HANDLING UNKNOWN WORDS IN NAMED ENTITY RECOGNITION USING TRANSLITERATION

Deepti Chopra[1] Sudha Morwal[2] and Dr. G.N. Purohit[3]
Department of Computer Engineering, Banasthali Vidyapith, Rajasthan, INDIA

deeptichopra11@yahoo.co.in
sudha_morwal@yahoo.co.in

ABSTRACT

Transliteration may be defined as the alteration of text from one language to another. In this paper, we have discussed how transliteration is useful in handling unknown words in Named Entity Recognition (NER). We have shown some of our results on unknown words handling in NER using transliteration.

KEYWORDS

Transliteration, NER, Unknown words, Performance Metrics

1. INTRODUCTION

Transliteration may be defined as converting word by word or letter by letter from one language to another. Transliteration is mapping alphabets or phonetic sounds of text written in one language to alphabets written in another language. Some of the applications of Transliteration include: Named Entity Recognition, Machine Translation, Information Retrieval, and Information Extraction. Named Entity Recognition (NER) is considered as one of the subtask of Information Extraction in which named entities or proper nouns are searched from a huge text and classified into various categories. Transliteration plays a crucial role in resolving the problem of unknown words in Named Entity Recognition. If a training of a given word is performed using example based learning in one language, then this word can be transliterated into other languages also.

e. g. नुसरत/PER सेब खाती है
दाापका/PER कानपुर/CITY म रहती है
दापा/PER नागपुर/CITY म रहती है

Above, Named Entities in Hindi are transliterated into English as: nusrat, deepika, Kanpur, deepa and nagpur.

2. RELATED WORK

Named Entity Recognition (NER) is the process in which Named Entities or proper nouns are identified and then categorized into different classes of Named Entity classes e.g. Name of Person, Sport, River, Country, State, city, Organization etc. The unknown words in Named Entity Recognition can be handled using many ways.

In one of the approaches, unknown word can be allotted a specific tag e.g. unk tag in order to handle unknown words in NER [1][2].
Capitalization information cannot be used to identify the POS tag, since Capitalization can exists in unknown words also. Also, Capitalization information is only restricted to identify Named Entities in English [3][4].

3. OUR APPROACH

In our approach, we have considered 'OTHER' tag to be 'Not a Named Entity' tag. Figure 1 depicts our approach during training in NER. We process each token of training data and check its tag. If the token is not tagged with 'OTHER' tag, then we transliterate the token into different languages and save these transliterated Named Entities along with their tags in a file. If the token is attached with 'OTHER' tag, then we simply process the next token in a sentence and continue this procedure until all the training sentences are not processed.

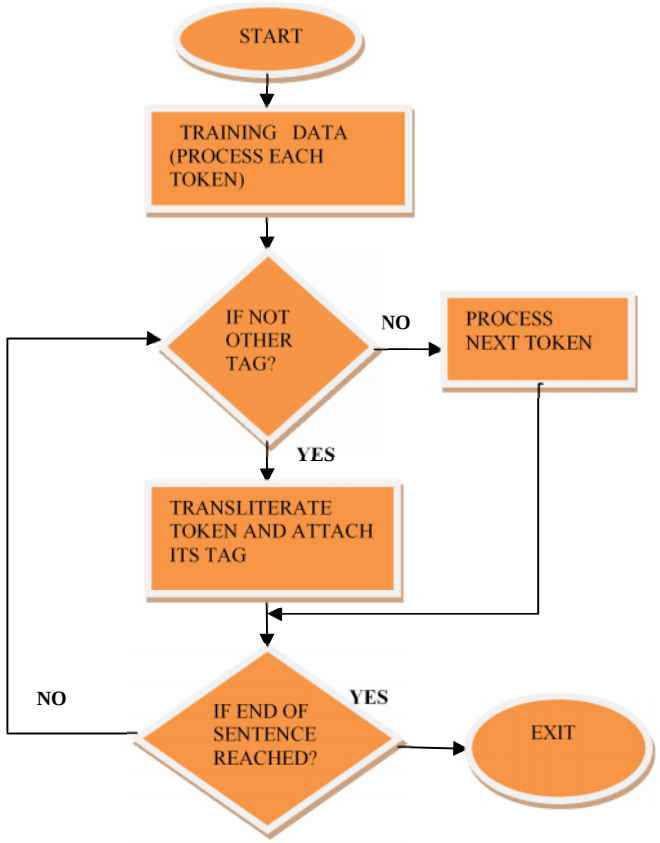

Figure 1 Steps taken while performing training in Named Entity Recognition

If we have the sound mapping of one language into any other language, then we can easily perform the transliteration task. e. g. If 'Ram/PER plays/OTHER cricket/OTHER' is a training sentence, then the transliteration task will transliterate Ram into other languages, since Ram is a Named Entity that is attached to 'PER' tag and not 'OTHER' tag. 'plays' and 'cricket' are not the Named Entities, since these are tagged by 'OTHER' tag, so transliteration is not performed on it. So, our approach is a language independent based approach, that is capable of recognizing Named Entities from a text written in any of the Natural languages, provided that it has been trained on a

Named Entity Recognition based system in any of the languages and there lie some sound mappings so that transliterated Named Entities in other languages can be achieved. Figure 2 displays the steps taken while performing testing in a Named Entity Recognition based system. In this approach, all the tokens in a testing sentence are processed one at a time. If the current token is not an unknown entity, then a tag is assigned to it using Viterbi algorithm, if Hidden Markov Model approach is used to perform Named Entity Recognition. If current token is an unknown entity, then the Transliteration file which was generated during training process is searched. If the unknown entity is found in transliteration file, then the corresponding Named Entity tag is attached to it, else 'OTHER' tag is allotted to it. This procedure continues until all the training sentences are not processed.

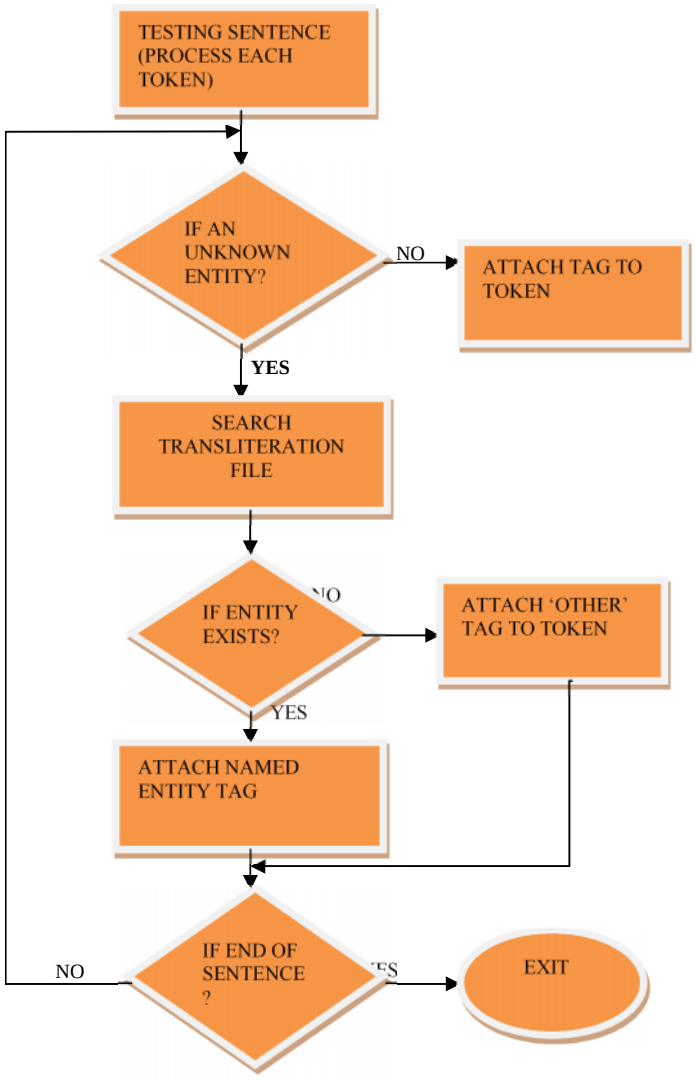

Figure 2 Steps taken while performing testing in Named Entity Recognition

Consider a testing sentence: राम क्रिकट खेलता है

If Training sentence is: Ram/PER plays/OTHER cricket/SPORT. Then, during training transliteration of Named Entities i.e. Ram and cricket takes place and the transliterated Named

Entities are stored in a transliterated file along with their tags. So, during testing राम and क्रिकट are given PER and SPORT tags respectively. खेलता and है tokens are unknown entities, since these tokens are neither trained during training process nor they exist in transliteration file.

3. RESULT

In Named Entity Recognition, handling of unknown words is a very crucial issue. We have developed a code that generates a list of Named Entities used in the training file. These Named Entities are then transliterated into different languages and stored in the respective transliterated files of different languages. TABLE 1 displays the transliteration of Named Entities in different languages obtained using our Transliteration code.

TABLE 1 Transliteration of Named Entities in different languages

ENGLISH	HINDI	FRENCH.	URDU	TELUGU	BENGALI
Ram	राम	Bélier	رام	[illegible]	[illegible]
Cricket	क्रिकट	Cricket	کرکٹ	[illegible]	[illegible]
India	भारत	Inde	بھارت	[illegible]	[illegible]
Ganga	गंगा	Ganga	گنگا	[illegible]	[illegible]

For a given testing sentence, we can initially check if all the tokens exist in already created words list. If for all the words, training has been performed, then by using Viterbi Algorithm, we can generate optimal state sequence. If one of the word is unknown that does not exists in training file, then we can check the transliterated file, if the word exists then the correct tag can be allotted to it and hence the problem of unknown word is resolved. If initially we have performed training in NER in English and during testing if we give a word in Hindi, then this would be treated as an unknown word for which we have not performed training yet. So, we can generate the transliterated list of Named Entities from English to other languages and handle the problem of unknown words in NER. TABLE 2 displays our results obtained while performing NER on multilingual corpus. The Training and Testing file include sentences from the following languages: English, Hindi, Marathi, Punjabi and Urdu. We have performed training on 142 tokens and testing on 212 tokens. Recall, Precision and F-Measure obtained are: 95.80%, 96.3% & 96.04%.

TABLE: 2 Unknown words handling in NER having 28 sentences in Training file

Tag set	{Person, City, Sport}
Number of Training sentences	28 (142 tokens)
Number of Testing Sentences	41(212tokens)
Number of sentence	0

tagged wrongly				
According to human frequency of tags (Correct Answe r)	42	6	1	163
Observed (System provided) frequency of tags (Experimental Results)	41	3	1	167
TAG SET	PER	City	Sport	Not-a- name
Recall(R)	R=Correct/(Correct+Incorrect+Spurious)=209/(209+8+1)=95.80%			
Precision(P)	P=Correct/(Correct+Incorrect+Missing)=209/(209+8+0)=96.3%			
F-measure	F-Measure = (2*P*R)/(P+R)=(2*96.3*95.8)/(96.3+95.8)=96.04%			

TABLE 3 displays the list of Named Entities on which transliteration is performed in different languages and then these transliterated Named Entities are stored in a file along with their tags that can be used further to resolve the problem of unknown entities in a Named Entity Recognition.

TABLE 3 Named Entities transliterated into different languages

SNO	TAGS
1	PER (Name of Person)
2	OZ (Name of Organization)
3	CN (Name of Country)
4	MAG (Name of Magazine)
5	WEEK (Name of Week)
6	LOC (Name of Location)
7	PC (Name of Personal Computer)
8	MONTH (Name of Month)
9	CITY (Name of City)
10	ST (Name of State)
11	SPORT (Name of Sport)

4. PERFORMANCE METRICS

Performance Metrics is measure to estimate the performance of a NER based system. Performance Metrics can be calculated in terms of 3 parameters: Precision, Accuracy and F-Measure. The output of a NER based system is termed as "response" and the interpretation of human as the "answer key" [5]. Consider the following terms:

1. Correct-If the response is same as the answer key.
2. Incorrect-If the response is not same as the answer key.
3. Missing-If answer key is found to be tagged but response is not tagged.
4. Spurious-If response is found to be tagged but answer key is not tagged. [6]

Hence, we define Precision, Recall and F-Measure as follows: [7][8][9]
Precision (P): Correct / (Correct + Incorrect + Missing)
Recall (R): Correct / (Correct + Incorrect + Spurious)
F-Measure: (2 * P * R) / (P + R)

5. CONCLUSION

In this paper, we have discussed about Transliteration. We have given our results obtained while performing Named Entity Recognition on multilingual corpus and handling unknown words using transliteration approach. We have also discussed about Performance Metrics which is a very important measure to judge the performance of a Named Entity Recognition based system.

ACKNOWLEDGEMENT

I would like to thank all those who helped me in accomplishing this task.

REFERENCES

[1] http://www.cse.iitb.ac.in/~cs626-460-2012/seminar_ppts/ner.pdf
[2] http://staff.um.edu.mt/cabe2/publications/nfHmms.pdf
[3] http://acl.ldc.upenn.edu/acl2002/MAIN/pdfs/Main036.pdf
[4] http://www.sigkdd.org/explorations/issues/7-1-2005-06/4-Fu.pdf
[5] Shilpi Srivastava, Mukund Sanglikar & D.C Kothari. "Named Entity Recognition System for Hindi Language: A Hybrid Approach" International Journal of Computational Linguistics (IJCL), Volume (2) : Issue (1) : 2011.Available at http://cscjournals.org/csc/manuscript/Journals/IJCL/volume2/Issue1/IJCL-19.pdf
[6] B. Sasidhar, P. M. Yohan, Dr. A. Vinaya Babu3, Dr. A. Govardhan,."A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.
[7] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay "Language Independent Named Entity Recognition in Indian Languages" .In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 33–40,Hyderabad, India, January 2008.Available at: http://www.mt-archive.info/IJCNLP-2008-Ekbal.pdf
[8] Darvinder kaur, Vishal Gupta. "A survey of Named Entity Recognition in English and other Indian Languages" .IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.
[9] G.V.S.RAJU, B.SRINIVASU, Dr.S.VISWANADHA RAJU, 4K.S.M.V.KUMAR "Named Entity Recognition for Telugu Using Maximum Entropy Model"

About Authors

Deepti Chopra is working as Assistant Professor in the Department of Computer Science at Banasthali University (Rajasthan), India. She has received B.Tech degree in Computer Science and Engineering from Rajasthan College of Engineering for Women, Jaipur, Rajasthan in 2011.She has done M.Tech in Computer Science and Engineering from Banasthali University, Rajasthan in 2013. Her research interests include Artificial Intelligence, Natural Language Processing, and Information Retrieval. She has published many papers in International journals and conferences.

Sudha Morwal is an active researcher in the field of Natural Language Processing. Currently working as Associate Professor in the Department of Computer Science at Banasthali University (Rajast han), India. She has done M.Tech (Computer Science) , NET, M.Sc (Computer Science) and her PhD is in progress from Banasthali University (Rajasthan), India. She has published many papers in International Conferences and Journals.

Dr. G. N. Purohit is a Professor in Department of Mathematics & Statistics at Banasthali University (Rajasthan). Before joining Banasthali University, he was Professor and Head of the Department of Mathematics, University of Rajasthan, Jaipur. He had been Chief-editor of a research journal and regular reviewer of many journals. His present interest is in O.R., Discrete Mathematics and Communication networks. He has published around 40 research papers in various journals