International Conference on "Recent Trends in Technology and its Impact on Economy of India"
Guru Nanak College for Girls, Sri Mukstar Sahib, Punjab (India) (ICRTTIEI-17)
24th October 2017, www.conferenceworld.in ISBN: 978-93-86171-74-0

Named Entity Recognition: Applications, Approaches and Challenges

Archana Goyal[1], Manish Kumar[2], Vishal Gupta[3]

[1]PG Department of Information Technology, GGDSD College (India).

[2]Department of Computer Science and Application,

Panjab University Regional Centre, (India)

[3]University Institute of Engineering & Technology,Panjab University (India)

ABSTRACT

Named Entity Recognition and classification is the task of identifying the text of special meaning and classifying into some predetermined categories. These categories may range from person, location, organization to dates, quantities, numeric expressions etc. Named entity recognition has been an important research area since 1996. Still it is a required field to be studied due to its important use in various natural language applications. In this article, we present a comprehensive survey of named entity recognition techniques, issues and challenges and application domains. The performance of some well-known rule-based and machine learning based named entity algorithms has been analyzed. This research work will help the researchers to select the most suitable named entity recognition technique in aspecific application and will also serve as a guide to identify the areas that need attention from the researchcommunity.

Keywords: Conditional Random Field, Hidden Markov Model, Maximum Entropy, Named Entity Recognition, Natural Language Processing, Support Vector Machine

I. INTRODUCTION

Named Entity Recognition (NER) aims to recognize mentions of rigid designators from text belonging to named entity types such as persons, locations, organizations etc. [1]. It has many applications. It acts as a standalone tool for information extraction and filtering. It also plays a key role in various natural language applications such as question answering, machine translation, automatic text summarization etc.

Information Retrieval and Extraction (IREX) Program [2], Automatic Content Extraction (ACE) Program [3], Conference on Natural Language Learning 2002 and 2003 (CONLL 2002 and 2003) [4, 5] have large contribution in emergence of NER. The task of recognizing named entities was fisrst considered in Sixth Message Understanding Conference (MUC-6) [6]. At that time, MUC was focusing on Information Extraction tasks. While defining this task, they realized that identification of information units such as person, location, organization, monetary values, numeric expressions etc. are essential for information extraction task. Since that time, research on NER has been a fascinating field to be studied.

ISBN: 978-93-86171-74-0

Different approaches are given by different researchers for NER task. Earlier rule based approaches [7, 8, 9] came into existence. These approaches focus on extracting names using a number of handcrafted rules. Rule based approaches provide better results for restricted domains only. Later machine learning approaches [10, 11, 12] i.e. supervised and unsupervised approaches overcome the difficulties of rule based approaches. These approaches are easily trainable and adaptable to different domains. But these techniques require large annotated corpus for training and testing. Now-a-days, hybrid approaches [13, 14, 15] are widely used. These approaches take the advantage of both rule based and machine learning based techniques.

Besides the technique to be used, there are a number of important factors that affect the performance of NER task such as language factor, textual genres or domain factor, entity type factor etc. [1]. Most of the NER research has been done in English and other European languages. These languages provide capitalization clue for identifying named entities. But fewer studies are there for Indian languages. Textual genres or domain factor also affects the accuracy. Porting of NER system developed for one domain into another domain is a great challenge.

In the next section, we highlights various applications where NER task has been successfully used and improved the performance of the system. A comprehensive review of various approaches proposed by different researchers for the NER task is also given. In the end, we highlighted some open issues and challenges faced by different researchers while recognizing named entities from different languages and different text genres.

II. APPLICATIONS OF NAMED ENTITY RECOGNITION

Named entity recognition is an important part of various natural language applications. It is employed in a pre-processing stage which extracts proper nouns required by many natural language applications to improve their performance. In this section, we discuss some systems where NER is employed.

2.1 Information Extraction Systems

Information extraction is the task of extracting relevant information as per the request made by the user. Named entities carry important information about the text itself so employing named entity recognition system significantly improves the information extraction accuracy. A number of extraction systems i.e. Protein-Protein Interaction (PPI) Information Extraction task [16], event extraction, relation extraction systems [17 ,18] start with recognizing named entities.

2.2 Question-Answering Systems

Question answering systems aim to find out the exact answers to the natural language questions in a large document collection. Answers of fact based questions are named entities so incorporating named entity system improves the speed and accuracy of getting correct answers [19, 20, 21].

2.3 Machine Translation Systems

Machine Translation is a subfield of computational linguistics that uses computer software to convert text or speech in one natural language into another. Different translation rules are applied on named entities and other words so extraction of named entities beforehand makes the task of translation systems quite easy [22, 23].

2.4 Automatic Summarization Systems

Automatic Summarizers produce condensed representation of the input text while preserving the important information content. Named entities are important information of the text and increase the performance of identification of text segments which are further included in summarized data. [24, 25]

2.5 Semantic Annotation

Semantic annotation enriches the unstructured or semi-structured data with a context that is further linked to the structured knowledge of a domain. It is related to the formal identification of concepts and their relations. By telling a computer how data items are related and how these relations can be evaluated automatically, it becomes possible to process complex filter and search operations. This automation is implemented with information extraction techniques, among which Named Entity Recognition is used to identify concepts to annotate. [26]

Besides the systems mentioned above, named entity recognition is also used in many other Natural Language Processing applications such as ontology population [27], opinion mining [28], semantic search [29], text clustering [30] and so on.

III. CATEGORIZATION OF NAMED ENTITY APPROACHES

The named entity recognition process has a rich literature, and a number of named entity recognizers of varying flavors have been developed over the decades. Current named entity recognizers belong to one of three categories: Rule Based, Statistical, and Hybrid. The performance of different named entity recognizers is presented in Table 1.

3.1 Rule-Based Named Entity Recognizers

Nayan*et al.* (2008) [7] introduced a novel technique i.e. Phonetic Matching technique for recognizing the named entities in Indian languages. This technique does not use statistical features rather it performs matching between strings of different languages i.e. Hindi and English on the basis of their similar sounding property.

Shaalan and Raza (2009) [8]have presented a named entity recognition system for Arabic (NERA) using the gazetteer of names, regular expressions and filtering mechanism. The purpose of the filter is to revise the system output by rejecting incorrect named entities.

Gupta and Lehal (2011) [9] developed a Named Entity Recognition system for Punjabi language text summarization. The authors developed various gazetteers lists like prefix list, suffix list, middle name list, last name list and proper name list by creating a frequency list from the Punjabi Corpus. These lists are used as lexical resources for implementing condition based algorithm which consist of 5 rules like prefix rule, suffix rule, middle name rule, last name rule and proper name rule.

Singh *et al.*(2012) [31] has developed a Named Entity Recognition system for Urdu language. The authors designed various rules and gazetteers look up to find 13 named entity tags. The accuracy of this system is reported good as compared to machine learning based NER system in Urdu.

Alfred *et al.* (2014) [32] have proposed a named entity recognition algorithm for Malay articles which is based on rule based POS tagging process as well as contextual feature rules. Several dictionaries were also constructed manually to detect 3 named entities i.e. person, location and organization.

Rahem and Omar (2015) [33] have used several heuristics and grammatical rules to develop a rule based named entity recognition system for drug related crime news documents. Five different named classes such as types of drugs, price of drugs, amount of drugs, drug hiding methods, the nationality of the suspect were detected.

Quimbaya*et al.* [2016] [34] have proposed exact matching, fuzzy matching and stemmed matching technique for extracting relevant named entities (diagnosis, treatment) from electronic health records. This task was quite challenging due to various issues like inclusion of images, test results, narrative text, variety of notes, diversity of language etc. in the text.

3.2 Machine Learning Based Named Entity Recognizers

Ekbal and Bandyopadyay (2008) [10] have detected four major named entity tags such as person, location, organization and miscellaneous using support vector machine in Bengali language. The authors first converted NE tagged Bengali News Corpus [35] into BIO format and then applied Feature set which includes context word feature, word suffix feature, word prefix feature, POS feature, Digit features, NE tag of Previous word (dynamic feature), First word, various gazetteer lists.

Goyal (2008) [11] introduced a Hindi language named entity recognizer using Conditional Random Field. The design of the system is divided into three modules: NER module, NEC module, NNE module which are responsible for recognizing named entities, classifying named entities and identifying nested named entities. The authors designed such an algorithm which can be portable for other South Asian languages.

Benajiba*et al.* (2009) [12] have done their study on Arabic language and found that the use of various lexical, contextual and morphological features affects the accuracy of named entity recognition task in different machine learning algorithms such as SVM, ME and CRF.

Saha*et al.* (2010) [36] proposed a SVM based composite kernel function which is the combination of Class Association function and hierarchical word clustering kernel function [37] for Hindi and Biomedical named entity recognition task. The authors found 3 NEs i.e. person, location, organization from Hindi data and 5 NEs i.e. DNA, RNA, Protein,Cell_type, Cell_line from Biomedical data.

Jung (2012) [38] has extracted named entities from the online streaming microtext from the social networking sites. The author used the concept of microtext clustering by applying three different contextual associations: semantic association, temporal association and social association. Four NEs are found out as Person, Location, Organization and Digital IDs using Maximum Entropy (ME) approach.

The performance of NER task is badly affected by the high dimensionality of features used in identification of named entities. Saha*et al.* (2012) [39] have proposedMaxEnt and CRF based NER system for Hindi and Bengali data. The authors explored various feature reduction approaches and found that NE class association metric is best to reduce the word features as well as N-Gram, Suffix and prefix features.

Ekbal*et al.* (2012) [40] have introduced anSVM based novel technique for annotating the data i.e. an active machine learning technique for NER. This technique extracts a large number of informative sentences out of the unlabeled documents and this process continues until the output of two consecutive steps become equal. Different combination of features are used to detect 5 NEs such as Person, Location, Organization, Miscellaneous and other from Bengali and Hindi data.

Liu and Zhou (2013) [41]developed a named entity recognition system for tweets using two stages labelled approach. The tweets are manually labelled by the two independent annotators using BILOU (Beginning, Inside, Last, Other, Unit Length). Four named entities are detected such as Person, Location, Organization and Product. In the first stage, initial results are obtained by tweet level labeler. Then these pre-labelled tweets are grouped into clusters (using hierarchical clustering) by finding the similarity among them. In the second stage, cluster level labeler is used to refine the pre-labelled results. At both stages, CRF model is used for training and testing.

Chopra and Morwal (2013) [42] introduced named entity recognition system in 6,680 English words [25 files] taken from Treebank Corpus found in NLTK. The authors used Hidden Markov Model technique for identification. Total 8 named entities such as Person, Organization, Country, Magazine, Week, location, Personal Computer, Month, Other are recognized. The accuracy is measured through F-measure which is 73.8% and more than 70% of accuracy in case of names of Persons.

Bam and Shahi (2014) [43] have identified named entities from Nepali text using Support Vector machine. Various features are used to detect 5 NEs such as Person, Location, Organization, Miscellaneous and other. The system is evaluated using intrinsic performance measures (precision, recall and f-score).

Banerjee *et al.* (2014) [44]introduced Margin Infused Relaxed Algorithm (MIRA) to extract named entities from Bengali language. The author used various language independent and dependent features which are the base for the success of the algorithm. The performance of MIRA is found to be the best model due to its better optimization technique.

Keretna*et al.* (2015) [45] have presented a medical named entity recognition system based on application of their new proposed SR (segment representation) technique. In this technique, the authors assigned a new class to those words which are ambiguous. So the new SR technique is IOBESA (inside, outside, begin, end, single, ambiguous). These ambiguous words are ignored while evaluating the performance. So this technique enhances the classification accuracy. The work is done on three medical NEs as treatment, problem and test. Eight different classifiers such as Naive Bayes, CRF, ME, k-NN, Random Tree, C4.5, Ada-Boost, and Random Forest are used for evaluating the performance. Out of the 8 classifiers, k-NN reduces the performance by 0.18% while other seven classifiers show improvement in the results.

Konkol*et al.* (2015) [46] have proposed a language independent NER system based on latent semantic features. These features are of unsupervised nature so these are useful for obtaining fruitful results. These features include automatic annotation of gazetteers, getting word similarity based on semantic spaces to cluster words, use of language independent unsupervised stemming etc. This CRF based system is evaluated on English, Spanish and Dutch CONLL corpora.

ISBN: 978-93-86171-74-0

Singh and Lehal (2015) [47] have introduced named entity recognition system for Punjabi language using two machine learning techniques: HMM and MEMM. Four named entities are detected in this study i.e. Person, Location, Organization, Date/Time. Various language dependent and independent features are used by the models. The results of MEMM are better than HMM.

Bhasuran*et al.* (2016) [48] have proposed a biomedical NER which is based on a stacked ensemble approach. The authors applied several domain specific, morphological, orthographical and contextual features as well as CRF based modelling along with two fuzzy matching algorithms for extracting disease named entities. Some post processing measures are also applied to enhance the performance of the model.

Adak *et al.* (2016) [49] have performed named entity recognition from offline unstructured handwritten document images without using any linguistic resources and any explicit word/character recognition. The datasets are first pre-processed andthen structural and positional properties of NEs areextracted from the pre-processed word images. Bidirectional Long-Short Term Memory (BLSTM) neural network classifier and some post-processing heuristics are also applied to improve the performance of the NER system.

3.3 Hybrid Named Entity Recognizers

Srikanth and Murthy (2008) [13] developed a Named Entity Recognition System for Telugu language. The authors first introduced CRF based noun tagger which identifies whether the current word is noun or not-noun. After introducing CRF based Noun tagger, heuristic based NER system is developed. In this system, various seed lists or gazetteers are prepared manually i.e. List of location nouns, organization names, small lists of Prefixes, suffixes and other contextual cues. With the help of these lists, named entities are found out by applying various rules and these named entities are added up to the corresponding gazetteers to enhance the Named Entity annotated database.

Kumar P and Kiran V (2008) [14] introduced a hybrid named Entity Recognition System for five South Asian languages i.e. Bengali, Hindi, Oriya, Telugu, Urdu. The authors proved that hybrid approach gives better results as compared to statistical techniques only. CRF and HMM based hybrid model was introduced which uses several language specific rules along with many features such as prefix and suffix feature up to length of 4, POS tag, Chunk information etc. HMM based hybrid model proved to be better than CRF with lexical f-measure of 39.77%, 46.84%, 45.84%, 46.58%, 44.75%, for Bengali, Hindi, Oriya, Telugu and Urdu.

Chandhuri and Bhattacharya (2008) [15] introduced an automatic named Entity detection system for Bangla language. In this study, the authors proposed a three step approach for NE detection. The steps involved are use of NE dictionary, rule based NE detection and n-gram statistical approach. The experiments are done on 10 data sets which gives the results same as that of manual evaluation.

Guanming*et al.*(2009) [50] have introduced a hybrid Chinese NER system which detects three named entities: person, location and organization. The system is first modelled on CRF and then transformation based learning and some rules are applied to get the better results.

Ekbal and Saha (2011) [51] have proposed a classifier ensemble based approach in which different classifiers are allowed to vote for each of the output class as per their reliability to each class. For classification,

multiobjective optimization technique named AMOSA (Archived Multiobjective Simulated Annealing) has been used which raised the performance of the system. The experiments are done for three languages: Hindi, Bengali, and Telugu.

Etkinson and Bull (2012) [52] have proposed a biomedical named entity recognition system in which the output of two classifiers (SVM and HMM) are assembled to get the better results. The authors did not use any external lexical resources (dictionary, ontologies) and post processing rules.

Küçük and Yazici (2012) [53] have proposed a Turkish hybrid named entity recognizer which uses all the features of rule based recognizer [54] for the recognition task as well as this system can enhance the information resources by rote learning from annotated data. The performance is evaluated on four genres of text i.e. news text, financial news text, child stories text and historical text through 10-fold cross validation.

Chopra *et al.* (2012) [55] have developed a named entity recognition system for Hindi language using hybrid approach. The NEs identified in this study are LOC, PER, QTY, Time, ORG, Sport, River, VEH, and Month. The authors first employed rule based heuristics on total 687 entities with different tags. The correct entities identified using this approach was 325. The accuracy of rule based approach was 47.5%. After that statistical HMM technique is applied on 362 undetected entities. The results obtained using statistical approach was 89.78%. By combining both approaches, the accuracy rises to 94.61% which is better than using individual approach.

Saha and Ekbal (2013) [56] have combined the output of seven different classifiers namely Naive Bayes, Maximum Entropy, Conditional Random field, Memory Based learner, Support Vector Machine, Hidden Markov Model, and Decision Tree on the basis of multiobjective optimization (MOO) technique named as NSGA-II (Non-dominated sorting Genetic algorithm) to identify the named entities from Hindi, Bengali and Telugu data.

Keretna*et al.*(2014) [57] have introduced a hybrid model for extracting drug named entities from the unstructured and informal medical text. The experiments are done on the dataset collected from i2b2 2009 medication challenge. The authors used lexicon/dictionary based and rule based techniques together to get the desired results. Firstly, the lexicon based techniques are used only which gives F-Score of 60.3%. Later, morphological and POS rules are also employed. Morphological features include Prefix and Postfix checking rule. Drug names confused with non-medical English text are also removed while processing. It increases F-Score measure up to 66.97%.

Munkhjargal*et al.* (2015) [58] have introduced a Mongolian named entity recognizer. Mongolian language has complex structure and agglutinative morphology so it is quite difficult to extract named entities from the text. The authors used statistical techniques namely Maximum Entropy, SVM and CRF and gazetteers as well as string matching patterns in order to handle the vocabulary words.

TABLE 1: Performance of Named Entity Recognizers

Category	Authors	Experimental Results
d E n t i	Nayan*et al.* (2008) [7]	Precision for all NEs: 80.2%

		Recall: Location: 74.6%, Person: 47.4%, Organization: 42.9%
	Shaalan and Raza (2009) [8]	F-Score: Person: 87.7% , Location: 85.9%, Company: 83.15% , Date: 91.6%, Time: 95.4% , Price: 98.6% , Measurement: 97.2% , Phone No: 91.3%, ISBN: 95.3%, Filename: 96.4%
	Gupta and Lehal (2011) [9]	Precision:89.32%, Recall: 83.4%, F-Score: 86.25%
	Singh *et al.*(2012) [31]	Data Set 1: Precision: 86.17%, Recall: 90.40%, F-Score: 88.1% Data Set 2: Precision: 58.15%, Recall: 62.05%, F-Score: 60.09%
	Alfred *et al.* (2014) [32]	Precision: 85%, Recall: 94.44% , F-Score: 89.47%
	Rahem and Omar (2015) [33]	Precision: 86%, Recall: 87%, F-Score: 87%.
	Quimbayaet al. (2016) [34]	Precision: 63.0%, Recall: 57.3%, F1: 60.0%, F2: 58.3%
Machine Learning based Named Entity Recognizers	Ekbal and Bandyopadyay (2008) [10]	10 fold cross validation are experimented on different feature combinations which gives the highest F-Score: 91.8%
	Goyal (2008) [11]	Test data 1 shows accuracy using F-Score with 49.2% for maximal entities and 50.1% for nested entities. Test data 2 shows accuracy using F-Score with 44.97% for maximal entities and around 43.70% for nested entities. English dataset shows F-Score of 76.19%.
	Benajiba*et al.* (2009) [12]	F-Score: 83.34%
	Saha*et al.* (2010) [36]	SVM based kernel outperformed. F-Score: Hindi:83.56%, Biomedical domain: 67.89%
	Jung (2012) [38]	F-Score: 90.3%
	Saha*et al.* (2012) [39]	F-Score: Hindi: 85.31% (CRF) and 80.2% (MaxEnt),

		Bengali: 70.75% (CRF) and 67.54% (MaxEnt), Biomedical data: 71.56% (CRF) and 67.24% (MaxEnt)
	Ekbal*et al.* (2012) [40]	Bengali NER: Precision: 91.65%, Recall: 91.66% F-Score: 91.65% Hindi NER: Precision: 90.22%, Recall: 89.41%, F-Score: 89.81%
	Liu and Zhou (2013) [41]	Precision: 84.8%, Recall: 80.4%, F-Score: 82.5%
	Chopra and Morwal (2013) [42]	F-Score:73.8%
	Bam and Shahi (2014) [43]	Precision: 86.85%, Recall: 98.53%, F-Score:9 2.31%
	Banerjee *et al.* (2014) [44]	Precision: 89.26%, Recall: 82.99%, F-Score: 86.01%
	Keretna*et al.* (2015) [45]	All classifiers except k-NN improved the performance
	Konkol*et al.* (2015) [46]	F-Score: English: 89.44%, Spanish: 83.08%, Dutch: 83.01%, Czech: 74.08%.
	Singh and Lehal (2015) [47]	F-Score: Person: 83.46%, Location: 82.20%, Organization: 86.13%, Dtime: 91.52% (HMM) Person: 87.93%, Location: 83.32%, Organization: 89.92%, Dtime: 93.74% (MEMM)
	Bhasuranet al. (2016) [48]	F-Score: NCBI corpus -94.66%, Bio creative corpus- 84.10%
	Adak et al. (2016) [49]	F-Score: 74.59%.
Hybrid Named Entity Recognizers	Srikanth and Murthy (2008) [13]	The performance of the system lies between 80% and 97% in different experiments
	Kumar P and Kiran V (2008) [14]	HMM based hybrid model proved to be better than CRF with lexical F-Score of 39.77%, 46.84%, 45.84%, 46.58%, 44.75%, for Bengali, Hindi, Oriya, Telugu, Urdu

	Chandhuri and Bhattacharya (2008) [15]	Precision: 94.24%, Recall:85.50%, F-Score: 89.51%
	Guanming*et al.*(2009) [50]	F-Score 93.49%. The system uses less resources and takes less time for training
	Ekbal and Saha (2011) [51]	F-Score: Hindi 92.80%, Bengali 94.55%, Telugu 89.85%
	Etkinson and Bull (2012) [52]	Precision: 91% , Recall: 80.1%, F-Score: 85.14%
	Küçük and Yazici (2012) [53]	F-Score: Child Stories: 92.47%, News Text: 90.13%, Historical Text: 80.66%, Financial Text: 76.80%
	Chopra *et al.* (2012) [55]	F-Score: 94.61%
	Saha and Ekbal (2013) [56]	Hindi: Precision: 90.63%, Recall: 99.07%, F-Score: 94.66% Bengali: Precision: 94.72%, Recall: 94.21%, F-Score: 94.74% Telugu: Precision: 95.18%, Recall: 82.79% , F-Score: 88.55%
	Keretna*et al.*(2014) [57]	F-Score: 66.97%
	Munkhjargalet al. (2015) [58]	F-Score: ME- 82.72% CRF-87.36% SVM- 87.43%

IV. ISSUES AND CHALLENGES IN NAMED ENTITY RECOGNITION

As the digital data is increasing day by day on the web so it is going to be difficult to store and extract the required information with great accuracy. Extraction of named entities is also challenging task. Besides the technique to be used, there are a number of important factors that affect the performance of NER task such as language factor, textual genres or domain factor, entity type factor etc. Most of the NER research has been done in English and other European languages. These languages provide capitalization clue for identifying named entities which is a great challenge for Asian and other Indian languages. Textual genres or domain factor also affects the accuracy. NER system developed for one domain is not easy to port into another domain.

Some languages are of agglutinative nature and have different morphological structure which is a great challenge for NER. Rule based NER systems developed for such kind of languages showed better results as compared to machine learning based NER as these languages rely mainly on language specific features. NER system for tweets, microblog, SMSes is quite difficult due to their short, noisy and dynamic nature.

The performance of NER system is highly dependent on some language resources such as POS tagger, morphological analyzer, chunker, parser etc. English and European languages are rich in these resources but some Asian and other languages such as Hindi, Punjabi, Urdu, Chinese, Mongolian, and Arabicare resource poor languages, thus making the NERC task more challenging.

Supervised named entity recognition systems require a large annotated corpora for further classification of named entities out of testing data. Annotation of large training data is very time consuming and requires domain experts to perform this task. However, this issue can be resolved by the use of unsupervised methods which need small amount of labelled data as seed examples for further classification.

Ambiguity in text is a major challenge for NER. Some words have different meaning in different contexts like the word Washington is considered as a name in one context and as a location in other context. So this ambiguity needs to be resolved to get the high performance of NER. Contextual features play an important role in resolving this issue.

Nested entities is also a great challenge in named entity recognition. Entities inside other entities are called nested entities. Different authors proposed different segment representation techniques such as IO, IOB1 [59], IOB2 [60], IOE1 [61], IOE2 [61], IOBE [61], IOBES [62] to resolve this issue. However, it needs more attention.

Above all, some other challenges such as spelling variation, non-local dependencies, capitalization issue needs to be considered to get the better performance of named entity recognition system.

V. CONCLUSION

Named Entity Recognition is a long-studied technology with a wide range of natural language applications. In this article, a review of different approaches proposed by different researchers for the NER task is given. These approaches range from rule based to machine learning and hybrid approaches. Besides it, importance of named entity recognition in different natural language applications has been presented. Moreover, the challenges and open problems related to named entity recognition outlined at the end of the article will give future directions to the researchers to enhance the performance of named entity recognition methods.

REFERENCES

[1] Nadeau, D. and Sekine, S., A Survey of Named Entity Recognition and Classification. *LingvisticaeInvestigationes, 30(1), 2007, 3-26.*

[2] Satoshi, S. and Hitoshi, I., IREX: IR and IE Evaluation Project in Japanese, In *Proceedings of the 2nd International Conference on Language Resources & Evaluation*, 2000.

[3] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R., The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation, In *LREC, 2, 1,* 2004.

[4] Sang, E. F. T. K., Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition, In *Proceedings of Natural language learning, 2002, 155-158,* Association for Computational Linguistics.

[5]Sang, E. F. T. K., Meulder, F. D., Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL, 4*, 2003,142-147, Association for Computational Linguistics.

[6] Grishman, R. and Sundheim, B., Message Understanding Conference-6: A Brief History. *In COLING, 96, 1996, 466-471*.

[7] Nayan A., Rao B. R. K., Singh P., Sanyal S., Sanyal R., Named Entity Recognition for Indian Languages, *In Proceedings of the IJCNLP-08 workshop on NER for South and Sound EastAsian Languages,2008, 97-104.*

[8] Shaalan, K. and Raza, H., Arabic Named Entity Recognition from Diverse Text Types, In *Advances in Natural Language Processing*, 2008, *440-451*, Springer Berlin Heidelberg.

[9] Gupta, V. and Lehal, G. S., Named Entity Recognition for Punjabi Language Text Summarization, *International journal of computer applications, 33(3), 2011, 28-32.*

[10] Ekbal A. and Bandyopadhyay S., Bengali Named Entity Recognition using Support Vector Machine,*In Proceedings of the IJCNLP-08 workshop on NER for South and South East Asian Languages,2008, 51-58.*

[11] Goyal A., Named Entity Recognition for South Asian Languages,*In Proceedings of IJCNLP-08 workshop on NER for South and Sound East Asian Languages*, 2008, *89-96.*

[12] Benajiba, Y., Diab, M., Rosso, P, Arabic Named Entity Recognition: A Feature-Driven Study, *IEEE Transactions on Audio, Speech, and Language Processing, 17(5), 2009, 926-934*.

[13] Srikanth P. and Murthy K. N., Named Entity Recognition for Telugu. *In Proceedings of IJCNLP-08 workshop on NER for South and Sound East Asian Languages*, 2008, *41-50.*

[14] Kumar P P. and Kiran V R., A Hybrid Named Entity Recognition System for South Asian Languages,*In Proceedings of the IJCNLP-08 workshop on NER for South and Sound East Asian Languages,2008, 83-88.*

[15] Chaudhary B. B. and Bhattacharya S.,An Experiment on Automatic Detection of Named Entities in Bangla, *In Proceedings of the IJCNLP-08 workshop on NER for South and Sound East Asian Languages*, 2008, *75-82.*

[16] Danger, R., Pla, F., Molina, A. andRosso, P., Towards a Protein–Protein Interaction Information Extraction System: Recognizing Named Entities, *Knowledge-Based Systems, 57, 2014, 104-118*. ELSEVIER.

[17] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. and Weld, D. S., Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1, 2011,541-550*.

[18] Riedel, S., Yao, L. and McCallum, A., Modeling Relations and their Mentions without Labeled Text, In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2010, *148-163*. Springer Berlin Heidelberg.

[19] Mollá, D., Van Zaanen, M. and Smith, D., Named Entity Recognition for Question Answering, In *Proceedings of the Australasian Language Technology Workshop (ALTW2006), 2006, 51–58.*

[20] Srihari, R. and Li, W., Information Extraction Supported Question Answering, In*8th Text Retrieval Conference (TREC-8), 500, 1999, 185–196.*

[21] Rodrigo, Á, Pérez-Iglesias, J., Peñas, A., Garrido, G. and Araujo, L., Answering Questions about European Legislation. *Expert Systems with Applications, 40(15), 2013, 5811-5816*, ELSEVIER.

[22] Babych, B. and Hartley, A., Improving Machine Translation Quality with Automatic Named Entity Recognition, In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*,2003, *1-8*. Association for Computational Linguistics.

[23] Chen, Y., Zong, C. and Su, K. Y., A Joint Model to Identify and Align Bilingual Named Entities, *Computational linguistics, 39(2), 2013, 229-266*.

[24] Nobata, C., Sekine, S., Isahara, H. and Grishman, R., Summarization System Integrated with Named Entity Tagging and IE pattern Discovery. *In LREC, 2002, 1742-1745*.

[25] Baralis, E., Cagliero, L., Jabeen, S., Fiori, A. and Shah, S., Multi-Document Summarization based on the Yago Ontology, *Expert Systems with Applications, 40(17), 2013, 6976-6984*. ELSEVIER.

[26] Mónica M., Julián U., Sonia S. C., Jorge M. and Juan M. G. B., Named Entity Recognition: Fallacies, challenges and opportunities,*Computer Standards & Interfaces, 35, 2013, 482–489.* ELSEVIER.

[27] Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T.,Soderland, S., ...& Yates, A., Unsupervised Named-Entity Extraction from the Web: An Experimental Study, *Artificial intelligence, 165(1), 2005, 91-134*. ELSEVIER.

[28] Popescu, A. M. and Etzioni, O., Extracting Product Features and Opinions from Reviews, In *Natural language processing and text mining*,2007, 9-28. SPRINGER.

[29] Habernal, I. and KonopíK, M., SWSNL: Semantic Web Search using Natural Language. *Expert Systems with Applications*, *40*(9), 2013, 3649-3664. ELSEVIER.

[30] Cao, T. H., Tang, T. M. and Chau, C. K., Text Clustering with Named Entities: A Model, Experimentation and Realization, In *Data mining: Foundations and intelligent paradigms*, 2012, *267-287*. Springer Berlin Heidelberg.

[31] Singh, U., Goyal, V. and Lehal, G. S., Named Entity Recognition System for Urdu,*In COLING, 2012, 2507-2518.*

[32] Alfred, R., Leong, L. C., On, C. K. and Anthony, P., Malay Named Entity Recognition based On Rule-Based Approach, *International Journal of Machine Learning and Computing, 4(3), 2014, 300.*

[33] Rahem, K. R. and Omar, N., Rule-Based Named Entity Recognition for Drug-Related Crime News Documents, *Journal of Theoretical & Applied Information Technology, 77(2), 2015.*

[34] Quimbaya, A. P., Múnera, A. S., Rivera, R. A. G., Rodríguez, J. C. D.,Velandia, O. M. M., Peña, A. A. G. and Labbé, C., Named Entity Recognition over Electronic Health Records through a Combined Dictionary-based Approach, *Procedia Computer Science, 100, 2016, 55-61*. ELSEVIER.

[35] Ekbal A. and Bandyopadhyay S., A Web based Bengali News Corpus for NamedEntity Recognition. *Journal of Language Resources and Evaluation, 42 (2), 2008, 173-182.* SPRINGER.

[36] Saha, S. K., Narayan, S., Sarkar, S. and Mitra, P., A Composite Kernel for Named Entity Recognition, *Pattern Recognition Letters, 31(12), 2010, 1591-1597*. ELSEVIER.

[37] Brown P. F.,Pietra V. J. D., DeSouza P. V., Lai J. C. and Mercer R. L., Class-based ngram models of natural language,*Journal of Computational Linguistics*, 18(4), 1992, 467–479, ACM.

[38] Jung, J. J., Online Named Entity Recognition Method for Microtexts in Social Networking Services: A Case Study of Twitter, *Expert Systems with Applications, 39(9), 2012, 8066-8070*, ELSEVIER.

[39] Saha, S. K., Mitra, P. and Sarkar, S., A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition, *Knowledge-Based Systems, 27, 2012, 322-332*, ELSEVIER.

[40] Ekbal, A., Saha, S. and Singh, D., Active Machine Leaning Technique for Named Entity Recognition,*In Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2012, 180-186*, ACM.

[41] Liu, X. and Zhou, M., Two-Stage NER for Tweets with Clustering, *Information Processing & Management, 49(1), 2013, 264-273*,ELSEVIER.

[42] Chopra D. andMorwal S., Named Entity Recognition in Punjabi using Hidden Markov Model,*International Journal of Computer Science and Engineering Technology, 3(12), 2012, 616-620.*

[43] Bam, S. B. and Shahi, T. B., Named Entity Recognition for Nepali Text using Support Vector Machines, *Intelligent Information Management, 2014*.

[44] Banerjee, S., Naskar, S. K. and Bandyopadhyay, S., Bengali Named Entity Recognition using Margin Infused Relaxed Algorithm, In *International Conference on Text, Speech, and Dialogue,2014, 125-132*, Springer International Publishing.

[45] Keretna, S., Lim, C. P., Creighton, D. and Shaban, K. B., Enhancing Medical Named Entity Recognition with an Extended Segment Representation Technique, *Computer methods and programs in biomedicine, 119(2), 2015, 88-100*,ELSEVIER.

[46] Konkol, M. and Konopík, M., Segment Representations in Named Entity Recognition, In *International Conference on Text, Speech, and Dialogue, 2015, 61-70*, Springer International Publishing.

[47] Singh J. andLehal G. S., Named Entity Recognition for Punjabi Language using HMM and MEMM, *In Proceedings of 21st IRF International Conference*, 2015, *4-8.*

[48] Bhasuran, B., Murugesan, G., Abdulkadhar, S. and Natarajan, J., Stacked Ensemble Combined with Fuzzy Matching for Biomedical Named Entity Recognition of Diseases, *Journal of Biomedical Informatics, 64, 2016, 1-9.* ELSEVIER.

[49] Adak, C., Chaudhuri, B. B. and Blumenstein, M., Named Entity Recognition from Unstructured Handwritten Document Images, In *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, 375-*380*, IEEE.

[50] Guanming, Z., Chuang, Z.; Bo, X. and Zhiqing, L.,Crfs-Based Chinese Named Entity Recognition with Improved Tag Set, In *2009 WRI World Congress on Computer Science and Information Engineering, 2009.*

[51] Ekbal, A. and Saha, S., A Multiobjective Simulated Annealing Approach for Classifier Ensemble: Named Entity Recognition in Indian Languages as Case Studies, *Expert Systems with Applications, 38(12), 2011, 14760-14772.* ELSEVIER.

[52] Etkinson, J. and Bull, V., A Multi-Strategy Approach to Biological Named Entity Recognition, *Expert Systems with Applications, 39(17), 2012, 12968-12974.* ELSEVIER.

[53] Küçük, D. and Yazıcı, A., A Hybrid Named Entity Recognizer for Turkish, *Expert Systems with Applications, 39(3), 2012, 2733-2742.* ELSEVIER.

[54]Küçük, D. and Yazici, A., Rule-Based Named Entity Recognition from Turkish Texts. In *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, 2009.*

[55] Chopra D., Jahan N. and Morwal S., Hindi Named Entity Recognition by Aggregating Rule Based Heuristics and Hidden Markov Model,*International Journal of Information Sciences and Techniques, 2(6), 2012, 43-52.*

[56] Saha, S. and Ekbal, A., Combining Multiple Classifiers using Vote Based Classifier Ensemble Technique for Named Entity Recognition, *Data & Knowledge Engineering, 85, 2013, 15-39,* ELSEVIER.

[57] Karetna S., Lim C. P. and Creighton D., A Hybrid Model for Named Entity Recognition Using Unstructured Medical Text. *In Proceedings of 9thInternational Conference on System of Systems Engineering (SOSE),2014, 85-90.*

[58]Munkhjargal, Z., Bella, G., Chagnaa, A. and Giunchiglia, F., Named Entity Recognition for Mongolian Language, In *International Conference on Text, Speech, and Dialogue*, 2015, *243-251*, Springer International Publishing.

[59] Ramshaw, L. A. and Marcus, M. P., Text Chunking using Transformation-Based Learning, *Association for the Advancement of Artificial Intelligence, 1995.*

[60] Ratnaparkhi, A., *Maximum entropy models for natural language ambiguity resolution*, Doctoral dissertation, University of Pennsylvania, Philadelphia, USA, 1998.

[61] Sang, E. F. and Veenstra, J., Representing Text Chunks, In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics,1999, 173-179*, Association for Computational Linguistics.

[62] Uchimoto, K., Ma, Q., Murata, M., Ozaku, H. and Isahara, H., Named Entity Extraction based on a Maximum Entropy Model and Transformation Rules, In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics,2000, 326-335*, Association for Computational Linguistics.