Named Entity Recognition Methods and Techniques: A Review

Rohit Narain[1], Neha Bathla[2]
Computer Science and Enginnering Department, Yamuna Institute of Engineering and Technology, Yamuna Nagar, Haryana

Abstract - In this research paper we have reviewed various named entity recognition approaches for extracting the entities in various languages that are spoken by humans. The results of various approaches have been reported later in the paper. We will also propose hybrid named entity recognizer which is capable of finding simple as well as complex entities such as (to date, from date, Credit amount, Debit Amount etc.) from the natural language text. During the Turing test, the accuracy achieved by the system is 98.2 % which is higher than any other system. We have also studied the problems and challenges associated with the various named entity recognition techniques.

I. INTRODUCTION

Now days many of the applications in the world needed the named entity recognition system are needed to process large amount of data, such applications includes question answering system, customer care support, search engine, automatic document summarization, efficient search algorithms, classifying contents for news papers. The named entity recognition is designed to extract the entities from the natural language text/sentences and machine readable text. The named entity recognition extracts entities such as name of person, name of organization, name of location, date, time, days, email etc. which is used further for text categorization, automatic text summarization, classification and recognition of various entities. However these problems cannot be solved by the traditional approaches because most of the data used today are unstructured in nature and is available in a very large amount which is time consuming. Many problems arise during designing these methods such as word sense disambiguation, recognition of foreign words in the text, agglutinative and inflectional nature of languages which will be discussed later in the paper. In this research paper we will review the various approaches that were used in designing the named entity recognition system and will also proposes the hybrid named entity recognizer which is capable of finding system as well as complex entities(e.g. to date, from date, credit amount, debit amount, promo amount etc) from the text.

II. LITERATURE REVIEW

In the recent years many researchers have worked on named entity recognition system and proposed many approaches. Although each approach have its own pros and cons. Some of these techniques are discussed below in detail:

2.1 Rule based Approach

Rule based approach is a type of approach in which set of rules for extracting the entities from the text. The system uses these rules to extract entities such as names, locations, time and date from the natural language text.

Kaur proposed a system for identifying the entities in the Hindi language. The proposed named entity recognition system was based on the hybrid approach which was the combination of the two approaches i.e. rule based approach and list look up approach. The system was compared with the supervised learning system known as NEC module of Freeling. The system achieved the accuracy of 90% as compare to FreeLing system. The accuracy of the system depends upon the number of entities and hand crafted rules stored in the database of the system which was the main limitation (Kaur et al., 2015).

Petasis et al. proposed a system for classification and recognition of entities using rule based approach. The machine learning approach was used in order to maintain the system. The system did not require any human intervention during the tagging of entities in the text. The system was tested using the two languages (Greek and French), that included 180,893 instances in 6,000 documents (Petasis et al., 2001)

2.2 Statistical Learning Approach:

Statistical Learning Approach uses statistical techniques to train the systems. The statistical learning uses probability measures like unigram, bigram and trigram. Statistical learning provides the capability to learn statistical regularities from the world.

Nie et al. introduced the interactive knowledge framework also known as the iKnoweb. The system was capable to work on the structured entities, named entities, entity facts and relations from the web. The system used the statistical extraction (Nie et al., 2011).

Jayan et al. proposed a system that used a hybrid statistical machine learning approach which was the combination of rule based machine learning and statistical approach. The system was compared with the two supervised taggers known as TnT and SVM, as per the results of both taggers, for known words

SVM showed better results and for unknown words TnT showed better results. The system achieved the accuracy of 73.42% (Jayan et al., 2013).

2.3 Hidden Markov Models:

Hidden Markov model can be called as the statistical markov model. The model can be described as the dynamic Bayesian model. Many researchers have applied Hidden Markov models in the named entity recognition system.

Etizioni et al. used hidden Markov models to present three ways for recall and extraction rate of entities. The extraction rate of entities was increased by automatically identified as the subclass. The recall of the system was increased by the Pattern learning, Subclass extraction and list extraction. The method improved the recall at the precision of 0.90 and discovered 10,000 cities missing from the gazetteer (Etizioni et al., 2005).

Florain et al. presented a classifier combination experimental framework for named entity recognition which was based on the four different classifiers such as robust linear classifier, maximum entropy, transformation based learning and hidden markov model. The system achieved the accuracy of 91.6 F-measures score (Florain et al., 2003).

Zhou et al. proposed the hidden markov model and HMM based chunk tagger for classifying and recognizing the entities in the system. The system achieved the accuracy of F-measures of 96.1% and 94.1% for both the MUC-6 and MUC-7 English named entity tasks (Zhou et al., 2002).

Zhang et al. improved the features of named entity recognizer that was based on the HMM technique and also studied the various characteristics of biomedical entities. The system introduced new features such as orthographic, morphological, part-of-speech and semantic trigger features. The proposed system with new features achieved the accuracy of 66.5 and 62.5 of F-measure score (Zhang et al., 2004).

Morwal et al. proposed a technique for the identification of named entities based on the hidden markov model approach. The proposed system was language independent and was very efficient and capable for Indian languages. The proposed system achieved the high accuracy of 90% during testing (Morwal et al., 2012).

Saha et al. proposed a system that was based on the trigram hidden markov model. The system was trained using the dataset extracted from the FIRE 2015 task. The system achieved the precision, recall and F-measures of 61.96, 39.46 and 48.21 (Saha et al., 2009).

2.4 Maximum Entropy Markov Models

The maximum entropy markov models are also known as discriminative models. Roy presented the hybrid system for named entity recognition. Maximum entropy model, language specific rules and gazetteers was used. The system was designed to recognize context patterns for Hindi and Bengali language. The system achieved the accuracy of f-value of 65.13 and 65.96% for both Hindi and Bengali, for Oriya, Telgu and Urdu the system achieved 44.65%, 18.74% and 35.47% respectively (Roy, 2012).

Collins et al. presented the algorithm for re ranking of top N hypotheses from a maximum entropy tagger, two approaches were used during the implementation of the system i.e. the boosting algorithm and voted perceptron algorithm. The presented algorithms give the significant better improvement i.e. 15.6% for boosting and 17.7% for voted perceptron over the maximum entropy baseline (Collins et al., 2002).

Saha et al. [13] presented the study on clustering of the words and selection based feature for named entity recognition. The study was based on the maximum entropy classifier. The system obtained f-value of 72.55 with the deep domain knowledge. The obtained f-value was 64.1, when the system was used only POS information as domain knowledge (Saha et al., 2009).

2.5 Conditional Random Field

Conditional Random Fields models are widely used model in the field of natural language processing. This is a type of discriminative modeling and based on the undirected graphs.

Rocktaschel et al. presented the named entity recognition system. The system was based on the hybrid approach that was the combination of conditional random field with the dictionary. The system was also known as chemspot. The system achieved the accuracy of F1 measure of 68.1% on the SCAI corpus (Rocktaschel et al., 2012).

Liu et al. presented a framework based on the combination of K-Nearest Neighbor classification and linear conditional random fields under the semi supervised learning. According to the results, predicted labels of the KNN classifier achieved the F1 as high as 80.2% while 22.2% accounted for all predictions. Similarly the baseline system that was based on the CRF model achieved the F1 of 75.4% (Liu et al., 2011).

Sobhana et al. proposed the named entity recognition system for extracting the geological text based on the conditional random field approach. The proposed system had the capability to extract 17 different types of classes of entities in the text. For the training purpose, 2 lakh words were manually fed into the dataset from the collection of scientific reports and articles of geology. The system achieved the accuracy of 75.8% F-measure (Sobhana et al., 2010).

Ekbal et al. proposed a named entity recognition system for Bengali language using that was based on the conditional random field. The proposed system used the different contextual information of words and different features for the extraction of named entities. For the training purpose, 150K words were manually fed into the dataset. The system achieved the accuracy of recall, precision and F-measure of 93.8%, 87.8% and 90.7% (Ekbal et al., 2008).

III. PROBLEMS AND CHALLENGES IN NAMED ENTITY RECOGNITION APPROACHES

Sometimes natural language that are spoken and written by humans are difficult to analyze and understand, due to which it becomes infeasible to extract and recognize entities from that language which in turns creates some problems in the system. This makes the named entity recognition system difficult to design and implement. However various approaches were used in designing the entity recognition but we still have some challenges. The main challenges in named entities recognition systems are described below:

3.1 Ambiguities

The ambiguities in classifying named entities are the biggest challenge. Some languages have many words that have multiple meanings or definitions that creates the problem. This makes the system difficult to analyze which word is to take or not. For better understanding we take the example the word 'Deenanath Mangeshkar' can be overlapped and have multiple meanings such as it is the name of person or name of organization.

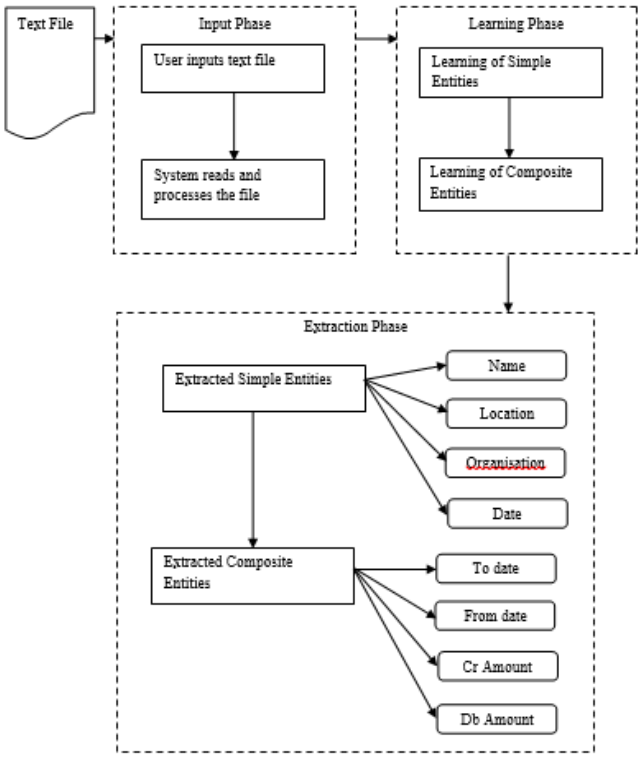

Fig1. Workflow of Proposed Hybrid Entity Recognizer

3.2 Foreign Words

The other challenges involves the foreign words recognition in the language, sometimes the language spoken by humans have many words that does not belong to one language which is very difficult to analyze as a entity. Consider the example, Adolf, Bruce Lee, Mercedes, Barack, Delhi etc. these words are not limited and can be coming anytime which makes the system difficult to recognize the entities.

IV. CONCLUSION

In this paper, we have reviewed all the approaches for named entity extraction for extracting the entities from the text. We have studied the accuracies, results of each approach and problem associated during the design and implementation of each approach. We have also proposed a hybrid approach for named entity extractor that is capable for extracting simple as well as composite entities from the text. The approach is a combination of statistical machine learning and memory based machine learning. The proposed approach is capable of learning any linguistic patterns from the text and easily learn new entities from the human. The proposed system achieves the accuracy of 98.2% which is significantly higher than the any other system in the Turing test.

V. REFERENCES

[1]. Collins M. (2002). Ranking Algorithms for Named–Entity Extraction: Boosting and the Voted Perceptron. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 489-496.

[2]. Etzioni O., Cafarella M., Downey D., Popescu A.M., Shaked T., Soderland S., Weld D.S., Yates A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. Artificial Intelligence, 165, 91-134.

[3]. Ekbal A., Haque R., Bandyopadhyay S. (2009). Named Entity Recognition in Bengali: A Conditional Random Field Approach. Linguistic Issues in Language Technology – LiLT, 2(1) 589-594.

[4]. Florian R., Ittycheriah A., Jing H., Zhang T. (2003). Named Entity Recognition through Classifier Combination. In CONLL '03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL, 4, 168-171.

[5]. Jayan J.P., Rajeev R.R., Sherly E. (2013). A Hybrid Statistical Approach for Named Entity recognition for Malayalam Language. In International Joint Conference on Natural Language Processing, 58–63.

[6]. Kaur Y., Kaur R. (2015). Named Entity Recognition (NER) System for Hindi Language Using Combination of Rule Based Approach and List Look Up Approach. *International Journal of scientific research and management*, 3(3), 2300-2306.

[7]. 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, 359-367.

[8]. Morwal S., Jahan N., Chopra D. (2012). Named Entity Recognition using Hidden Markov Model (HMM). International Journal on Natural Language Computing (IJNLC), 1(4), 15-23.

[9]. Nie Z., Wen J. R., Ma W.Y. (2011). Statistical Entity Extraction from Web. *IEEE*, 1-12.

[10]. Petasis G., Vichot F., Wolinski F., Paliouras G., Karkaletsis V., Spyropoulos C.D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. In *ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 426-433.

[11]. Roy S. (2012). Named Entity Recognition. AKGEC International Journal of Technology, 8(2), 38-41.
[12]. Rocktäschel T., Weidlich M., Leser U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640.
[13]. Saha S.K., Sarkar S., Mitra P. (2009). Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of Biomedical Informatics, 42, 905-911.
[14]. Sobhana N.V., Mitra P., Ghosh S.K. (2010). Conditional Random Field Based Named Entity Recognition in Geological Text. International Journal of Computer Applications, 1(3), 119-122.
[15]. Zhou G.D., Su J. (2002). Named Entity Recognition using an HMM-based Chunk Tagger. In *the 40th Annual Meeting of the Association for Computational Linguistics (ACL) proceedings*, 473-480.
[16]. Zhang J., Shen D., Zhou G., Su J., Tan C.L. (2004). Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37, 411-422.