International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

IDENTIFICATION AND EXTRACTION OF DOMAIN SPECIFIC ENTITIES FROM THE CORPUS DATA

M Eliazer[1], Parvathy S[2], Akshara Santharam[3], Biswas Sreya Monobikash[4]

[1]Assistant Professor, Dept. of Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India
[2,3]Student, Dept. of Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

---

***Abstract*:-** *The extraction of information from datasets has become a huge challenge due to the increase in availability of data. It becomes more challenging when information has to be extracted from specific domains for eg., like a Wealth Management System as it contains millions for data sets with same occurring words and entities. We use NLP algorithm that can trace entities in the sentence like person, location, date, time etc. The model requires a user to upload advise set documents and define attribute of mapping rules. The model scans the entire uploaded documents & identifies attributes & extracts its corresponding values.The model relies on NLP engine to keep track of the domain specific relationships.*

This paper discusses the various methods and challenges in NER as well as try to identify & extract domain specific entities from a Wealth Management System especially from large datasets.

***Key Words*: Domain Specific Entity, Named Entity Recognition, Machine Learning, Word Embeddings, Extraction, Wealth Management, Advise set documents**

1. INTRODUCTION

Named entity recognition is also called as entity extraction or entity chunking and is a major task in NLP. It is been widely used in all areas and research including medicine, finance, banking etc. Entity extraction is usually done from a text document. It often mentions many entities like people, locations, organizations, places. These contain different values like numbers, addresses, amounts, links, emails etc. All these values and entities can provide information about a particular piece of text and thus improve the overall analysis of that text document. Ex. Ram is from Chennai ,in this Ram and Chennai is named substances where a NER system has to identify RAM as a 'name of the person' and CHENNAI as the 'name of a place'. The basic steps involved in identifying entity from a raw text is given below TEXT->SENTENCE SEGMENTATION->TOKENISATION->POST TAGGING -> ENTITY DETECTION

This paper presents a review on NER methods and techniques, current and existing methods, limitations of current methods and how we can build a system or technique that can overcome the existing techniques.

2. LITERATURE REVIEW

Shuwei Wang et al.,**[1]** uses a novel method to identify named entities in financial system mainly using three steps.As first step the domain dictionary is applied to recognize stock names. Secondly the full form FNEs are identified, a classifier based on Conditional Random Field is used for this. Third, abbreviation FNE candidates are recognized using mutual information boundary entropy and context features. The approach successfully improves precision effectively, and identify named entities very effectively .

Waleed Zaghloul et al.,**[2]** has proposed to build high precision extractors for entities such as person and organization in order to train and learn in machine-learning systems for different categories and also across domains. A Rule based entity extractor is being used on 52,000 set documents for both the entities. Precision, extraction and quality of the extraction system and retrieved instances is then checked. The paper concludes that a high precision, quality extractor can be made, which can be used across knowledge domain.Generalizability is maintained in the extraction system.

Payal Biswas et al.,**[3]** develops a NER for agriculture domain namely AGNER . They used linguistic and domain specific knowledge base for developing the system. The two main process involved are Linguistic processing and Domain specific

Tagging. The research concludes by developing a entity tagger for agriculture domain, also used linguistic and domain specific knowledge base for developing the system.

Dan Klein et al.,**[4]** uses character & character n-grams NER models as an important part of data representation. Characterlevel HMM with minimal context information and maximum entropy conditional Markov model with richer context features are being employed. In this paper, it is concluded that switching from word model to character level model gave 30% less error comparing to other models. This paper also shows that character sub strings are valuable.

Vivek Kulkarni et al.,**[5]** proposes methods to detect and analyze semantic differences in word usage across multiple domains. Linguistic variations across domains is analyzed and then methods to capture domain specific semantic of word usage is done. Generic word2Vec embedding, Domain/Sense specific word embedding are employed. The baseline method used involves CoNLL-ONLY method ,Feature sub setting and online-FLORSFEMA. The method concludes by capturing domain specific entities from unlabeled text and scale well to large web scale data sets.The method also boost performance on NLP tasks like NER on domains with very little training sets as well.

Dr. K.S.Wagh et al**[6]**., proposes the use of Machine Learning and Conditional Random Field to extract named entities. The main approach in this research is to extract terms such as DNA, RNA, cell type, cell line, protein and other bio medical terms. The dataset used in this research is GENIA corpus. Collection of Medline abstracts which represents the literature of molecular biology. The research successfully concludes that named entity recognition has been completed precisely with the help of machine learning models and CRF algorithm.

Ken Yano et al.,**[7]** projects the use of "end-to-end" character-based recurrent neural network with its modality judgment by using bidirectional LSTM coupled with CRF that extracts disease named entities. The dataset used is the Japanese medical text. The paper concludes that the proposed method can recognize the disease named entities accurately with advantages like simplified processing and automatic learning of relevant features effective for Disease Named Entity (DNE) extraction.

Meizhi Ju et al.,**[8]** uses a novel dynamic neural model for nested entity recognition, without relying on external knowledge. The datasets used to support the research are GENIA and ACE2005. The paper shows the comparison of novel dynamic neural network with state-of-the-art method and concludes that proposed neural method out performs the other methods. The proposed model also gained more improvements in ACE2005 than GENIA because ACE2005 has deeper nested entities and has more nested entities.

Rohit Kumar et al.,**[9]** implemented a model which uses CRF with Kernel to produce precise results. The dataset used is CONLL-2003. The research used a model based on CRF with Kernel function and compared it with CRF model using dataset CONLL-2003 and concludes that the proposed model produces better results in terms of accuracy, precision and recall.

Hutchatai Chanlekha et al.,**[10]** proposes a Maximum Entropy Model with Simple Heuristic Information to perform Thai named entity recognition and solve the boundary problem of multi-word NE by using heuristic from rules, dictionary and statistic of word co-occurrence. The dataset used is corpus domain political news, with size 110,000 words for training task, and 25,000 words for testing. The research concludes that the results of the experiment are quite acceptable and, in the future, more powerful NE boundary detection approach and longer-distance information, to find names that are not captured by the proposed model can be incorporated.

Mai Oudah et al.,**[11]** present a paper where strengths and weaknesses of two approachesrule-based learning and machine learning based techniques along with a system of integrated methodologies is analyzed in order to develop an NER system for Arabian language that recognizes 11 entities(Person, Location, Organization, etc). The techniques were employed on the following data setsACE 2003, ACE 2004, ACE 2005, ANERcorp, ATB Part1 v 2.0 and their own corpus(consisting of list of file names, phone numbers and ISBN numbers). It was finally concluded that the hybrid approach is the best methodology as it combines strengths of both rule-based and machine learning features. The future work involves improving the system by adding more list so that they can enhance the gazetteers and rectify grammatical errors

Kozerenko E.B et al., **[12]** deals with the problem of determining and establishing between text segments consisting of semantic components under a particular domain. The datsets of this paper cannot be specified as it does not take any single domain into consideration. It further discusses how this can be used for semantic similarities between legal texts and Constitutional Court and Ruling Legal Acts using Pullenti based engine. In order to do so various legal and mass media texts are analyzed. The methods here discussed involve a hybrid approach(rule-based+ example-based approach), distributional semantic techniques

and a two-step semantic algorithm. The final conclusion drawn was that since NER technique in itself is just a part of Information extraction and Discovery, in order to make a system that powerful, it is beneficial to involve additional data structures and algorithms.

Shamima Parvez **[13]**, the focus is primarily on the improvement of issues faced in the Bengali NER system due to lack of annotated data and less accurate POS tagger. The proposed system provides a comparatively less complicated alternative to deal with this issue in three stages. The technique involves creation of a new POS tagger whose data is collected from a verified electronic medium and then HMM training and testing technique to identify the desired entities. However, there were certain hindrances in this model such as it was not highly accurate, the limited corpus availability and ability to correctly identify each entity. The future work of this paper aims towards discovering and developing a better model of evaluation to accurately perform entity recognition.

Gurinder Pal Singh Gosal **[14]** discusses the role of NER approaches in the biomedical domain. The idea of BERs is to perform information extraction, relation extraction and event extraction that is discovered biological entities in a text and discover relationships that lie between them. Biological dictionaries, Biomedical literature, Biological corpus like protein name recognizer play a very important role in order to create the tagger. There are three approaches presented in the paper- Dictionary based, supervised learning and unsupervised learning.

Bailin Wang et al., **[15]** approaches cases where entities are mentioned in a recursive or a nested structure and introduces a transition method solution to handle this. It is modeled with Stack-LSTM technique that is used to represent the state of system. The purpose of doing so is to identify the dependencies between the nested elements efficiently. Also, the character based components help to capture and analyses the words at letter-level. This proposed model is used to conduct experiments in three datasetsACE-04, ACE-05, GENIA and is said to have generated best results on Ace and relatively mediocre results in GENIA**.**

3. CHALLENGES

The limitation in the current model is that it is not suitable for large sets of data consisting of thousands of advice set documents.

(i)The confusions about whether the word in a text corpus denotes the name of a person, a location, an organization etc.

(ii)Another limitation in the current model is that it is not suitable for large sets of data consisting of thousands of advice set documents.

(iii)Precision and quality of extracted data is poor using existing models.

(iv)It takes lot of quality time in extraction and identifying process.

(v)Boundary detection is difficult to control and it hence decreases the performance of entity recognition.

(vi)The present system is single layered i.e., the entire NER process is performed at a single stage making the task more complex and thus slowing down the processing speed.

3.1 PROPOSED SYSTEM

The idea of the proposed system is to eliminate the limitations in the existing system and develop a rather more optimal version. The techniques that we plan to use are:

(i)Structure extraction – identifying fields and blocks of content based on tagging

(ii)Identify and mark sentence, phrase, and paragraph boundaries – important when doing entity extraction and NLP since they serve as useful breaks within which analysis occurs (Open source possibilities : Lucene Segmenting Tokenizer,Open NLP sentence and paragraph boundary detectors.).

(iii)Language identification –Language detectors are critical to determine what linguistic algorithms and dictionaries to apply to the text(Open source possibilities : Google Language Detector or the Optimized Language Detector).

TABLE 1- CLASSIFICATION OF THE STUDIES INVESTIGATED

Author	Methods Employed	Data sets involved
Shuwei Wang, Ruifeng Xu*,Bin Liu,Lin Gui and Yu Zhou [1]	Classifier based on conditional random fields, mutual information boundary entropy &context features	5500 datasets(from 15,000 financial news texts),85% FNEs in full-form and 15% abbreviations
Waleed Zaghloul and Silvana Trimi [2]	Rule based entity extractor	52,000 unclassified training news, reports and stories
Payal Biswas ,Aditi Sharan and Ashish Kumar [3]	Linguistic processing and Domain specific Tagging	Dataset by crawling the agriculture website http://www.agriculturalproductsindia.com/ Dataset of size 2096KB having 1,55,882 words
Dan Klein,Joseph Smarr, Huy Nguyen, Christopher D. Manning [4]	Character-level HMM with minimal context information and maximum entropy conditional markov model	English development set, german data sets and english test set
VivekKulkarni,Yashar Mehdad and Troy Chevalier [5]	Generic word2Vec embeddings ,Domain/Sense specific word embeddings, CONLL-ONLY method ,Feature sub setting and online-FLORS-FEMA .	Unlabeled data :All sentences of wikipedia,random articles from 1 million articles from Yahoo Labeled data : ConLL, Yahoo Finance, Yahoo sports
Dr. K.S.Waghet [6]	Machine Learning and Conditional Random Field	GENIA corpus
Ken Yano [7]	character-based recurrent neural network by using bidirectional LSTM coupled with CRF	Japanese medical text
Meizhi Ju [8]	novel dynamic neural model	ACE2005 , GENIA
Rohit Kumar [9]	CRF with Kernel, CRF model using dataset CONLL-2003	CONLL-2003
Hutchatai Chanlekha [10]	Maximum Entropy Model with Simple Heuristic Information	corpus domain political news, with size 110,000 words for training task, & 25,000 words for testing
Mai Oudah· Khaled Shaalan [11]	Integrated Rule based and Machine Learning Approach, Decision Tree	ANERcorp, GENIA c 3.0.2, JNLPBA, AnCora Datasets, ACE 2003, ACE 2004, ACE 2005, ATB part 1 v 2.0 , Own corpus(consisting of file names, phone numbers and ISBN numbers.)
Kozerenko E.B., Kuznetsov K.I. , Morozova Yu.I. and Romanov D.A. [12]	Hybrid approach (comprising linguistic rules and example-based learning techniques), two-step Semantic Expansion Algorithm, Distributional Semantics methods	legal and mass media texts, regulations issued by the Constitutional Court and Ruling Legal Acts.
Shamima Parvez [13]	POS tagging, Hidden Markov Modelling,	Bengali newspaper, Bengali Wikipedia, Electronic news document
Gurinder Pal Singh Gosal [14]	Bio-NLP, Dictionary based approach, Learning based approach(supervised, unsupervised, semi-supervised)	Biological dictionaries, Biomedical literature, Biological corpus like protein name recognizer
Bailin Wang, Wei Lu, Yu Wang and Hongxia Jin [15]	Transition modelling, Stack-LSTM	ACE-04, ACE-05, GENIA

5. CONCLUSION

Named Entity Recognition is a very important method in Natural Language Processing and is very useful in all domains including Finance, Medicine, Banking etc. It overcomes the existing limitations in wealth management systems like low precision and quality of extraction and difficulties in using extremely large datasets. This paper mainly talks about the existing methods for entity extraction and also defines some methods for expanding the capabilities of existing model and hence build a higher quality domain specific entity extractor.

REFERENCES

[1] Shuwei Wang,Ruifeng Xu*,Bin Liu,Lin Gui and Yu Zhou : Financial NER based on conditional random fields and information entropy Procedings of the 2014 International conference on Machine Learning and Cybernetics,Lanzhou,13-16 July ,2014

[2] Waleed Zaghloul and Silvana Trimi : Developing a innovative entity extraction method for unstructured data Zaghloul and Trimi International Journal of Quality Innovation (2017)-3:3 Published online -22 May 2017

[3] Payal Biswas, Aditi Sharan and Ashish Kumar AGNER : Entity Tagger in Agricultural Domain 2015 2nd International conference on Computing for Sustainable Global Development (INDIACom)2015-IEEE paper

[4] Dan Klien, Joseph Smarr, Huy Nyugen, Christopher D Manning- Named Entity Recognition with character-level models

[5] Vivek Kulkarni, Yashar Mehdad and Troy Chevalier - Domain Adaption for NER in online media with word embeddings arXiv:1612.00148v1 [cs.CL] 1 Dec 2016

[6] Dr. K.S.Wagh , Aishwarya Kulkarni, Shraddha Kashid, Neha Kirange, Pratiksha Pawar CRF based Bio-Medical Named Entity Recognition International Journal of Emerging Technology and Computer Science Volume: 3 Issue: 2 April – 2018;14-18

[7] Ken Yano Neural Disease Named Entity Extraction with Character-based BiLSTM+CRF in Japanese Medical Text arXiv:1806.03648v1 [cs.CL] 10 Jun 2018

[8] Meizhi Ju1,3, Makoto Miwa2,3 and Sophia Ananiadou1,3 A Neural Layered Model for Nested Named Entity Recognition Proceedings of NAACL-HLT 2018, pages 1446–1459

[9] Rohit Kumar, Priya Batta, Deepshikha Chhabra and Lokesh Pawar Improved Speech Recognition, Classification, Extraction using Conditional Random Field with Kernel Approach International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 1027-1040

[10] Hutchatai Chanlekha, Asanee Kawtrakul Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information

[11] Mai Oudah , Khaled Shaalan Studying the impact of language-independent and language-specific features on hybrid Arabic Person name recognition Journal: Natural Resources and evaluation Vol 51 Issue 2, June 2017

[12] Kozerenko E.B., Kuznetsov K.I. , Morozova Yu.I. , and Romanov D.A. Semantic Proximity Establishment in the Tasks of Knowledge Extraction and Named Entities Recognition International Conference of Artificial Intelligence(ICAI) 2017

[13] Shamima Parvez Named Entity Recognition from Bengali Newspaper Data International Journal on Natural Language Computing (IJNLC) Vol. 6, No.3, June 2017

[14] Gurinder Pal Singh Gosal A Survey of Biological Entity Recognition Approaches International Journal on Recent and innovation trends in computing and communication (IJRITCC), September 2015

[15] Bailin Wang, Wei Lu, Yu Wang and Hongxia Jin A Neural Transition-based Model for Nested Mention Recognition arxiv: 1810.01808 [cs.CL] 3 Oct 2018