EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Big Data and Named Entity Recognition Approaches for Urdu Language

Qudsia Jamil [1], Muhammad Rehman Zafar[2]

Department of Computer Science Bahria University, Islamabad, Pakistan

1 qudsi.ch@gmail.com, [2] rehmanzafar.bui@gmail.com

Abstract

Nowadays data is stored in digital form and Terabyte of data is generated on daily basis. It is difficult task to extract useful information from Big data efficiently. From unstructured text Information extraction is a technique which used to extract information. Named Entity Recognition (NER) is an essential component of information extraction in the field of Natural Language Processing (NLP). Further, Urdu language has various challenges to NER due to its agglutinative, inflectional nature and rich morphology. Therefore, NER systems for Urdu language are not mature yet due to lack of resources and ambiguities. This paper specifically addresses the different approaches to NER and explore the existing work for NER in Urdu language.

Keywords: Big Data, Named Entity Recognition, Urdu Text Processing, Natural Language Processing(NLP)

Received on 13 October 2017, accepted on 03 January 2018, published on 13 April 2018

doi: 10.4108/10.4108/eai.13-4-2018.154469

1. Introduction

In this digital era Enormous amount of data is available. Analyst has collection of unstructured, semi-structured and structured datasets called big data. The volume, complexity and rate of growth of data is very vast. To capture, manage, process and analyze this enormous amount of data is difficult task for analysts and different programming tools and applications needed to process this data. Data is available in different forms such as textual, video, image, audio, web page log files, blogs, tweets, location information, sensor data.To make intelligent and faster decision big data is very valuable for any organization. According to International Data Corporation (IDC) Big Data is: “Big Data technologies describe new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery and/or analysis”[1] For structured data, data source is Business Applications in big data e.g; retail, nance and bio informatics etc. For semi structured data, data source is Web Applications which can be web logs, email and web pages. For unstructured data, data source is Audio, Video, Images, Sensor data, Blogs and Tweets. In January 25thTwitter was available for their users in four different languages named as: Arabic, Farsi, Hebrew and Urdu1. Geo News Urdu has 1.88 million tweets 2 which are published in Urdu, BBC Urdu has 16.7K tweets3 in Urdu and Dawn news has 95.1K tweets4 published in Urdu. Almost 115 websites and blogs which are available in Urdu language 5 but news sites are actively participating in producing large amount of data in Urdu on daily basis.

2 Approaches to implement NER

NER implementation involves three main approaches and the selection is based on their corresponding efficiency. These techniques are named as ruled based, statistical and hybrid approaches. The hybrid approach which involves a combination of the previous two ruled based and statistical approaches.

2.1 Rule Based Approach

Rule based approach is applied on the textual data, whenever the system gets input in textual format, it finds named entity and compares it with the rules mentioned in the dictionary mapping and linguists. Thereafter, an output is generated by pulling each mapped individual named entity classification with each rule set of linguists [2].

2.1 Rule Based Approach

Rule based approach is applied on the textual data, whenever the system gets input in textual format, it finds named entity and compares it with the rules mentioned in the dictionary mapping and linguists. Thereafter, an output is generated by pulling each mapped individual named entity classification with each rule set of linguists [2].

2.2 Statistical Approach

Statistical approach uses machine learning techniques to implement a linguist in a semi-automatic way instead of constructing it manually. These techniques effectively extract part of speech tagging, assign categories to text and carry out sentence parsing. Furthermore, this approach employs a training module that is trained on previously constructed corpus to perform NER with term frequency calculation and context understanding.

Models in Statistical Approach

Implementation of statistical NER further involves different statistical models which employ specific statistical methodologies to train and work with corpus to achieve a required outcome.
– Hidden Markov Model (HMM) is a graphical model that uses the conditional probability distributions determined on the bases of predefined limited history i.e. the Markov property. HMM further splits into two main conditional contexts i.e. Hidden contexts that refers to latent conditions or states of related context, and the other one is observation contexts which in turn describes the related contextual observations. HMM finds its applications in parts of speech tagging, speech recognition and machine translation [5].

– Maximum Entropy Model (MEM) is the maximum entropy represents the largest entropy state of the current knowledge base of the implemented system. It does not put any assumptions or preconditions regarding data distributions in training data, hence, considered as the best for performing various kinds of experiments. It has applications in parts of speech tagging, speech recognition, NER and machine translation [6].

– Conditional Random Field (CRF) is considered as a Markov random field that was trained in a more of discriminative fashion. Distribution over mostly observed variables is not needed, in this way, more complex variables can also be added in the model. It has applications in shallow parsing, NER and gene finding work [7].

Models in Statistical

The statistical approach makes it possible to train the system for efficient NER and involves three methodologies to learn a system [8].
– Supervised Learning involves a pre-built and pre-annotated state of the art corpus to train the system according to the aforementioned models. In supervised, the learning system reads the available corpus and memorizes it to further process the text fed to the system. For a better performance, availability of a very large corpus is necessary.
– Semi Supervised Learning refers to training some initial entities termed as seeds into the system. Afterwards, the system searches for trained seeds to identify other entities within the same context.
– Unsupervised Learning is a learning environment in which a large number of entities appearing within the similar context are grouped into one unit called as a cluster. Later on system is made to learn all of these clusters and is able to identify the matching entities as in learned clusters.

2.3 Hybrid Approach

Hybrid approach involves a combination of multiple approaches such as rule based and statistical approaches to combine both machine learning techniques and manually constructed linguist rule sets. The main advantage of applying this technique is to overcome the shortcomings of the traditional approaches for NER [9].

3 Related Work

IJCNLP-08 workshop concentrated on developing NER systems for Hindi, Bengali, Oriya, Telugu and Urdu languages. In [10] author developed the Hybrid NER system for five languages. In another study conditional random field based NER is proposed in which machine learning approaches with heuristics are used. This NER system is developed for Bengali, Oriya, Telugu, Urdu and Hindi by Karhik Gali et al.2008 [11]. Moreover, Name Entity (NE) helps to extract information from text and accurate recognition and classification of NE is essential. Rule-based NE recognizer [12], New York University's MENE [13] MEM based recognizer and statistical NE

recognizer such as BBN's HMM based IdentiFinder [13] is developed. Unfortunately, these recognizers are not developed for Urdu language. For five languages including Urdu Kumar et al 2008 [14] proposed NER system using hybrid approach which includes CRF and HMM. For all languages HMM model shows better performance than hybrid CRF model. In another study Rule Based approach was used to develop a system by Kashif Riaz et al. 2010 [2] and results were more accurate than the other systems developed in IJCNLP-2008 workshop. Moreover, for Urdu language an information extraction tool was developed by Mukund et. al in [15]. This tool also contains a submodule for NER which were developed with the combination of HMM and CRF models. Rules based approach for NER in Urdu is used by Riaz [2]. Different rules constructed from 200 documents of Becker-Riaz Urdu corpus [16] is used to formulate different rules. Out of 2,262 documents 600 documents are choose. In 2012 Singh, et. al. [17] developed rules based Urdu NER system. The system used IJCNLP corpus for thirteen NEs. Test set 1 uses the 12032 tokens and Test set 2 uses 150243 tokens. In 2008 Shah, et. al. [18] developed MEM based system for the NER. This NER system focused on five Indian languages. In this study, rules and gazetteers are not built for Urdu language. Test data of 12,805 words was used and Nested, maximal and lexical precision, F-measure and recall is measured. In another study Gali, et. al. 2008 [11] two stage hybrid approach used for NER for South and South East Asian Languages which included Telugu, Hindi, Urdu, Bengali and Oriya without using any language specific resources. The dataset of 35,000 Urdu words are used to train the system. In [19] author developed a statistical NER system for Urdu by using unigram and bigram models with gazetteer lists. System used training data that contained 2313 name entities and test data contained 104 name entities. In [3] author developed a hybrid NER system with n-gram model, rules (prefix and suffix characters are used) and gazetteers for Urdu language. Author used the IJCNLP named entity and CRL named entity corpora.

4 Performance Evaluation

The NER approaches discussed in Section 3 are evaluated by calculating the precision, recall and f-measure. Moreover, approaches presented in IJNLCP 2008 workshop measured the performance of the system by dividing into three categories such as maximal matches, nested matches and lexical item matches. The maximal matches considers the longest possible match of NEs while in nested matches, the longest match of nested NEs are considered. Rest of the systems are evaluated for individual NEs tag set and the overall performance of the systems are evaluated along with gazetteer lists and handcrafted rules. Table 1 shows the overall maximum f-measure achieved by the system. The system [10], extracted fifteen different features and included the different contextual information to identify the various classes of NEs. In this approach different variations of numbers with special characters such as number with comma, hyphen, period, slash, percentage are handled. Furthermore, the system obtained the f-measure with maximum, nested and maximum lexical matches is 30.35%, 28.55% and 35.52% respectively. Furthermore, in [14] approach, authors tagged the corpus using HMM model. They used two layer approach such as statistical and rule based to extend the tool for other languages rather than Hindi which includes Urdu too. In first layer, the model is trained on annotated data and class of each word is identified. Second layer is used to validate the system with the chunks of test data and define rules the rules for each class of NEs. Further, in this study CRF is used for initial tagging with out analyzing morphology due to limited availability of the Urdu language resources. The system obtained the f-measure with maximum, nested and lexical item matches is 33.17%, 31.78% and 38.25% respectively.

Table 1. F-Meassure of NER systems for URDU language

Author	Approaches	Precision	Recall	F-measures (%)	Corpus
[18]	ME+rules	37.58	33.58	35.47	35,447 tokens
[11]	CRF	48.96	39.07	43.46	35,447 tokens
[10]	CRF	54.45	26.36	35.52	35,447 tokens
[14]	CRF+rules HHM+rules	52.35 56.21	30.13 37.15	38.25 44.73	35,447 tokens
[15]	ME	-	-	53.3	55,000 tokens
[2]	Hand crafted rules	91.5 -	90.7 -	91.10 81.60	2,262 documents 35,447 tokens
[17]	Rules based	86.17	90.40	88.1	1,50,243 tokens
[19]	Hybrid Unigram Hybrid Bigram	65.21 64.58	88.63 84.54	75.14 73.23	46019 Tokens
[3]	Hybrid Unigram Hybrid Bigram	89.57 88.8	95.57 96.76	92.47 92.65	35,447 tokens

In another relevant study [18], authors initially used ME model and than language specific rules are constructed to identify the nested entities. Moreover, for Urdu language, the POS information, language specific rules, morphological information and gazetteers are not used to tune the system due to unavailability of enough resources. However, the system obtained the f-measure with maximum, nested and maximum lexical matches is 27.79%, 28.59% and 35.47% respectively. Moreover, in [11] authors used CRF with rules to identify the NEs. The rules are constructed for Hindi and Bangali languages. Due to lack of resources and domain knowledge of other languages the rules are not constructed. By combning the rules with CRF, the system obtained the f-measure with maximum, nested and maximum lexical matches is 39.86%, 39.01% and 43.46% respectively. This approach is obtained the highest results for Urdu language between all proposed approaches in IJNLCP 2008 workshop. In study [2], authors proposed the rule based approach and used 6-gram model. The authors used finite state automata with lexical information which analyzes the states of the character. Moreover, they defined three rules based on heuristic, corpus and grammar and weighted in which order they are applied. The evaluation of this system is performed on two Urdu corpus (Becker-Riaz and IJCNLP

2008). The system obtained the 91.1% and 81.6% f-measure respectively. Further, in [15] the focus of the authors was to developed the complete tool for Urdu text analysis which also include the NER module. Further, authors used ME model with 10-fold cross validation. The system is evaluated on Computing Research Laboratory (CRL) dataset of 55,000 words and obtained the f-measure of 69.21%. Further, this system also evaluated on IJCNLP 2008 dataset and obtained the f-measure of 34.2%. The system is not performed well enough on IJCNLP 2008 dataset due to different domain of the dataset. In [19], authors used statistical model for NER. They used uni-gram and bi-gram models to estimate the probabilities of the class labels. Moreover, to tune the system they used backoff smoothing and gazetteer lists to identify the NEs. The system is evaluated on a dataset of size 2417 NEs and obtained the f-measure with uni-gram is 75.14%, with bi-gram and smoothing is 75.83%. Moreover, in another relevant study [17], authors follows the rule based approach with 12 NE tags which were used in IJCNLP 2008 workshop. Moreover, gazetteer lists also used to identify several NEs. The system is trained on two different datasets. The domain of dataset 1 is extracted from news articles related to politics while in dataset 2, science and business domains are selected. With dataset 1 and 2, the system obtained the f-measure of 60.09% and 88.1% respectively. Furthermore, in [3], authors have used hybrid multilayer approach which combines the rules, gazetteers, uni-gram and bi-gram models in order to identify the NEs. The system is evaluated on two corpora (IJCNLP 2008 and CRL) in Urdu text. By using the hybrid uni-gram model the system obtained the f-measure of 92.65% and by using hybrid bi-gram the system obtained the fmeasure of 87.6% with IJCNLP 2008 corpus. While, with CRL corpus the system obtained the 92.47% and 86.83% f-measure respectively.

5 NER Tools

There are many NER tools and libraries are available such as Stanford CoreNLP6, Apache OpenNLP7, NLTK8, NERSuite9 and many more. But unfortunately these tools are not applicable on Urdu text documents. Although some of these tools are multilingual but only applicable to English, European and some Asian languages. Unfortunately, there is no tool available publicly which identifies the NEs for Urdu language.

6 Conclusion

Urdu language has various challenges to NER due to its agglutinative, inflectional nature and rich morphology. Therefore, NER systems for Urdu language are not mature yet due to lack of resources and ambiguities. As IJCNLP dataset has not large amount of NEs, statistical approaches did not performed well as compare to rule based approaches. Here it is worth mentioning that, the machine learning approaches are trained and evaluated on IJCNLP dataset which contains only 36,000 tokens. They did not performed well due to insufficient size of corpus because machine learning approaches needs a large set of annotated data to outperform. In future we will work on NER extraction for Urdu language in parallel processing such as Spark and hadoop to attain the fast extraction and reduce storage cost with effective performance.

References

[1] Gantz, J., Reinsel, D.: Extracting value from chaos. IDC iview 1142(2011) (2011) 1–12

[2] Riaz, K.: Rule-based named entity recognition in urdu. In: Proceedings of the 2010 named entities workshop, Association for Computational Linguistics (2010) 126–135

[3] Naz, S., Umar, A.I., Razzak, M.I.: A Hybrid Approach for NER System for Scarce Resourced Language-URDU: Integrating n-gram with Rules and Gazetteers. Mehran University Research Journal of Engineering & Technology 34(4) (2015)

[4] Becker, D., Riaz, K., Bennett, B.H., Davis, E., Panton, D.: Named entity recog- nition in urdu: A progress report. In: International Conference on Internet Com- puting. (2002) 757–761

[5] Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Proceedings of the fifth conference on Applied natural language processing, Association for Computational Linguistics (1997) 194–201

[6] Borthwick, A.: A maximum entropy approach to named entity recognition. PhD thesis, Citeseer (1999)

[7] Li, W., McCallum, A.: Rapid development of hindi named entity recognition using conditional random fields and feature induction. ACM Transactions on Asian Language Information Processing (TALIP) 2(3) (2003) 290–294

[8] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1) (2007) 3–26

[9] Fresko, M., Rosenfeld, B., Feldman, R.: A hybrid approach to ner by memm and manual rules. In: Proceedings of the 14th ACM international conference on Information and knowledge management, ACM (2005) 361–362

[10] 1 Ekbal, A., Haque, R., Das, A., Poka, V., Bandyopadhyay, S.: Language independent named entity recognition in indian languages. In: IJCNLP. (2008) 33–40

[11] . Gali, K., Surana, H., Vaidya, A., Shishtla, P., Sharma, D.M.: Aggregating machine learning and rule based heuristics for named entity recognition. In: IJCNLP. (2008) 25–32

[12] . Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C.D., Stamatopoulos, P.: Rule-based named entity recognition for greek financial texts. In: Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000), Citeseer (2000) 75–78

[13] Bikel, D.M., Schwartz, R., Weischedel, R.M.: An algorithm that learns what's in a name. Machine learning 34(1-3) (1999) 211–231

[14] Kumar, P.: A hybrid named entity recognition system for south asian languages. NER for South and South East Asian Languages (2008) 83

[15] Mukund, S., Srihari, R., Peterson, E.: An information-extraction system for urdu— a resource-poor language. ACM Transactions on Asian Language Information Processing (TALIP) 9(4) (2010) 15

[16] Becker, D., Riaz, K.: A study in urdu corpus construction. In: Proceedings of the 3rd workshop on Asian language resources and international standardization- Volume 12, Association for Computational Linguistics (2002) 1–5

[17] Singh, U., Goyal, V., Lehal, G.S.: Named entity recognition system for urdu. In: COLING. (2012) 2507–2518

18. Saha, S.K., Chatterji, S., Dandapat, S., Sarkar, S., Mitra, P.: A hybrid approach for named entity recognition in indian languages. In: Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. (2008) 17–24