problem will consist of specific terminological categories where the classifiers fail to obtain good performance. It also lies in the improvement of the recall by allowing multi-word terms to be reordered, shortened or term expansion. We can also expand the evaluation to the full CRAFT corpus using various other resources to train the system.

3.12 Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora

The motive of this paper is to give importance to the languages apart from English in clinical text mining. Medical entity recognition is the core of clinical information retrieval in any language. The medical retrieval problem is not completely solved and still faces challenges in the English language. By research, it was found that Swedish and Spanish were a widely explored language in natural language processing which gives us an upper hand during information retrieval.

From this project, we were able to learn the impact on each language caused by using the same methods in Spanish and Swedish. Prez e al. used a baseline and three types of taggers on different feature-sets. We have used unsupervised learning techniques on medical entity recognition in EHRs. This consisted of ensembles of brown trees and semantic spaces unclustered form. To form the clusters, brown trees were pruned and to obtain the code-words vector quantization was used. It was found out that, in order to eliminate data sparsity, clustering was the best approach. We should also give importance to the present tools present in the biomedical domain. The models on being trained were all fit to perform on-line entity recognition. The achieved precision and recall is high but there is still a lot to be improved. This can be done using alternative methods for feature extracting and employing ensemble learners such as stacking.

3.13 Integrating Bilingual Named Entities Lexicon with Conditional Random Fields Model for Arabic Named Entities Recognition

Arabic is a rich language and is linguistically far ahead than most of the languages. In this paper, Hkiri et al. propose a hybrid NER system that applies conditional random fields (CRF), bilingual NE lexicon and grammar rules to the task of Named Entity Recognition in Arabic languages.

Initially, the data is cleaned and annotated manually by people in an XML format. The dataset is retrieved from the United Nations dataset and the ANERcorp dataset. The main approach to the solution is based on a combination of both rule-based and ML-based components collectively called the hybrid system. In the rule-based system since the rules for the combinations are difficult to handle in Arabic a suitable mapping is created between the English and the Arabic named entities. This reduces a lot of time that was consumed in the old approaches. The number of gazetteers is halved which results in fewer annotations and hence more response. The ML component then receives the output from the rule-based component and then it predicts the correctness of the rule-based approach for Named entities. The supervised model of the ML component contains the conditional random fields. This involves feature extraction and the features used for NE classification are rule-based characteristics, morphological features, PoS feature, Gazetteer features, punctuation and so on.The results obtained in the above approach has better precision than the previous approaches. This is mainly due to the combination of the rule-based component, bilingual NE lexicon and the ML component based on the CRF model. The use of the rule and ML-based approach has increased precision in finding the NE in the Arabic language which was initially complex due to the nature of the language. The rule-based component is enriched with the help of NELexicon extracted from the DBpedia linked datasets. The ML component helped in the quality of annotations by the rule-based approaches further approaches should focus on increasing the grammatical rules developed within the rule-based components. Also, a different ML component can be used in place of CRF.

3.14 Medical Entity Recognition using Conditional Random Field (CRF)

In this paper, Herwando et al. propose Medical Entity Recognition (MER) with the help of Conditional Random Field for online medical question answering based on their entities. The first step is to get the content from the online forums and then filter the responses received. Filtering is done by eliminating multiple responses from the same user and

then users providing the same content. Then the conditional random field is used to find the entities with parts of speech a, medical dictionary and words. After filtering the content the input is subjected to POS tagger. The tagger is based on the Indonesian text since the input is based on the Indonesian language. After tagging the output is subjected to BIO encoding where the particular entity type is found based on the combination of features. The training data is made ready with the help of manual annotation and manual tagging by people.

The results obtained in the MER approach has better precision than the previous approaches. This is mainly due to internal factors such as errors during the tokenization process. The result is due to the fact that inconsistencies are removed in the POS tagging phase. This approach does not use formal medical documents and cannot be reliable as the previous approaches. Further improvements can be some to make this approach get data from the formal document rather than from online discussion forums.

3.15 A method for named entity normalization in biomedical articles: application to diseases and plants

In this paper an approach is used for normalizing the biological entities such as disease names and plant names by Cho et al. This is done with the help of word embedding's to represent semantic spaces. The training data for the disease names are acquired from the NCBI disease corpus and the unlabeled data is received from PubMed abstracts which are used to construct the word representations. For the name of plants, the training corpus is manually created and unlabeled data is created with PubMed abstract to create the word vectors.

The results obtained in the approach has better precision than the previous approaches. It has been showed that the normalization accuracy has been improved by the model even when the dictionaries are not competent enough. When normalization is done only for the disease names they outperformed the existing approaches and indeed the best system in the task. In this study, they integrate training data and unlabeled data for word representation in entity name normalization and verified that the proposed normalization outperforms the existing tools. But since the approach depends on a comprehensive dictionary the research should be done to tag those names that are not part of the dictionaries and thus a whole diverse class of entities can be normalized

3.16 Identifying Disease-related Expressions in Reviews Using Conditional Random Fields

In this paper, Z. Sh et al. use the Conditional Random Fields algorithm to tag the comments as it is better than most of the other approaches used in tagging. The disease-related entity extraction is treated as a sequence labelling problem. The CRF algorithm takes in a series of inputs and then calculates the probabilities of the predefined labels and at last assigns the text to a label that has the largest probability outcome. The tagging system that is used is called the BIO tagging system. These are identified at the document level. Initially, the features like word, POS, suffix and prefix and so on are calculated and labelled. The dictionaries are used in the expansion of some words like acronyms. Words are represented in two forms via cluster based and distributed. The training data contains a huge number of user comments from various resources and then the duplicate texts are removed and then all the letters are lowercased. The words are partitioned into 150 clusters with the help of Brown hierarchical clustering algorithm.Word2vec is used to train embedding's on Health Dataset.

The above-specified method is trained with the help of CADEC corpus that annotated with Drug and Disease entities at the sentence level. The model is evaluated on the basis of two baseline methods a knowledge-based one and a bidirectional RNN.CRF and RNN extract 92% of the annotations correctly and the NER does not exist here because of the use of RNNs. CRF produced the best results ranging up to 79.4% which led to the investigation of evaluation of the CRF features. The dictionaries based on the domain texts made a significant rise in the performance of CRF.

This paper relates to recognizing the opinion expressions about a disease in the public and then the social media. It has been found that the CRF algorithm gave the best results compared to the dictionary and the RNN methods. Additional research should be done to find out whether the above effort may be as a result of medication or another lifestyle. The data should be annotated to be made available in many languages to facilitate opinion recognition throughout the whole world.

3.17 Arabic Named Entity Recognition Using Topic Modeling

Bazi et al. used 2 approaches: Topical Word Embedding's (TWE) features and Topical Prototype (TP) features in this paper for Arabic NER. Topical Word Embedding's is a multi-prototype word embedding prototype. It performs LDA with Gibbs sampling to obtain topics. The three models are TWE-1, TWE-2, TWE-3. In this study, we only make use of TWE-1 and TWE-3 as they produce the best results.

The primary idea in Topical Prototypes is that similar words have a high probability to be tagged with the same entity label. The topic distributions are obtained by performing LDA on the text-corpus. The Normalised Point wise Mutual Information of each label is calculated and the top 'm' words are chosen as the prototypes of each label. The prototypes are the features to our supervised model.

The NER model used here is a Conditional Random Fields classifier. The feature set includes the word, part-of-speech tag, affixes. The morphological features include aspect, case, gender, number, and norm word. The Arabic Wikipedia Named Entity Corpus (AQMAR) consists of 74000 tokens and 2687 sentences. One half was used as development corpus and the other as the testing corpus. The testing corpus was used in the training phase.

The performance measure to evaluate NER used here is F1 measure. The best results are achieved by combining Brown clusters with world-class and topical prototype feature with 73.85% which is four and a half points higher than the baseline. Thus, novel features based on LDA are designed. Future works include incorporating more appropriate ways of including LDA into NER

3.18 A co-training based entity recognition approach for cross-disease clinical documents

Chen et al. propose a method here that is based on co-training and LSTM-CRF is selected as the basic learner for cross-disease clinical documents. The system makes use of 1)a partial annotation corpus of the single disease where dependency parsing is used to mark the corpus and reduce

time consumption and 2)entity recognition by cross-disease co-training-where we use LSTM for feature extraction and the results are added to the CRF layer. This is followed by a feedback mechanism.

The LSTM model is preferred as it can store the associations between related clinical documents and can save contextual information. The gate structure is used here. The sequential labelling is done by CRF as it is proved to be better at computing the joint probability and optimizing sequence labels. The results are combined with that of the LSTM results and the softmax function is computed. Thus the score of the correct tag sequence can be maximized. Cross-disease co-training model is a method of using different kinds of clinical characteristics of multi-document view. The model was evaluated on a collection of thyroid nodules ultrasound. The experimental data is derived from RUIJIN Hospital. The test set contains 19378 patients.

The performance measures to evaluate NER used here are accuracy, recall, and F1-score. Among the artificial and mechanical construction methods for tagging datasets, the accuracy of the artificial method is high but it is time consuming while the accuracy and time consumed by the mechanical method are less. Hence we use a combination of manual and syntactic analyses. As the data sets increase the accuracy of recognition also increases. While the accuracy of CRF is higher than LSTM and HMM, the recall is lower. This shows that LSTM is better than CRF in big data sets. Accuracy and recall are high when we combine both methods.

When compared to previous systems the performance achieved is by far the best. Future methods involve optimizing the feedback mechanism to make the results better and improve accuracy. The basic learner could also be improved.

3.19 A comparative study for biomedical named entity recognition

As one of the fundamental biomedical text mining tasks, Named Entity Recognition (NER), aiming at identifying chunks of text referring to specific entities of interest, plays a key role in disease-treatment relation extraction, gene function identification and semantic relation extraction between concepts in a molecular biology ontology. In this paper, Wang et al. compare 6 biomedical NER tools.

ABNER has a preventative graphical interface and contains two modules for tagging entities (e.g. protein and cell line) trained on standard corpora. This tool is written in Java and employs graphical window objects in the Swing library. With optimization for biomedical text, such as MEDILINE abstracts, GENIA tagger functions well in the biomedical domain. It is a good option to extract information from biomedical documents because it is trained on three corpora, the Wall Street Journal corpus, the GENIA corpus and the PennBioIE corpus, respectively. The developers apply the bidirectional algorithms and achieved equivalent performance compared with other machine learning models. The algorithm finds the highest probability sequence and the corresponding decomposition structure in polynomial time among all the possible enumerated decomposition structures. The tagging result of GENIA Tagger contains five entities, which are protein, DNA, RNA, Cell Line and Cell type. To sum up, Gimli performs better than all the other five tools on GENETAG corpus, achieving the accuracy of 90.22% and F-measure of 87.17%, which are 1.56 and 0.74% improvement over the second best tool, BANNER, respectively. LingPipe is a tool kit for processing text using computational linguistics. LingPipe contains a model trained on GENETAG corpora, which makes it capable of recognizing named entity like proteins, genes, etc. in biomedical text. LingPipe and BANNER do not support JNLPBA corpus. For all the five categories of biomedical name entities, Gimli achieves the highest F-measure, which is 0.86 and 116% higher than the other ones in top 3. Gimli achieves 1.94 and 2.53% higher performance than NERSuite correspondingly. Because of the complex solutions that include the application of linguistic, lexicon features and the combination of various CRF methods, Gimli outperforms the other five tools and achieves the highest overall performance, again. Besides the comparison of F-measure, we regarded the processing speed as a critical evaluation criterion considering the burst amount of recently published literature. BANNER is not as fast as ABNER, but it ranks first considering speed and F-measure, thus it is a suitable option for biomedical named entity recognition.

4. CONCLUSION

This paper provides a review of Named Entity Recognition methods in the biomedical field. It explains the need for Named Entity Recognition in this field along with the various problems in the process including the challenges of extracting information from narrative text and the lack of performance of Named Recognition works designed for English in this biomedical field. In this paper, we discuss different techniques like CRF, HMM, Perceptron, Logistic Regression, Naive Bayes, Supervised models, Unsupervised models, Semi-supervised models that have been used for NER in the Biomedical field. Out of these categories, it has been observed that machine learning based approaches are the best suited and most popular. This paper also discusses the limitations of these approaches and the future works that can be tried out for better performance.

5. REFERENCES

[1] Mnica Marrero, Julin Urbano, Sonia Snchez-Cuadrado, Jorge Morato, Juan Miguel Gmez-Berbs ”Named Entity Recognition: Fallacies, Challenges and Opportunities” Published 2013 in Computer Standards and Interfaces.

[2] Shaodian Zhang, Nomie Elhadad ”Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts” Journal of Biomedical Informatics 46 (2013) 10881098.

[3] Rezarta Islamaj Dogan, Robert Leaman, Zhiyong Lu ”NCBI disease corpus: A resource for disease name recognition and concept normalization” Journal of Biomedical Informatics 47 (2014) 110

[4] Yaqiang Wang, Zhonghua Yu, Li Chen, Yunhui Chen, Yiguang Liu, Xiaoguang Hu Yongguang Jiang ”Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study” Journal of Biomedical Informatics 47 (2014) 91104.

[5] Robert Leaman, Chih-Hsuan Wei, Zhiyong Lu ” tmChem: a high-performance approach for chemical named entity recognition and normalization” Leaman et al. Journal of Cheminformatics 2015, 7(Suppl 1): S3.

[6] Shengyu Liu, Buzhou Tang, Qingcai Chen and Xiaolong Wang ”Drug Name Recognition: Approaches and Resources” Information 2015, 6, 790-810.

[7] Yukun Chen, Thomas A. Lasko, Qiaozhu Mei, Joshua C. Denny, Hua Xu ”A study of active learning methods for

named entity recognition in clinical text" Journal of Biomedical Informatics 58 (2015) 1118. corporais Korkontzelos, Dimitrios Piliouras, Andrew W. Dowsey, Sophia Ananiadou "Boosting drug named entity recognition using an aggregate classifier" Artificial Intelligence in Medicine 65 (2015) 145153.

[8] Ioannis Korkontzelos, Dimitrios Piliouras, Andrew W. Dowsey, Sophia Ananiadou "Boosting drug named entity recognition using an aggregate classifier" Artificial Intelligence in Medicine 65 (2015) 145–153

[9] Alexandra Pomares Quimbaya, Alejandro Sierra Munera "Named Entity Recognition over electronic health records through a combined dictionary-based approach" Conference on Enterprise Information Systems / International Conference on Project Management / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016

[10] Ahmed Sultan AL-Hegami, Ameen Mohammed Farea Othman, Fuad Tarbosh Bagash "A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set" IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.1, January 2017

[11] Marco Basaldella, Lenz Furrer, Carlo Tasso and Fabio Rinaldi "Entity recognition in the biomedical domain using a hybrid approach" Journal of Biomedical Semantics (2017) 8:51

[12] Alicia Prez, Rebecka Weegar, Arantza Casillas, Koldo Gojenola, Maite Oronoz, Hercules Dalianis "semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora" Journal of Biomedical Informatics 71 (2017) 1630

[13] Raditya Herwando, Meganingrum Arista Jiwanggi, Mirna Adriani "Medical Entity Recognition using Conditional Random Field (CRF)" 2017 14th IAPR International Conference on Document Analysis and Recognition

[14] Emna Hkiri, Souheyl Mallat, Mounir Zrigui "Integrating Bilingual Named Entities Lexicon with Conditional Random Fields Model for Arabic Named Entities Recognition" WBIS 2017 978-1-5386-2038-0/17

[15] Hyejin Cho, Wonjun Choi and Hyunju Lee "A method for named entity normalization in biomedical articles: application to diseases and plants" BMC Bioinformatics (2017) 18:451

[16] Miftahutdinov Z. Sh., Tutubalina E. V., Tropsha A. E. "Identifying Disease-related Expressions in Reviews Using Conditional Random Fields" Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2017

[17] Ismail El Bazi, Nabil Laachfoubi "Arabic Named Entity Recognition Using Topic Modeling" International Journal of Intelligent Engineering Systems

[18] Dehua Chen, Nannan Che, Jiajin Le, Qiao Pan "A co-training based entity recognition approach for cross-disease clinical documents" Article in Concurrency and Computation Practice and Experience.

[19] Xu Wang, Chen Yang, Renchu Guan "A comparative study for biomedical named entity recognition" Springer-Verlag Berlin Heidelberg 2015