BUILDING NAMED ENTITY RECOGNITION AND CLASSIFICATION SYSTEM FOR RESOURCE POOR LANGUAGES USING PARALLEL CORPUS: ENGLISH-URDU AS CASE STUDY

Muhammad Kamran Malik, Aasim Ali, Khawar Mahmood
Punjab University College of Information Technology (PUCIT), University of the Punjab, Lahore Pakistan

***ABSTRACT*:** *In this paper we propose an Urdu Named Entity Recognition and Classification (NERC) system and annotate an untagged Urdu corpus using alignment mapping technique given English-Urdu parallel corpus. First four probable alignments have been considered which produce accuracies of 57%, 18%, 7% and 4% respectively.*

Keywords: *Urdu NER, Urdu Machine Translation, NER for Resource Poor languages, Application of NER.*

1. INTRODUCTION

A Named Entity Recognition and Classification (NERC) is the task of identifying and classifying the Person Name, Organization Names, Location, Date, Numbers and other entities from the text. Named entities (NEs) are central elements in texts and their correct recognition and disambiguation are an essential part of successful Information Extraction (IE), Information Retrieval (IR), Question and Answering system (Q&A), Machine Translation (MT), and so on. NER problem is solved using standard techniques like supervised machine learning algorithm, unsupervised machine learning algorithm, rule base approach and semi supervised machine learning algorithm. For most of renowned approaches, we require large amount of manually annotated data which is both expensive and time consuming. [1,2,3,4] build an efficient and good performing NERC system using a supervised learning approach with large amount of manually annotated data. The situation gets even worse for resource poor languages like Urdu in which large amount of annotated data is difficult to find and hence building NERC system for resource poor languages become very expensive and difficult. Problems of word segmentation and capitalization in Urdu along with indistinguishable Urdu proper nouns from common nouns and adjectives, a lookup approach relying on proper noun dictionaries will not work. Further, Urdu is a morphologically rich language which poses additional challenges for the NER task. E.g. Kamran is the name of a person (proper noun) in English, whereas it is either the name of a person (proper noun) or an adjective in Urdu. Similarly the word Omer is the name of the person in English, whereas it either represents a name or age in Urdu. In this paper, we propose a multilingual NERC system using alignment mapping by generating a large amount of NE annotated data of source language from target language given the parallel corpus of source and target language exists. We have considered an English-Urdu parallel corpus of 25000 sentences for experimentation purpose [5]. We use English NER system to annotate Urdu NE data. By using this technique coupled with word alignment generated by English-Urdu parallel corpus, we build a reasonably accurate Urdu NERC system. The rest of the paper is organized as follows. Section 2 discusses the Related Work. Section 3 describes Corpus Preparation mechanism and section 4 describes Building NERC System Using Parallel Corpus. Discussion and Results, Conclusion and Future work are explained in Section 5 and Section 6 respectively.

2. Related Work

Two approaches are commonly used for Machine Translation (MT) i.e. statistical approach [6,[7,8,9] and symbolic approach [10,11]. To gain insight of translation knowledge, both approaches require parallel corpus which is used to acquire both word and phrase alignment.

[12] used Chinese-English corpus for NER experimentation. After annotating 80,000 sentences of bilingual text, Chinese NER performance F-measure increased by 3%. They used two types of constrains, hard and soft. In hard constraints they assumed that each word alignment pair would have same NE and in soft constraints they assumed that one pair may have different NE.

In [13] authors proposed multi-view approach for NER on English and German sentence pairs. They used 10000 sentences for training and Europol 2006 and 2007 news wire for testing. Their bilingual model improved the German NER F-measure by 16.1%.

Others [14], used English-Bulgarian and English-Korean for NER. They used Wikipedia information to solve NER problem. [15] used bilingual corpus annotated with NER labels to improve the performance of monolingual tagger. An experiment was conducted on Chinese-English parallel corpora. Results were improved from 87.9% F-measure to 89.7%.

Others in [16] used English Chinese parallel corpora to improve the parsing accuracies using shift reduce parser.

Some [17], used Maximum Entropy (ME) approach for word alignment to generate good results for Named Entities. English Chinese parallel corpus is used for experimentation. They used four feature i.e., transliteration score, translation score, source and target NE co-occurrence score and a distortion score for distinguishing identical NEs in the same sentence.

Elsewhere [18], proposed a new approach based on the information of initially detected NE type and a constraint that the entities within aligned pair have same type. They selected English-Chines corpus to perform experiments. Results show that F-Measure of identified NE pairs was improved from 68.4% to 81.7%.

Authors in [19] used English-Chines parallel corpus to identify NE translation dictionary to improve monolingual NE annotation quality for English and Chines. They used the alignment information to improve the NE tagging of both languages.

Authors in [20] used word alignment approach with projection method to transform high-quality results of one

language to other languages. They build NE tagger from English to French and achieved good accuracies.

Others [21] tried to extract Chinese NE from parallel corpus of English-Chinese language. They proposed a framework of four components which consist of Alignment, English NER, NE Candidate Generation and Training Data selection.

3. Corpus Preparation

A parallel corpus of approximately 25000 sentences has been considered for experimentation [5]. Data has been taken from the Islamic domain of Ahadeeth. Total Urdu and English words are 500618 and 384218 respectively. We consider English as source and Urdu as target language. We generate Mapping Lexicon (ML) by using GIZA++ [22], a state-of-the-art package for the task of word alignment in the field of statistical Machine Translation. ML is a file which contains English-Urdu word mappings along with their probabilities. Some of the sample entries of ML are given in Table 1.

Table 1: Some examples of Mapping Lexicon

English word	Urdu word	Probability
Ubaidullah	اشاره	0.0312500
Ubaidullah	زائد	0.0131579
Ubaidullah	عبيدالله	0.0645161

We considered following three options to solve NERC problem using parallel corpus.

1. By using the existing NERC system like Stanford NER [23], we tag English side of ML. Then we use the mapping defined in ML to identify corresponding NE in Urdu language.
2. We define our own NEs and manually tag the English side of ML. Then we extract the corresponding Urdu side NEs using ML.
3. We define our own NEs and tag the English side of ML using list approach. We train a model on English side of ML using any of the statistical tagging approach (CRF or HMM).Then we use this model to again tag the English side of ML. Then we extract the corresponding Urdu NEs using ML

In this paper, we have adopted 2nd approach to identify NEs from our Ahadeeth data. 10 NEs have been defined based on the type of data. Tagger trained on data annotated with our defined NEs can subsequently be used for IR, IE, MT and Q&A. Due to the nature and type of our data, 1st option is not viable for NER task as most of the people are interested to know about Prophets, Angels, Books etc. Details of each NE with example are given in Table 2. To the best of our knowledge, no such NERC system exists which is trained on our defined NE annotated and Ahadeeth data. Third approach can also be used to build NERC system and result can be compared with our approach.

4. Building NERC System Using Parallel Corpus

Following are the steps for building NERC using parallel corpus.

Table 2: Statistic of Name Entities

Named Entity	Counts	Examples
Angle	82	جبریل
Book	197	قرآن
Location	1105	عراق
Day	25	جمعرات
GOD Names	75	خدا
Item	1585	جنت
Prophet	6041	اسماعیل
Sahabi	21471	اسامہ
Tribe	639	امیہ
Prayer	121	مغرب

1. In this section we define a framework for building NERC system using parallel corpus. We assume that target language is resource poor and source language is resource rich.
2. After collecting English-Urdu parallel corpus data of Quran Ahadeeth [5], we define our NEs (Allah, Prophets, Books, and Locations etc.).
3. We Assign Part of Speech (POS) tags to English text using Stanford POS tagger [24].
4. Three persons were used to manually tag and review the English text with NEs. Two persons manually tagged the data with NEs. Third person manually reviewed the annotated data and resolved ambiguities.
5. We Build word-phrase table of English-Urdu translation using parallel corpus [22].
6. We select highest probability Urdu word corresponding to English NEs from word-phrase table of English-Urdu translation. We select English NEs and find corresponding Urdu translation with highest probability from word-phrase table of English-Urdu translation.
7. We assign English NEs to highest probability Urdu word selected in previous step.
8. After the manual review of Urdu annotated data, we compare the predicted results with actual Urdu NEs to calculate accuracies.
9. We repeat step 6 to 8 multiple times. In each step, we select next highest probability Urdu word corresponding to English NEs from word-phrase table of English-Urdu translation.

5. DISCUSSION AND RESULTS

We use English-Urdu Parallel corpus to obtain word level alignment. After word alignment we considered top four probabilities to calculate the results. The main reason of using more than one probabilities (other than highest probabilities) is that if we have huge amount of parallel corpus then automatically English side NE will be aligned with Urdu side NE and the need to select multiple probabilities words will cease. In our experiment parallel corpus size is less that's why we are considering multiple probabilities for results.

We define 10 NEs and manually tag English and Urdu data as listed in table 2. When using the highest probability sequence, accuracy of Angle, Book, Location, Day, GOD Names, Items, Prophet, Sahabi, Tribe and Prayers are 34%, 55%, 57%, 36%, 67%, 56%, 58%, 57%, 66% and 40% respectively. Overall accuracy achieved is 57%. When we increase our window size for calculating accuracies from

highest probability sequence to second highest probability sequence, then accuracies increase from 34% to 67%, 55% to 73%, 57% to 73%, 36% to 72%, 67% to 89%, 56% to 73%, 58% to 81%, 57% to 80%, 66% to 78% and 40% to 60% for Angle, Book, Location, Day, GOD Names, Items, Prophet, Sahabi, Tribe and Prayers respectively. Detailed results are mentioned in table 3

Table 3: Details of results of NER

Named Entity	P1	P2	P3	P4
Angle	34%	33%	33%	0%
Book	55%	18%	0%	0%
Location	57%	16%	7%	5%
Day	36%	36%	0%	27%
GOD Names	67%	22%	0%	0%
Item	56%	17%	6%	6%
Prophet	58%	23%	9%	3%
Sahabi	57%	23%	9%	3%
Tribe	66%	12%	10%	5%
Prayer	40%	20%	20%	0%
Overall	57%	18%	7%	4%
Overall Accumulative	57%	75%	82%	86%

P1= Highest Probability, P2= Second Highest Probability,
P3 = Third Highest Probability, P4 = Fourth Highest Probability

6. CONCLUSION AND FUTURE WORK

Above experiments show that we can build NE tagged data and NERC system with reasonably good accuracies using significantly large parallel corpus of two languages, one of which is resource rich and other is resource poor.

To the best of our knowledge, no one has used NER information to improve the BLUE score of MT system from English to Urdu. By using above approach we may build parallel corpus with NEs and then using parallel corpus we build alignment model. Tagged data (Urdu in our case) as a result of this method may then be used for further NER experimentations. To do such experiments, we divide the complete Urdu tagged data into training and testing files and then by using any Machine Learning algorithm like HMM, CRF etc we can build NERC system with good accuracies.

REFERENCES

1. Zhou, G., & Su, J. (2002, July). Named entity recognition using an HMM-based chunk tagger. In*proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 473-480). Association for Computational Linguistics.
2. Chieu, H. L., & Ng, H. T. (2002, August). Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics.
3. Takeuchi, K., & Collier, N. (2002, August). Use of support vector machines in extended named entity recognition. In *proceedings of the 6th conference on Natural language learning-Volume 20* (pp. 1-7). Association for Computational Linguistics.
4. Settles, B. (2004, August). Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (pp. 104-107). Association for Computational Linguistics.
5. Ali, A., Hussain, A., & Malik, M. K. (2013). Model for english-urdu statistical machine translation. *World Applied Sciences*, *24*, 1362-1367.
6. Probst, K., & Brown, R. (2002, July). Using similarity scoring to improve the bilingual dictionary for word alignment. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*(pp. 409-416). Association for Computational Linguistics.
7. Och, F. J., & Ney, H. (2000, October). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 440-447). Association for Computational Linguistics.
8. Cherry, C., & Lin, D. (2003, July). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 88-95). Association for Computational Linguistics.
9. Wang, W., Huang, J. X., Zhou, M., & Huang, C. N. (2002, August). Structure alignment using bilingual chunking. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*(pp. 1-7). Association for Computational Linguistics.
10. Huang, J. X., & Choi, K. S. (2000, October). Chinese-Korean word alignment based on linguistic comparison. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*(pp. 392-399). Association for Computational Linguistics.
11. Ker, S. J., & Chang, J. S. (1997). A class-based approach to word alignment. *Computational Linguistics*, *23*(2), 313-343.
12. Che, W., Wang, M., Manning, C. D., & Liu, T. (2013). Named Entity Recognition with Bilingual Constraints. In *HLT-NAACL* (pp. 52-62).
13. Burkett, D., Petrov, S., Blitzer, J., & Klein, D. (2010, July). Learning better monolingual models with unannotated bilingual text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (pp. 46-54). Association for Computational Linguistics.
14. Kim, S., Toutanova, K., & Yu, H. (2012, July). Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 694-702). Association for Computational Linguistics.
15. Li, Q., Li, H., Ji, H., Wang, W., Zheng, J., & Huang, F. (2012, October). Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1727-1731).
16. Huang, L., Jiang, W., & Liu, Q. (2009, August). Bilingually-constrained (monolingual) shift-reduce parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3* (pp. 1222-1231). Association for Computational Linguistics.
17. Feng, D., Lü, Y., & Zhou, M. (2004, July). A New Approach for English-Chinese Named Entity Alignment.

In *EMNLP* (Vol. 2004, pp. 372-379).

18. Chen, Y., Zong, C., & Su, K. Y. (2010, July). On jointly recognizing and aligning bilingual named entities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 631-639). Association for Computational Linguistics.
19. Huang, F., & Vogel, S. (2002). Improved named entity translation and bilingual named entity extraction. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on* (pp. 253-258). IEEE.
20. Yarowsky, D., Ngai, G., & Wicentowski, R. (2001, March). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*(pp. 1-8). Association for Computational Linguistics.
21. Fu, R., Qin, B., & Liu, T. (2011). Generating Chinese Named Entity Data from a Parallel Corpus. In *IJCNLP* (pp. 264-272).
22. Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19-51.
23. Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363-370). Association for Computational Linguistics.
24. Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173-180). Association for Computational Linguistics.