IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Pattern based Named Entity Recognition using context features

Mukta S. Takalikar[1,2*], Manali M.Kshirsagar[3], Kavita R. Singh[4]

[1]Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India
[2]Department of Computer Engineering , Pune Institute of Computer Technology, Pune, India
[3]Department of Computer Technology , Yeshwantrao Chavan College of Engineering, Nagpur, India
[4]Department of Computer Technology , Yeshwantrao Chavan College of Engineering, Nagpur, India

Corresponding Author: muktapict@gmail.com, Tel.: +91 9860076259

Available online at: www.ijcseonline.org

Received: 05/Apr/2018, Revised: 10/Apr/2018, Accepted: 22/Apr/2018, Published: 30/Apr/2018

***Abstract*—** In Natural Language Processing research, Named entity recognition acts as an important tool. To improve the quality of search results, while searching through the internet , the automatic Named entity recognition(NER) and classification in the text plays very important role. Many natural language processing applications like question answering, document clustering, document summarization uses the output of Named Entity Recognition. Even today, the highly accurate Named Entity Recognition (NER) is a challenge, In this paper, a novel approach using unsupervised learning is proposed to automatically create gazette for Named entity recognition and Named entity extraction**.** The main purpose of approach is to automate the named entity recognition task, as manually recognizing named entities is cumbersome. Manually labeling so huge number of entities is effort intensive and can lead to wrong classification of entities.

***Keywords*—** Named Entity Recognition, Named Entity Extraction, Natural Language Processing, Machine Learning, Information Retrieval

I. INTRODUCTION

The named entities are units of the text that carry a well-defined semantics. The structured information referred as named entities are extracted from the unstructured text. Thus Named entity recognition and classification tasks are referred as information extraction subtask. Overall, the named entities are classified as Generic named entities and domain specific named entities[6] .The generic named entities are the names of persons, locations, organizations, PAN numbers, phone numbers ,dates etc. while for biological domain ,the domain-specific named entities are names of organisms, proteins, enzymes, genes, cells etc. Similarly for computing domain, the domain-specific named entities are technology , application domains, tools, build tools, operating systems and many more.

The effective NER systems use large amount of common-sense knowledge. With the rapid technological advancements at every instance there is some new technology, new tool, new operating system. A number of factors make automating NER with high accuracy a challenging task. The challenges are: Open nature of vocabulary, Clues such as capitalization, Overlap between NE Types, Indirect occurrences of NE, Different ways of referring to same entity. Because of complex interrelations among various parts of sentence and the variety of languages (e.g. Hindi or Marathi does not have capitalization clues), building NER systems that performs exactly as that of human is a challenge.

Descriptors or characteristic attributes of words designed for algorithmic consumption are referred as features. Named entity recognition task can use following features

- Word form and POS tags (if available)
- Orthographic features: Like capitalization, decimal, digits
- Word type patterns: Conjunction of types like capitalized, quote, functional etc.
- Bag of words: Word forms, irrespective of position
- Trigger words: Like New York **City**
- Affixes Like Hyderabad, Nagpur, Mehdipatnam, Tiruchirapally
- Gazetteer features: class in the gazetteer
- Left and right context
- Token length: Number of letters in a word
- Previous history: Classes of preceding Named Entities

Rest of the paper is organized as follows, Section II contains the related work on different approaches used for Named Entity Recognition, section III explains the methodology with considerations and main steps of the proposed algorithm, Section IV describes results and discussion , and Section V concludes research work with future directions

II. RELATED WORK

NER is treated as a classification problem with labelled training dataset as input used by the classification algorithm for the discovery of set of rules .Various supervised approaches to NER Classification Algorithm uses machine learning, pattern recognition and statistical literature. The models used are Hidden Markov Model i.e.HMM (Bikel *et al* 1999), (Seymore *et al* 1999), (Collier et al 2000), (Miller et al 1998), (Klein *et al* 2003). HMM approach has also been used for NER in languages other than English and as well for Domain Specific Named Entity recognition; e.g., biomedical domain (Shen et al 2003), (Zhang et al 2002), (Zhao 2004); Liu *et al* 2005 used HMM for identifying NE such as product names. The other models used are Maximum entropy model, Support Vector Machines, decision trees and conditional random fields[7].

In the unsupervised work Watanabe et al 2003 uses CRF to create gazetteers from Wikipedia. Jimeno *et al* 2008 compares various NER methods for automatically creating a gazetteer as well as an annotated NER corpus for disease names in medicine. Given a seed list of NE type examples, (Talukdar *et al* 2006) learns a pattern (as an automaton) from their contexts (*k* words before and after)[17]. The contexts are pruned using the IDF measure and then an automaton is induced from the context using a grammatical induction algorithm[7].

Some of the well known rule based NER systems are Univ. of Sheffield's LaSIEII (Humphreys *et al* 1998), ISOQuest's NetOwl (Krupka and Hausman 1998), Facile (Black *et al* 1998), SRA (Aone *et al* 1998) and Univ. of Edinburgh's LTG system (Mikheev *et al* 1999) and FASTUS (Appelt 1998) for English NER[7]. Recently lot of work has been done using deep learning for cross lingual named entity recognition [1,2]

Aim of the proposed work in the paper is to correctly classify the named entity instances for Indian resume. It maximizes the accuracy of classification by using non-seed based entities and generating a feature pattern. The feature patters are ranked and the feature pattern filtering is done based on feature pattern score computed [14]

III. METHODOLOGY

A. *Considerations:*

- Keeping Indian resume as template , the important 28 named entity classes are considered out of which 8 are seed based entities that are dependent on seed values and others are independent entities that do not need any seed values
- The corpus containing large number of English language resume is taken as input
- In any new resume coming as input , unknown entity instances can be there
- Contextual features are used.

B. *Description of the Proposed Algorithm:*

The proposed work consists of following main steps.

1. Pre-processing of text documents
2. Sentence boundary detection algorithm used to detect sentence boundaries
3. Configuration of Named entities and providing Seed values and forming clusters of named entities
4. Feature pattern Generation Algorithm
5. Feature pattern ranking using feature score and filtering
6. Feature pattern based gazette Creation Algorithm
7. Gazette post processing
8. Scalable implementation of Named Entity Recognition task

IV. RESULTS AND DISCUSSION

Dataset containing nearly 5000 English language resume is created and given as input to the algorithm along-with few (four to ten) seed values for seed based entities. The separate gazette for all 28 types of named entities that are typically found in Indian Resume, are created. On the basis of obtained result, the Table 1 shows count of retrieved entries against relevant entries of 'Technology" gazette. The precision for all the significant named entities is computed and is found to be satisfactory as per the plot shown in Figure 1 . The total time required for creating the separate gazette of all named entities using parallel implementation vs. sequential implementation is plotted in Figure 2 and the plot is showing the reduction in time in case of gazette creation using parallel threads.

V. CONCLUSION AND FUTURE SCOPE

As different persons write resume differently the correct named entity recognition was a challenge, but in future, the approach can be extended for any Indian language document input. Choice of accurate seed values act as the limitations of the research as the overall accuracy of algorithm is dependent on seed based named entities.

Table 1. Precision : Technology Gazette

Documents	Technology		
	Retrieved	Relevant	Percentage (%)
100	31	31	100
200	72	70	97.22
400	145	141	97.24
600	187	180	96.25
800	217	210	96.77
1000	241	226	93.77
2000	462	443	95.88
3000	653	615	94.18
4000	826	770	93.22
5000	992	950	95.76

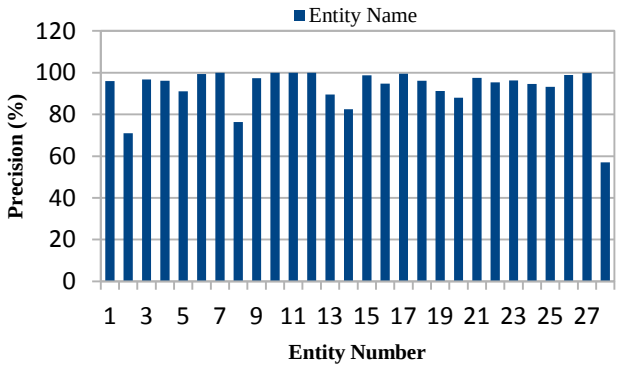

Figure 1.Precision for all significant Named Entities

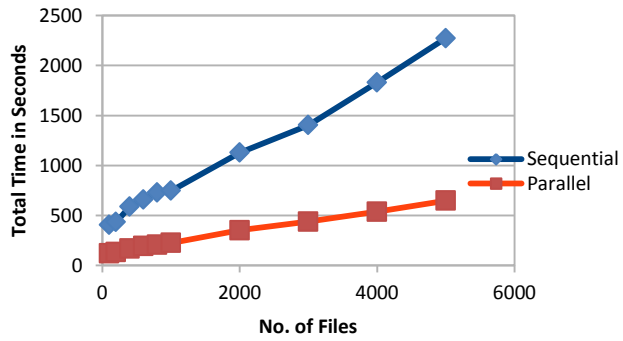

Figure 2. Parallel Implementation Vs.Sequential Implementation

ACKNOWLEDGMENT

This work would not have been possible without the support and guidance from Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India .

REFERENCES

[1] Rudra Murthy V , Pushpak Bhattacharyya," A deep Learning solution for NER" ,17th International Conference on Intelligent Text Processing and Computational Linguistics,Turkey, April 2016

[2] Rudra Murthy V , Mitesh Khapra,Dr.Pushpak Bhattacharya,"Sharing Network Parameters for Cross-lingual Named Entity Recognition ", July 2016

[3] Guillaume Lample,Miguel Ballesteros,Sandeep Subramanian, Kazuya Kawakami ,Chris Dyer, "Neural Architectures for Named Entity Recognition", Proceedings of NAACL-2016

[4] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He,Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig," Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding",Proc. IEEE/ACM transaction on audio, speech and language ,VOL. 23(2015) pp530-539

[5] Sangameshwar Patil , Sachin Pawar , Girish Palshikar, "Named entity extraction using Information distance", 1st Indian Workshop on Machine Learning, IIT Kanpur,India,2013

[6] Pawar Sachin,Rajiv Srivastava,Palshikar G.K.2012, "Automatic Gazette Creation for Named Entity Recognition and Application to Resume Processing"", In Proceedings of ACM COMPUTE 2012 Conference, Pune, India.

[7] Palshikar, G.K., 2011, "Techniques for named entity recognition: a survey", TRDDC Technical Report, pp.191-217

[8] Nadeau, D., & Sekine, S. (2007),"A survey of named entity recognition and classification. Lingvisticae Investigations", 30, 3–26. doi:10.1075/li.30.1.03nad

[9] Ekbal, A., & Bandyopadhyay, S. (2010),"Improving the performance of a NER system by post-processing and voting. In Structural, Syntactic and Statistical Pattern Recognition", LNCS 5342 (pp. 831−841), Springer

[10] Krishnarao, A., Gahlot, H., Srinet, A., & Kushwaha,D. (2009),"A comparison of performance of sequential learning algorithms on the task of named entity recognition for Indian languages", In proceedings of the International Conference on Computational Science(ICCS 2009), LNCS 5544, (pp. 123–132), Springer.

[11] Nadeau D., Turney P. Matwin S. 2006,"Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity", Proc. 19th Canadian Conf. Artificial Intelligence.

[12] Zornitsa Kozareva, "Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists" ,Proceedings of 11th Conference of European chapter of ACL,2006

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

[14] Thelen M. AND Riloff E. 2002," A bootstrapping method for learning semantic lexicons using extraction pattern contexts", Conference on Empirical Methods in natural Language Processing (EMNLP 2002).

[15] Bikel, D. M., Schwartz, R., Weischedel, R. M. (1999),"An algorithm that learns what's in a name. Machine Learning", 34, 211–231. doi:10.1023/A:1007558221122 .

[16] Collins M. AND Singer Y. 1999,"Unsupervised models for named entity classification", Proc. EMNLP,pp.100-110.

[17] Talukdar, P., Brants, T., Liberman, M., & Pereira, F. 2006,"A context pattern induction method for named entity extraction" ,In proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-2006), pp. 141–148.

Authors Profile

Mrs. Mukta Takalikar has done Bachelor of Engineering(Computer Science and Engineering) from Marathwada University in 1993 and Master of Engineering(Computer Engineering) from Pune University in year 2005. She is currently pursuing Ph.D. and working as an Assistant Professor in Department of Computer Engineering at Pune Institute of Computer technology,Pune,India. Her research interests are Multidisciplinary natural language processing, theory of computation,Compiler construction and machine learning.

Dr.Mrs. Manali Khirsagar has done Bachelor of Engineering(Computer Science) from Nagpur University in 1992 and Master of Engineering(Computer Science) from Amravati University in year 2001. She has done Ph.D.Computer Science from SHAUTS,Allahabad in 2009.Currently she is working as Professor in Department of Computer Technology at Yeshwantrao Chavan College of

Engineering, Nagpur, India. Her research interests are Data Mining, Data Warehousing, Bioinformatics.

Dr.Mrs.Kavita R.Singh has done Bachelor of Engineering, Master of Technology and aquired Ph.D. in Computer Science from SVNIT,Surat, in 2013.Currently she is working as an Associate Professor in Department of Computer Technology at Yeshwantrao Chavan College of Engineering, Nagpur, India. Her research interests are Image Processing, Database Management System, Rough Sets, Computer Vision, Soft Computing.

