URDU NAMED ENTITY RECOGNITION AND CLASSIFICATION SYSTEM USING CONDITIONAL RANDOM FIELD

Muhammad Kamran Malik, Syed Mansoor Sarwar
Punjab University College of Information Technology (PUCIT), University of the Punjab, Lahore Pakistan
Corresponding Author: Kamran.malik@pucit.edu.pk

ABSTRACT: *Named Entity Recognition (NER) system for the Urdu language based on Conditional Random Field (CRF) is described. Only three Named Entities, i.e., Person, Organization and Location names, are considered to obtain results for precision, recall, and f-measure. Our system yields 63.72%, 62.30%, and 63.00% as values for precision, recall, and f-measure, respectively. These are the best-reported results for the Urdu language using any statistical model. We also identify some language independent features to show that a NER system can be developed for languages that have limited linguistic resources.*

Keywords: *Statistical NER, Indian language NER, domain independent NER, Language independent NER, URDU NER.*

1. INTRODUCTION

Urdu is written from right to left by using Arabic script in the Nastalique writing style [1]. It is the national language of Pakistan and one of the state languages of India. It has more than 60.6 million first language speakers and over 490 million total speakers in more than 22 countries [2]. Reading, writing (displaying), storing (encoding), analyzing, and synthesizing text and speech of Natural Languages (NL) are generally included in their computational aspects. Work on the computational aspects of Urdu started in early 1980s [3].

Natural Language Processing (NLP) is applied on NL to obtain computable linguistic artifacts, including Part of Speech (POS), stem, noun, and phrases. NLP is one of the challenging fields of Artificial Intelligence (AI) because natural languages are philosophical, psychological, and conceptual in nature so it is very difficult to process them. NLP has many applications including spell checker, grammar checker, sentiment analysis, and information retrieval. Some of these applications consider computational aspects and other require manipulation of conceptual and psychological knowledge.

Among other areas of NLP, intelligent information retrieval is an emerging field and the first step for most Information Extraction (IE) systems is the detection and classification of the named entities in a given text. The term Named Entity (NE) is widely used in NLP and Named Entity Recognition and Classification (NERC) system has been of interest to computational and corpus linguistics for over fifteen years. It is one of the important subtasks of IE. Much more work has been done both on the computational aspects and conceptual aspects of the English language, but very little work has been done on the Urdu Language. The need for a NERC system for Urdu is due to the sudden increase in the Urdu websites. The main feature of the Urdu text is its variability in structure, style, and vocabulary.

The task of a NERC system is to identify proper names and classify those entities according to their types, for example, names of persons, organizations, and locations. Numeric expressions including time, date, money, and percent expression are also example of entities. Instances of proper names that can be classified as person, organization, location name, amount, etc., are referred to as named entity mentions. For example, in the sentence "Pakistan is our country", the word "Pakistan" is a named entity mention and location (i.e., country in this case) is a named entity.

The term NE was first time used in the sixth Message Understanding Conference (MUC-6) [4]. MUC-6 was mostly about IE related efforts, i.e., extraction of structured information from unstructured text of a company's activities and defense related activities. In the process of performing this task, it was observed that if a software system is able to recognize information units such as proper nouns, then the system might be able to convert unstructured text into structured information.

2. LITERATURE REVIEW

The NER task has been performed on many languages, including English, German, Dutch, Spanish, French, Chinese, Italian and Hindi, as discussed in [5]. For South Asian languages, the results of NER systems are not good and efforts to develop efficient NER systems are still under investigation. IJCNLP-08[1] workshop played a vital role in the development of NER systems for Indian languages. Five languages were targeted in this workshop: Bengali, Hindi, Oriya, Telugu, and Urdu. Statistical and hybrid approaches were used for the development of the NER systems of these languages.

[6] Describes a NER system for the Hindi language. This system uses the Maximum Entropy (ME) approach to perform the NER task. The system was run on the data collected from the "Dainik Jagaran" newspaper. The accuracy of their system in terms of f-measure is 81.52%.

[7] Describes ME based Hindi NER system that tries to identify features by using the transliteration approach. In this paper in order to make English list useful for Hindi NER a two-phase transliteration approach is used and then this approach is also used to build Bengali NER to get better results. By using gazetteer list built by using transliteration based approach improve the f-measure results from 75.89% to 81.2%.

[8] Describes the application of Conditional Random Field (CRF) with feature induction to a Hindi NER system. The system discovers relevant features like word suffixes and word prefixes of length 2, 3, and 4, presence in gazetteer list, previous and next sequence by providing a large array of lexical tests and using feature induction to construct the language features that increase the conditional likelihood of NER.

[1] http://ltrc.iiit.ac.in/ner-ssea-08/

[9] Explains the results of CRF model for developing a Hindi NER system. It shows some features like POS feature, context feature and context pattern to show results. It also describes different approaches for NER. Tourism domain data was used for training and testing and training data was manually tagged in the Inside Outside Begin (IOB) format.

[10] Describes a hybrid approach to perform NER for Indian languages. The statistical approach called ME and language specific rules are used with the help of gazetteers. The system was designed and implemented for the International Joint Conference on Natural Language Processing (IJCNLP) NER shared task competition. A dataset was annotated for training using 12 types of NEs: person, designation, title-person, organization, abbreviation, brand, title-object, location, time, number, measure, and term. Several suitable features for performing the Hindi NER task were identified, including orthographic features, suffix, prefix information, morphology information, part of speech (POS) information, and context of a word (i.e., surrounding words) and their tags. Five Indian languages, Hindi, Bengali, Oriya, Telugu and Urdu were selected for the NER system. First, a baseline system was created using ME and then language specific information was added to improve NE accuracy. Data was annotated using the SSF standard and converted into the IOB format. The training data for Hindi contained more than 500,000 words, for Bengali about 160,000 words, and about 93,000, 64,000 and 36,000 words for Oriya, Telugu and Urdu, respectively. The overall accuracy of the system in terms of f-measure was 65.13%, 65.96%, 44.65%, 18.74%, and 35.47% each for Hindi, Bengali, Oriya, Telugu, and Urdu, respectively.

[11] Discusses a NER system for Bengali by combining the classifiers CRF and Support Vector Machine (SVM). Unlabeled corpus of 10 million word forms was used to generate lexical patterns. Restful show the effectiveness of the approach with values of 91.33%, 88.19%, and 89.73%, for recall, precision and f-measure respectively.

[12] Describes a voted NER system for Bengali by using unlabeled data. ME, CRF, and SVM were used to identify language independent features like contextual and orthographic word level features along with the language dependent features. Context patterns are also learned from the unlabeled data. Overall recall, precision, and f-measure values of 93.81%, 92.18% and 92.98%, respectively, show the effectiveness of this method.

[13, 14] Use a Genetic Algorithm (GA) approach to build a NER system for Bengali, Hindi, Telugu, and Oriya. [13] Uses the search capability of a GA to develop the NER system. It is assumed that each classifier differs among the various NE classes while performing prediction. An attempt is made to find the appropriate weights of voting for each class in each classifier using a GA. Results of using this approach for the four Indian languages are given in Table 1.

Table 1: Accuracies of Indian Languages

Language	Precision	Recall	F-measure
Bengali	92.08	92.22	92.15
Hindi	96.07	88.63	92.20
Telugu	78.82	91.26	84.59
Oriya	88.56	89.98	89.26

The results of Urdu NER of the IJCNLP share task competition are summarized in Table 2.

Table 2: Accuracies of Urdu Language

Language	Precision	Recall	F-measure
Baseline	51.72	18.94	27.73
Saha [7]	37.58	33.58	35.47
Gali [15]	48.96	39.07	43.46
Kumar [16]	56.21	37.15	44.73
Ekbal [17]	54.45	26.36	35.52

3. METHODS OF IMPLEMENTING NAMED ENTITY RECOGNITION AND CLASSIFICATION SYSTEM

In this section, we describe briefly different approaches described in the literature for implementing a NERC system.

3.1 Hand crafted rules

In a handcrafted system, rules are derived manually using linguistic knowledge. These rules are then used to find named entities in the system. The drawback of this approach is that it requires in-depth linguistic knowledge.

3.2 Supervised learning algorithms

In this approach, initially a large annotated corpus is developed manually. Then, different supervised machine learning algorithms, including Hidden Markov Model (HMM), Decision Trees (DT), Maximum Entropy (ME), Support Vector Machine (SVM), and Conditional Random Fields (CRF) are used to learn patterns or rules from the data.

3.3 Semi-supervised learning algorithms

The main technique used in Semi-supervised Learning algorithms are called “bootstrapping”. It involves a small degree of supervision. Initially, a set of seeds (i.e., manually annotated data) is used for starting the learning process and the system learns rules from this data. These rules are then used to annotate more data. Wrong annotations are corrected manually and corrected data is again used in learning additional rules.

3.4 Unsupervised learning algorithms

Clustering is the primary technique in unsupervised learning algorithms to identify named entities of the same types. Basically, the technique relies on lexical resources (e.g., WordNet), lexical patterns, and statistics computed on a large un-annotated corpus.

4. MATRICES USED IN NER SYSTEM

The following is the detail of metrics used in NER system.

$$\text{Recall} = \frac{\text{Correct named entities identified by the system}}{\text{Actual named entities present in the document}} * 100$$

$$\text{Precision} = \frac{\text{Correct named entities identified by the system}}{\text{Total named entities identified by the system}} * 100$$

$$\text{F-measure} = \frac{(\beta^2 + 1) * Recall * Precision}{\beta * Precision + Recall} * 100$$

β is the weight between precision and recall. Typically, β = 1 when recall and precision are evenly weighted, i.e., when β = 1, F- measure is called F1-measure. For example, suppose there are 20 named entities in a testing document. If the NERC system identifies a total of 16 named entities out of which 12 are correct named entities, then precision, recall and F-measure of the system can be calculated as follows.

Correct named entities identified by the system = 12
Total named entities identified by the system = 16
Actual named entities present in the document = 20

Recall = $\frac{12}{20}$ x 100 = 60 %

Precision = $\frac{12}{16}$ x 100 = 75%

F-measure (β = 1) = $\frac{2*0.6*0.75}{0.6+0.75}$ x 100 = 66 %

5. DETAILS OF CORPUS

The corpus for our experiment was taken from IJCNLP-08 NERSSEAL shared tasks datasets. In this task NER systems have to identify nested NEs. For example, in the University of the Punjab, Punjab is Location and University of the Punjab is Organization. NE system was required to recognize both NEs.

For annotation purpose, the first step was to identify whether a word is NE or not. For example, whether the word ”Fazal” is NE or not depends on its context. In the Urdu language, there is no concept of capitalization. Thus, in the sentence ”Us per khuda ka fazal hai (he has the blessing of God)”, fazal (blessing) is not an NE, whereas in ”fazal aik laek talibilm hai (Fazal is a competent student)”, Fazal is an NE (PERSON). The next step was to tag maximal entity. For example, ”Quaid-e-Azam Library” should be tagged as Location. It should not be marked ”Quaid-e-Azam” as Person.

In total, three persons were used in the manual tagging process. Two persons were asked to tag the corpus and if in tagging process a conflicts occurs the third person was asked to resolve the ambiguity and his decisions were considered final.

12 tags were used for tagging the dataset. The details of tagset are given below:

- NEP (Person): ’Quaid-e-Azam Muhammad Ali Jannah’, simply ’ Quaid-e-Azam ’, ’Allama Iqbal’ etc.
- NED (Designation): ’Prime Minister’, ’President’ (as in ’President Musharaf’), ’General’ (as in ’General Raheel) etc.
- NEO (Organization): ’State Bank of Pakistan’, ’DELL or ’Al Qaida’, ’The Ministry of Defense’ etc.
- NEA (Abbreviation): ’PU’ (P.U.), ’CRF’, ’AJK’, ’LTV’ etc.
- NEB (Brand): ’Pepsi’, ’Windows’ etc.
- NETP (Title-Person): ’Mr.’, ’Sir’, ’Field Marshall’ etc.
- NETO (Title-Object): ’The Seven Year Itch’, ’American Beauty’, ’1984’ (as in ’1984 by George Orwell’), ’One Hundred Years of Solitude’ etc.
- NEL (Location): ’Lahore’, ’Islamabad’, ’Punjab’ etc.
- NETI (Time): ’19 May’, ’1965’, ’6:00 pm’ etc.
- NEN (Number): ’Fifty five’, ’3.50’, ’ten lac’ etc.
- NEM (Measure): ’10 kg’, ’32 MB’, ’five years’ etc.
- NETE (Terms): ’Horticulture’, ’Conditional Random Fields’, ’Sociolinguistics’, ’The Butterfly Effect’ etc.

In CONLL 2003, only four tags Person name, Organization name, Location name and miscellaneous were used. In MUC-6, three main kinds of NE based on ENAMEX (persons, location and organization), TIMES (time expression) and NUMEX (number expression) were used, but in IJCNLP shared task, more refined targsets were defined to improve the accuracies of the MT system using NER.

6. JUSTIFICAITON OF WORK

There are many reasons that make the development of a NERC system for the Urdu language of even greater interest.

1. It is a member of the Indo-Aryan family of languages for which no high accuracy NERC system has been developed yet.
2. It is hoped that the experience of creating a high accuracy NERC system for Urdu may allow us to do the same for other Indo-Aryan languages.
3. The nature of Urdu as an Indo-Aryan language has been influenced very strongly by Persian and Arabic. It is widely spoken in India and Pakistan, and is an important minority language in the Middle East, Europe, North America, and many other parts of the world.

With the development of a high accuracy NERC system, we may be able to build better English to Urdu translation systems, questioning and answering systems, text summarization systems, and Urdu information retrieval systems.

We face the following problems in the Urdu language:

1. Unavailability of resources, including POS tagger and morphological analyzer with high accuracy for Urdu.
2. The Urdu language is highly inflectional in nature.
3. In Urdu there is no concept of capitalization, which is a major clue for NEs.
4. Urdu is a free word-order language i.e., a sentence can be written using Subject-Object-Verb or Object-Subject-Verb. For example, “Ali ne paani ka aik glass piya” and “Panni ka aik glass Ali ne piya”, both translate to “Ali drank a glass of water”.
5. The Urdu language is Agglutinative in nature, which means that by adding additional features to a word more complex words can be formed.

7. METHODOLOGY, RESULTS AND DISCUSSION

We use CRF [31] to train our Urdu NER system. IJCNLP workshop data is converted into IOE2 tagging format because it is designed to handle postpositional languages like Urdu. NER on English languages with highest accuracy used BILOU tagging scheme [27] but BILOU tagging scheme is inspired from IOB2 tagging scheme, which is designed for prepositional languages like English. The whole process can be described in following steps.

1. Urdu tagged data is selected.
2. Preprocessing is performed to convert the tagged data into the IOE2 scheme.
3. Word Normalization is performed on Urdu.
4. The CRF algorithm is used to build the training model.
5. Test data is used to predict NEs.
6. Results are generated.

The use of finer tags allows us to get better accuracies for the MT system [18]. We use the IJCNLP workshop data for training and testing purposes. For our experiment, we use only five tags: Person Name, Abbreviation, Organization, Designation and Location. The details of training and testing data are summarized in Table 3.

Table 3: Statistics about Urdu Data

NE	Training Data + Testing Data
NEP (PERSON)	365 + 145 = 510
NEA (ABBRIVATION)	39 +3 = 42
NEO (ORGANIZATION)	155 + 40 = 195
NED (DESIGNATION)	98 + 41 = 139
NEL (LOCATION)	1118 + 468 = 1586
NEs	1638 + 635 = 2273
Words	35447 + 12805 = 58252
Sentences	1508 + 498 = 2006

In our experiment, we attempt to find language independent features of Urdu to improve Urdu NER results for Person, Organization and Location. For Abbreviation and Designation we try to find language independent and language dependent rules for the improvement of Urdu NER. For our experiments, we use CRF for training and testing Urdu data and used IOE tagging scheme. Without using any feature of Urdu like word formulation, context of a word, precision, recall, and f-measure are 55.32%, 28.77%, and 37.85%, respectively for the IOE tagging scheme. Precision, recall, and f-measure after applying Bag of words + previous History + left right context are 63.72%, 62.30%, 63.00%, respectively. Rules to identify Abbreviation and Designation applied in all experiments except baseline. When the model is trained by using the word formation information, a very small increase in f-measure is observed. We repeat all the experiments after normalizing Urdu data. We perform character level normalization on Urdu data because in Urdu there are many character that can be written using different Unicode. For example آ can be written using Unicode (0627+0653) ا+ٓ or 0622.

After character level normalization we repeat our all experiments and there is significant improvements. In baseline f-measure increases from 37.85% to 39.10 %. After applying formation of words f-measure increases from 40.55% to 41.58%, after applying Bag of words + previous history f-measure raised from 55.17% to 53.72% and using Bag of words + previous history + left right context f-measure increases from 63.00% to 65.51%. The final results are shown in Table 4.

Table 4: Results of Urdu NER using IOE2 tagging

Parameter Used	Precision	Recall	F-measure
Baseline	55.32	28.77	37.85
Formation of words	42.08	39.13	40.55
Bag of words	41.83	67.93	51.77
Bag of words + Previous history	48.66	63.69	55.17
Bag of words + Previous history + Left right context	63.72	62.3	63.00
After Normalization			
Baseline	58.33	29.40	39.10
Formation of words	43.17	40.11	41.58
Bag of words	45.12	66.38	53.72
Bag of words + Previous history	51,92	65.73	58.01
Bag of words + Previous history + Left right context	64.11	66.98	65.51

8. CONCLUSION AND FUTURE WORK

In this paper, results of Urdu NER are shown without using linguistic knowledge like stemming, POS tagging, and external list. One can build a NER system with better accuracies simply by selecting appropriate features of the target language, even for resource scared languages like Urdu. To our knowledge, no one has ever carried out experimentation on Urdu data with the IOE2 scheme and supervised learning using CRF. Our methodology has produced the best-reported values for precision, recall, and f-measures for the Urdu language using any statistical model.

In future, we can use Urdu POS by [19], NP chunker by [20, 21] and behavior of the word 'kaa' [22] to show improved results. Results of Urdu NER model can be used to improve alignment of the English Urdu translation system described in [23, 24]. Case systems may be used for improving Urdu NER as mentioned in [25].

REFERENCE

1. Hussain, S. \www.LICT4D.asia/Fonts/Nafees Nastalique" *In the Proceedings of 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society*, Asian Media Information Center, Singapore, 2003.
2. BBC-Languages, A Guide to Urdu - 10 facts, key phrases and the alphabet [Internet],[cited 2012 May 2] Available from: http://www.bbc.co.uk/languages/other/urdu/guide/.
3. S. Hussain, Resources for Urdu Language Processing, *The 6th Workshop on Asian Language Resources* 99-100 (2008).
4. R. Grishman and B. Sundheim, Message Understanding Conference - 6: A Brief History, *In*

Proc. International Conference on Computational Linguistics **96** 466-471 (1996).

5. D. Nadeau and S. Sekine, A survey of named entity recognition and classi_cation, *Linguisticae Investigationes*, **30(1)** 3-26 (2007).
6. S. Saha, P. Ghosh, S. Sarkar, and P. Mitra, Named Entity Recognition in Hindi using Maximum entropy and Transliteration, *Research journal on Computer Science and Computer Engineering with Applications,* 33-41 (2008).
7. S. Saha, S. Sarkar and P. Mitra, A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition, *In Proceedings of the 3rd International Joint Conference on NLP, Hyderabad, India*, 343-349 (2008).
8. W. Li and A. McCallum, Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper), *ACM Transactions on Computational Logic,* 290-294 (2003).
9. P. Gupta and S. Arora, An Approach for Named Entity Recognition System for Hindi: An Experimental Study, *In Proceedings of ASCNT- CDAC, Noida, India,* 103-108 (2009).
10. S. Saha, S. Chatterji, S. Dandapat, S. Sarkar and P. Mitra, A Hybrid Approach for Named Entity Recognition in Indian Languages, *In Proceedings of IJCNLP Workshop on NER for South and South East Asian Languages* 17-24 (2008).
11. A. Ekbal and S. Bandyopadhyay, Improving the Performance of a NER System by Post- processing, Context Patterns and Voting, *In Computer Processing of Oriental Languages,* 45-56, (2009).
12. A. Ekbal and S. Bandyopadhyay, Voted NER system using appropriate unlabeled data, *In Proceedings of the Association for Computational Linguistics Named Entities Workshop: Shared Task on Transliteration*, 202-210 (2009).
13. A. Ekbal and S. Saha, Weighted Vote Based Classi_er Ensemble for Named Entity Recognition: A Genetic Algorithm Based Approach, *In ACM Transactions on Asian Language Information Processing* **10 (2)** (2011).
14. A. Ekbal and S. Saha, Classi_er Ensemble Selection Using Genetic Algorithm for Named Entity Recognition, *Research on Language and Computation Journal, Springer* **8 (1)** 73-99 (2010).
15. Gali, K., Surana, H., Vaidya, A., Shishtla, P., & Sharma, D. M.. Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition. *In IJCNLP* (pp. 25-32) (2008, January).
16. Kumar, P. P., & 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 South East Asian Languages* (pp. 83-88) (2008, January).
17. Ekbal, A., Haque, R., Das, A., Poka, V., & Bandyopadhyay, S. Language Independent Named Entity Recognition *in Indian Languages. In IJCNLP* (pp. 33-40) (2008, January).
18. Singh, Anil Kumar. "Named Entity Recognition for South and South East Asian Languages: Taking Stock." *IJCNLP*. (2008).
19. Kamran Malik, M., Ahmed, T., Sulger, S., Bögel, T., Gulzar, A., Raza, G., ... & Butt, M. (2010). Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. *In LREC, Seventh International Conference on Language Resources and Evaluation* (pp. 2921-2927) (*2010).*
20. S. Siddiq, S. Hussain, A. Ali, M.K. Malik, W.Ali, Urdu Noun Phrase Chunking-Hybrid Approach, *In proceedings of IEEE International Conference on Asian Language Processing* 69-72 (2010).
21. W. Ali, M.K. Malik, S. Hussain, S. Siddiq, and A. Ali, Urdu noun phrase chunking: HMM based approach, *In the proceedings of IEEE International Conference on Educational and Information Technology* 2 (2010).
22. M.K. Malik, A. Ali, and S.Siddiq, Behavior of Word `kaa' in Urdu Language, *In proceedings of IEEE International Conference on Asian Language Processing* (23-26) (2010).
23. N. Karamat,M.K. Malik, and S. Hussain, Improving Generation in Machine Translation by Separating Syntactic and Morphological Processes, *In the Proceedings of 9th International Conference on Frontiers of Information Technology (FIT)* 195-200 (2011).
24. A. Ali, S. Siddiq, and M.K. Malik, Development of Parallel Corpus and English to Urdu Statistical Machine Translation, *International Journal of Engineering and Technology IJETIJENS*, **10(05)** 31-33 (2010).
25. M. Butt and T. Ahmed. The redevelopment of Indo-Aryan case systems from a lexical semantic perspective, *Morphology* **21**, 545-572 (2011).