International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 22 (2018) pp. 15451-15455

Hybrid Named Entity Recognizer for Recognizing Entities

Rohit Narain[1], Neha Bathla[2]

[1,2]Computer Science and Engineering Department, Yamuna Institute of Engineering and Technology, Yamunanagar
[1]C-4/1499, Jaroda Gate, Nr. Old Grain Market, Jagadhri, Yamuna Nagar, Haryana - 135003, India.

Abstract
This research paper presents the hybrid named entity recognizer for presenting the recognizing and extracting the named entities from the natural language text. The proposed system is based on the hybrid approach that uses statistical machine learning and memory based learning for the training of the system. The system is enough capable to recognize simple as well as complex entities from the text such as (to date, from date, credit amount, debit amount) etc. The proposed system requires less intervention of human in learning process. To analyze accuracy of the proposed approach, 1,000 sentences from newspapers, chat conversations have been used. During the Turing test, the proposed approach achieved the 98.2% of accuracy which is significantly higher than other systems available in the market.

1.1 INTRODUCTION

The Named Entity Recognition is used to extract or recognize entities from the natural text that are spoken and written by the humans. The entities involve the important text from the sentences such as Name of Person, Name of Organization, Locations, Time, Date, money and numerical quantities etc. The named entity recognition is now days becomes a very important part in many applications such as efficient search algorithms, classification contents for newspapers provider, customer support, question answering systems etc.

Most of the approaches have been proposed and developed in order to automatically extract and recognize the entities from the text. The main problems associated with these approaches were the word ambiguities, recognition of foreign words and entities, extraction of complex entities and Agglutinative and Inflectional nature of Languages.

In this research paper, we presents a hybrid named entity recognizer that is capable of extracting the simple as well as complex entities i.e. it can easily recognize sub-entities from the text such as to date, from date, credit amount, debit amount etc. The system requires less intervention of human in feeding and training of systems. The proposed system uses the hybrid approach that combines the statistical machine learning and memory based learning together.

1.2 RELATED WORK

In the recent year most of the approaches on named entity recognition and extraction have been proposed by the researchers. Some of the researchers proposed the named entity recognition which was based on the rule based approach, in which set of rules were used in order to extract entities from the text.

Kaur et al. [1] presented a system for identification of the entities in the Hindi language. The system was developed using the hybrid approach that 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 main limitation of the system was that the accuracy of the system depends upon the number of entities and hand crafted rules stored in the database.

Petasis et al. [2] 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.

Some of the researchers used the statistical methods in order to train the systems which provide the ability to learn statistical regularities from the world using the probability measures (unigram, bigram and trigram).

Jayan et al. [3] 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%.

In order to increase the extraction rate and accuracy of the entities, some researcher used Hidden Markov model which can be described as the Dynamic Bayesian model. Etizioni et al. [4] 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.

Florain et al. [5] 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.

Zhou et al. [6] 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.

Zhang et al. [7] 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.

Some of the researchers have used the maximum entropy markov models for re ranking of entities and context patterns. These models are also called as discriminative models; Roy [8] 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.

Collins et al. [9] 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.

Saha et al. [10] 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.

1.3 PROPOSED HYBRID ENTITY RECOGNIZER

The proposed system is based on the hybrid approach which is the combination of two approaches i.e. statistical learning approach and memory based learning approach. The proposed system uses these two approaches as a two different parts. One is for extracting the simple entities and other is for extracting the composite entities from the natural language text. During the extracting phase the system uses statistical learning approach for extracting the simple entities such as name of person, organization, location etc. from the text. Finally the composite entities (to date, from date, credit amount and debit amount etc.) are extracted from the simple base entities using the memory based learning approach. The entire workflow of the proposed system is as shown in fig.1 below

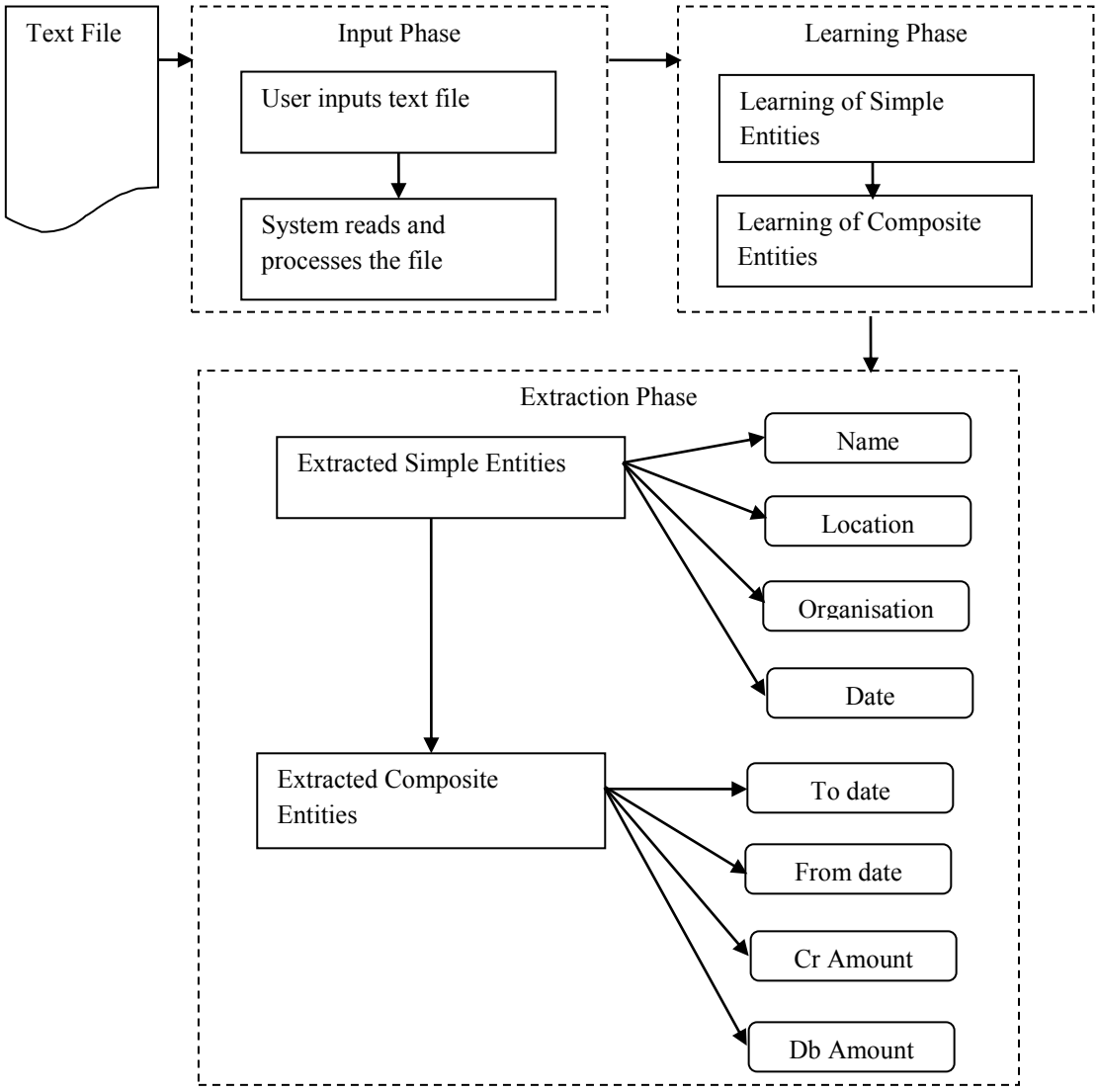

Fig1. Workflow of Proposed Hybrid Entity Recognizer

1.3.1 Dataset used

For training of the proposed system we use the dataset of 1,000 sentences from various newspapers and chat conversation in order to learning the entities to the system. Each sentence is made up of various specific words like name of person, name of organization, name of location, date (to date, from date), amount (credit amount, debit amount, promo amount).

For analyzing the accuracy of the proposed system during the testing phase, we used a dataset of 1,000 sentences from various newspapers and chat conversations for extracting the entities from the sentences. The output of the extracted entities from various sentences is shown in table 1 below.

Table 1: Dataset of sentences used during testing on the system

S.No.	Input Sentences	Actual Evaluated Results				
		Name	Organisation	Date	Location	Currency
1	The payment I made on date 12-Oct-2016	I		12-Oct-2016		
2	The $100 I paid in Oct month	I		Oct		100$
3	The payment that I made in the Oct month	I		Oct		
4	The payment that I made last month	I		PromoStartDate: [01-03-2017], PaymentEndDate: [31-03-2017], PromoEndDate: [31-03-2017], PaymentStartDate: [01-03-2017]		
5	The USD100 payment that I made last month	I		PromoStartDate: [01-03-2017], PaymentEndDate: [31-03-2017], PromoEndDate: [31-03-2017], PaymentStartDate: [01-03-2017]		USD 100
6	The $ 100 payment that I made on Oct month	I		PaymentEndDate: [oct], PromoStartDate: [oct], DATE: [oct], PromoEndDate: [oct], PaymentStartDate: [oct]		100$
7	The 100 dollar payment that I made on 12-Oct-2016	I		PromoStartDate: [12-oct-2016], DATE: [12-oct-2016], PaymentEndDate: [12-oct-2016], PromoEndDate: [12-oct-2016], PaymentStartDate: [12-oct-2016]		100 dollar
8	The promo that expires on month of March			PromoStartDate: [01-03-2017], DATE: [march], PaymentEndDate: [31-03-2017], PromoEndDate: [31-03-2017], PaymentStartDate: [01-03-2017]		
9	The promo that expires on 25-Mar-2017			PromoStartDate: [25-mar-2017],DATE: [25-mar-2017], PaymentEndDate: [25-mar-2017], PromoEndDate: [25-mar-2017], PaymentStartDate: [25-mar-2017]		
10	The $780 promo that expires March month			DATE: [march],		$780

1.3.2 Algorithm used to implement Proposed System

The algorithm for the proposed model is described below:

Step1: Start

Step2: The sentence is inputted into the system.

Step3: The system passes the sentence to the text processing module for extracting the entities.

Step4: (a) the text processing passes the sentence to Syntax net for Parts of Speech (POS) tagging.

(b) The text processing passes the sentence to rule based entity extractor for applying rules on the sentence.

Step5: The rule based entity extractor checks the correspondent rules or meaning of entities in the entity corpus.

Steps6: The processed sentence is then feed into the Machine learning based entity extractor.

Step7: The basic entities are extracted.

Step8: The Basic entities are separated into the base entity buckets and memory based linguistic patterns are other separated. The composite entities are then extracted from the bucket.

Step9: Stop.

1.4 RESULT AND DISCUSSIONS

As per the results, the proposed system is capable of acquiring simple as well as composite entities. The system performs well with the text that consists of composite entities such as (to date, from date) from the text. The proposed system achieves the accuracy of 98.2% during the Turing test. In this research paper, the comparison between the approach used in proposed system and the existing approaches is done and given below in the table 2. Similarly, the accuracy of proposed system with the different system available in the market is compared, which is appended in table 3.

Table 2: Comparison of Accuracies with Various Approaches with Proposed System

S.No	NER Approach	Model Used	Algorithms Used	Used In	Accuracy
1.	Rule Based Approach		Rule based and list lookup algorithms	Recognition of Entities in Hindi Language [2]	92%
				Named entities extraction and relation extraction [46]	90%
2.	Statistical Learning Approach		Hybrid Statistical Machine Learning Approach	[60]	73.42%
3.	Hidden Markov Model	HMM	Hidden Markov Model	three ways for recall and extraction rate of entities [15]	64%
4.	Maximum Entropy Markov model	MEM	Hybrid approach with MEM	to recognize context patterns for Hindi and Bengali language [13]	65.96%
			the boosting algorithm and voted perceptron algorithm	Re ranking of entities [25]	15.6%
5	Conditional Random Field	CRF	Hybrid Approach	Named Entity Recognition [31]	68.1%
			K-Nearest Neighbor classification and linear conditional random fields under the semi supervised learning	Framework for named entities recognition [39]	75.4%
6	**Proposed Model**	**Hybrid Approach**	**Combination of Statistical Machine learning and Memory based Learning**	**For recognition of simple and composite named entities in the text**	**98.2%**

Table 3: Comparison of accuracy of various system in market with the proposed system

S.No	Name of Systems	Features/ Approaches			
		Approach	**Types of entities**	**Limitations**	**Accuracy**
1	Stanford NER	Linear chain Conditional Random Field (CRF) sequence models	Recognizes named (Person, Location, Organization, Misc), numerical (money, number, ordinal, percent), temporal (Date, time, duration, set).	Less customizability, Possible non-determinism, Problems with case less models	89.96%
2	Apache Open NLP	Rule based	Person names, locations, companies, dates, emails	Problem in tokenization	90%
3	Gate	BIO representation and CRF	Person names, locations, organization	Genre difference in entity, Capitalization is not indicative, unusual spellings, acronyms, social media conventions	**93%**
4	Spacy				**92.6%**
5	NLTK				
6	**Proposed System**	**Combination of Statistical Machine learning and Memory based Learning**	**Simple Entities (Person name, locations, organization, date, email, pin code), Composite Entities(To date, from date)**		**98.2%**

REFERENCES

[1]. 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), pp. 2300-2306.

[2]. 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, pp. 426-433.

[3]. 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, pp. 58–63.

[4]. 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, pp. 91-134.

[5]. 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, pp. 168-171.

[6]. 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, pp. 473-480.

[7]. 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, pp. 411-422.

[8]. Roy S., 2012, "Named Entity Recognition", AKGEC International Journal of Technology, 8(2), pp. 38-41.

[9]. 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), pp. 489-496.

[10]. Saha S.K., Sarkar S., Mitra P., 2009, "Feature selection techniques for maximum entropy based biomedical named entity recognition", Journal of Biomedical Informatics, 42, pp. 905-911.