Improving the Quality of Machine Translation through Proper Transliteration of Name Entities

Deepti Bhalla
Department of Computer Science, IEC College of Engineering & Technology, Greater Noida, India
deeptibhalla0600@gmail.com

Nisheeth Joshi and Iti Mathur
Department of Computer Science, Banasthali University, Rajasthan, India
{nisheeth.joshi, mathur_iti}@rediffmail.com

***Abstract*—Machine Translation is the study of system that translates the given input text in source language to the output text in the target language. The source language and the target language are Natural Languages. Machine Translation is a very difficult problem; especially name entity translation has always been a challenge for the machine translators because of different spelling variations in translation of name entities. In this paper we focus on improving the quality of our machine translated output. For this, at first we recognize the name entities and then transliterate them. We have calculated the phoneme based N-Gram Probabilities for all the name entities. Using these probabilities we are transliterating our name entities from English to Punjabi through syllabification in which we divide the word in to syllables.**

***Index Terms*—transliteration, translation, MT quality, punjabi, english, name entity.**

I. Introduction

Machine Translation is a very interesting and challenging field as often it gets very difficult to obtain the correct translations, if our system is not able to recognize the name entities properly. In the area of natural language processing it is very important to recognize the name entities that belong to certain predefined classes such as person, location, organization, percent, money, time and date which are called name entity recognition. Name entity recognition has many applications in the area of natural language processing such as information extraction, question answering, medical sciences and bioinformatics. The incorrect translations of name entities not only abjure the meaningful context of language but also deteriorate the quality of translated output. Because of this reason transliteration of name entity proves to be very beneficial in the field of machine translation. Machine Transliteration has come out to be an emerging research area in the field of natural language processing. Our proposed work improves the quality of machine translated output to a great extent. For this we have used syllabification based approach. For example, if we have a name entity 'ray' it should be translated to '**ਰੇ**' in Punjabi instead of '**ਕਿਰਣ**'.

The remainder of this paper is organized as follows: In section II we described related work. Section III gives a brief introduction about script of Punjabi language. Section IV explains our approach and experimental set up. Section V shows evaluation and results. Finally in section VI we have concluded the paper.

II. Related Work

A lot of work has been carried out in the field of machine translation and machine transliteration. Singh et al. [1] presented statistical analysis of syllables and have shown that how this statistical analysis will proves to be helpful in selection of syllables for the speech database. Gupta et al. [2] implemented a condition based named entity recognition algorithm. For condition based system they have done in-depth analysis of output of over 50 Punjabi news documents. Kamal Deep and Goyal [3] have addressed the problem of transliterating Punjabi to English language by using rule based approach. They have proposed transliteration scheme using various approaches such as Grapheme based approaches (G), phoneme-based approaches (YP) and hybrid approaches (H) to model the transliteration problem. The hybrid technique achieved accuracy of 93.22%. Babych and Hartley [4] presented the result of an experiment in which Machine Translated input was processed using output from the name entity recognition module of sheffield's GATE information extraction system. They have shown the result which indicates that combining IE technology with Machine translation has a great potential for improving the state of art in output quality. Yasar AL-Onaizan and Knight [5] presented a novel algorithm for translating Name Entity phrases using easily obtainable monolingual and bilingual resources and the results obtained are compared with the results obtained from human translations and a commercial system for the same task. Hassan et al. [6] evaluated the quality of the

doi:10.4304/jetwi.6.3.354-358

extracted translation pairs by showing that it improves the performance of a named entity translation system.

They used this approach to build a large dictionary of Arabic/English named entity translation pairs. Alek et al. [7] presented an approach to improve machine translation of named entities by using Wikipedia. In this the author addressed various problems by using Wikipedia to translate NEs and presented to the MT system. They experimented with English to Czech translation. The translation of a NE is done by looking up the Czech version of the English Wikipedia article about the named entity. Kamal deep et al. [8] proposed hybrid (statistical +rules) approach for Transliteration from Punjabi to English. The Authors used letter to letter mapping as baseline and tried to find out the improvements by using statistical methods. In this paper they have discussed the Punjabi to English Machine transliteration system which is forward transliteration system.They achieved an accuracy of 95.23%. Gupta et al. [9] have shown the previous work done in English and other European languages. A survey is given on the work done in Indian Languages i.e. Telugu, Hindi, Bengali, Oriya and Urdu. They have listed and categorized the features that are used in recognition of NE and also provided an overview of the evaluation methods that are in the use of NER accuracy. Joshi et al. [10] showed a method to transliteration of text using frequency based technique for English-Hindi language pair. They showed that using this technique an effective database can be made which can be used as an input system. Their system could capture non standardized Hindi spellings. Their approach was based on edit distance based methodology and attained an accuracy of 66%. Joshi and Mathur [11] showed an approach of transliterating text using phonetic mapping. They created a table for all possible phonemes for English-Hindi language pair. This approach was also based on edit distance based methodology and attained an accuracy of 68%. Bhalla et al.[12] showed the experiment using the statistical machine translation tool. They have shown the extraction of syllables from the input text and transliteration of name entities from source text to target text.They have also calculated the probabilities with the help of MOSES toolki and attained the accuracy of 88.19%. Ameta et al.[13] have worked on improving the quality of Gujrati-Hindi Machine Translation. They have proposed the use of Stemming and Part of Speech Tagging for improving the quality of Translation. Bhalla et al.[14] have worked on improving the quality of MT output using novel name entity translation scheme. They have calculated the probabilities for all the syllables that have been extracted with the help of a syllabification algorithm. They experimented with the test set and acheived the accuracy of 86.98%. Their experimental results have shown that recognition of name entities can prominently improves the performance of machine translation from English to Punjabi.

III. Analysis of Punjabi Script

Punjabi is one of the Indo-Aryan languages which are mainly used in regions of Punjab. Its script is Gurumukhi which is based on Devanagri and is written from left to right. In Gurumukhi all consonants are followed by an ‘a’ sound. According to Gurumukhi script, Punjabi language has 38 consonants and 19 vowels (non-nasal vowels and nasal vowels). Some independent vowels can be constructed using the three basic characters Ura (ੳ), Aira (ਅ) and Iri (ੲ). Punjabi is also one of the constitutional languages of India and official language of Punjab. It is world’s 14th widely spoken language. Along with Punjab it is also spoken in neighboring states such as Haryana, Delhi and Himachal Pradesh.

IV. Approach And Experimental Setup

In our work for the transliteration of name entities we have used English-Punjabi parallel corpus of name entities. We have calculated translation probabilities based on relative frequency. In this we are transliterating the name entities using syllabification and for this we are using rule based approach. Here we are testing our system whether it generates the correct transliteration for the recognized name entities.

Phoneme is the basic unit of phonology and combination of these phonemes led to higher unit i.e. syllable. The syllable is the combination of vowel and consonants. A syllable should have a vowel because without vowel a syllable cannot exist. Some possible combinations of vowel and consonants are V, CV, VC, VCC, CVC, CCVC and CVCC (where V and C represent vowel and consonants respectively). Almost all the languages have VC, CV or CCVC structures so we are using these vowel and consonants as the basic phonological unit. Table I shows some possible combinations for syllable extraction.

After the name entities are syllabified they will be transliterated to their target text. Our transliteration system follows the steps which are represented in Fig. 1.

TABLE I.
Possible Syllable Combinations

Syllable Structure	*Example*	*Syllabified form (English)*	*Syllabified form (Punjabi)*
V	Iri	[i] [ri]	[ਇ] [ਰੀ]
CV	Maneesh	[ma] [neesh]	[ਮ] [ਨੀਸ਼]
VC	Azmat	[az] [mat]	[ਅਜ਼] [ਮਤ]
CVC	Karnal	[kar] [nal]	[ਕਰ] [ਨਾਲ]
CCVC	Shanti	[shan] [ti]	[ਸ਼ਾਨ] [ਤੀ]
CVCC	fauj dar	[fauj] [dar]	[ਫੌਜ] [ਦਾਰ]
VCC	Ani	[a] [ni]	[ਅ] [ਨੀ]

V = Vowel; C = Consonant

Algorithm for syllable extraction

1. Input sentence is taken in source language i.e. English.

2. {a,e,i,o,u} is defined as vowel set and rest other are defined as consonants. This defines the basic vowel set.
3. A vowel and its following consonants are treated as a separate syllable.
4. If Consonants are followed by a single vowel then they form an individual syllable.
5. If Consonants are followed by more than one vowel then consider both vowels as separate syllables and consonant as separate syllable
6. Other remaining characters sets are regarded as separate syllables.

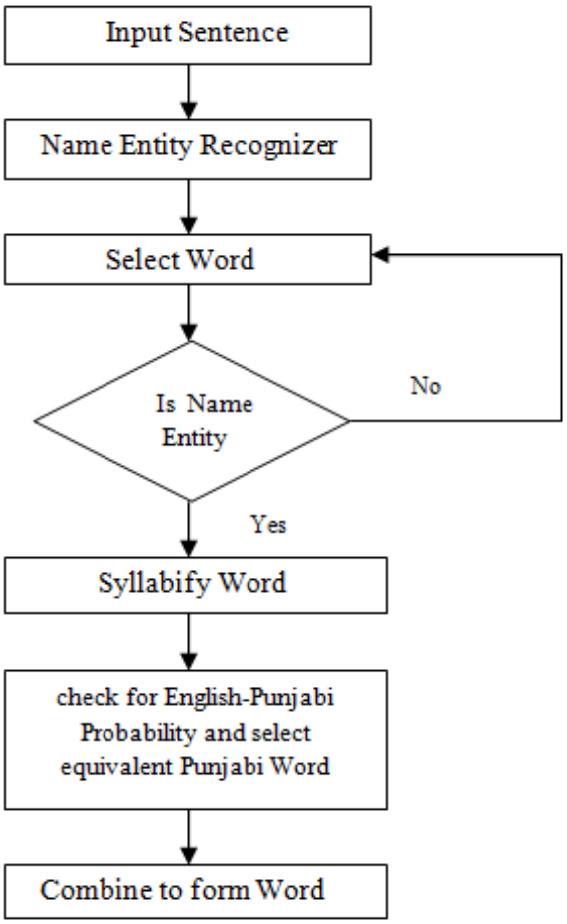

Figure 1. System architecture

For translating our English input text to equivalent Punjabi output we first recognize the name entities from our input sentence. For this we have used Stanford NER tool [15] which gives us recognized name entities in English text. The text entered by the user is first analyzed and then preprocessed. After the extraction of name entity, a word will be selected from the recognized name entities. If the selected input is some name entity, it is passed to the syllabification module through which the syllables are extracted otherwise a new word will be selected. When a name entity is recognized and is syllabified then each equivalent English syllables is searched in the syllabified database. When an English syllable has two corresponding Punjab syllables, the one with highest probability is selected. Table II shows the snapshot of this database. Table III shows the statistics of English-Punjabi parallel corpus that we have used for performing our experiment.

TABLE II.
SNAPSHOT OF SYLLABLE DATABASE

English Syllable	*Punjabi Syllable*	*Probability*
ra	ਰਾ	0.8764
m	ਮ	0.6543
ba	ਬ	0.5763
za	ਜ਼ਾ	0.6758
r	ਰ	0.8753

TABLE III.
TEST CORPUS

	English Words	*Punjabi Words*	*English Sentences*
Training	22,000	22,000	-
Testing	13023	-	2000

Next testing is performed to test our system. Name entities have been extracted and transliterated using the parallel corpus of syllabified name entities. We have calculated N-gram probabilities based on relative frequency for all the extracted syllables. Fig. 2 shows the testing phase.

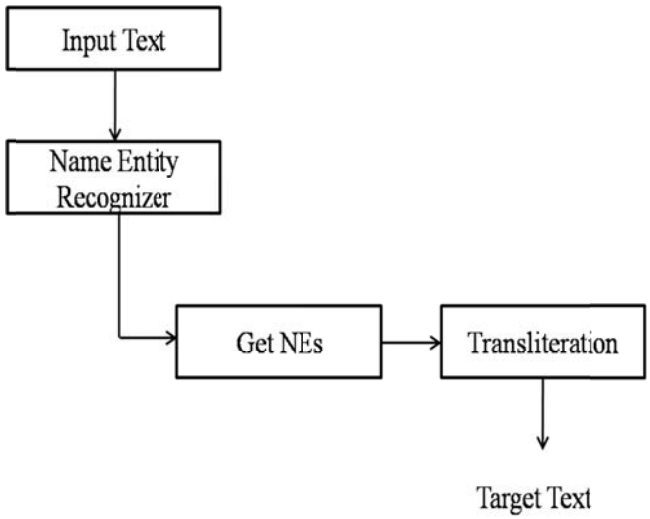

Figure 2. Testing architecture

For calculating the N-gram probability we have used the following formula [16].

$$P(S_n|S_{n-1}) = \frac{C(S_{n-1}|S_n)}{C(S_{n-1})} \quad (1)$$

Where S_n is the current syllable, S_{n-1} is the previous syllable. Conditional Probability of current syllable (S_n) is calculated by counting the no. of times current syllable (S_n) and the previous syllable (S_{n-1}) is seen together in the corpus upon the no. of times previous syllable (S_{n-1}) occurred independently, i.e. with or without current syllable.

For example: For a name Dileep probabilities for every syllable can be as follows:

$Prob(\text{ਦਿ}|di) = \frac{C(di,\text{ਦਿ})}{C(di)} = 0.9519231$

$Prob(\text{ਲੀਪ}|leep) = \frac{C(leep,\text{ਲੀਪ})}{C(leep)} = 0.5789474$

Final Score= Prob(ਦਿ|di)×Prob(ਲੀਪ|leep) = 0.551113404

After selecting the equivalent Punjabi syllable on the basis of probability calculation these syllables are joined together to form the corresponding word and final output is obtained.

V. Evaluation

We have used the corpus of 2000 sentences which consists of 23678 words. Among them 55% were name entities i.e. 13023 were proper names, locations, organization, percent, money, date & time. Among these 92% were name entities of type person & location and organization i.e. 11981 were person names, location names or organization names and rest were percent, money, date and time From these our system correctly transliterated 11689 name entities. This provided us with an accuracy of 89.76%. Fig. 3 shows the Evaluation Results. Some of the names which are correctly transliterated by our system are shown in Table IV and some which are incorrect are shown in Table V.

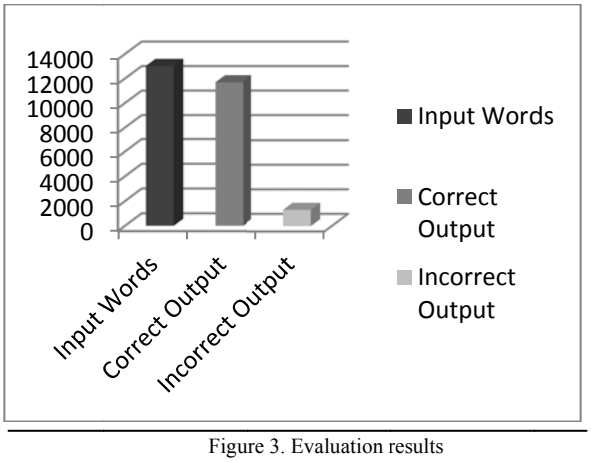

Figure 3. Evaluation results

TABLE IV
CORRECT OUTPUT

Input Word	*Output Word*
Khimram	ਖਿਮਰਾਮ
Gurpreet	ਗੁਰਪ੍ਰੀਤ
Dilsukh	ਦਿਲਸੁਖ
Haryana	ਹਰਿਆਣਾ
Mathurawale	ਮਥੁਰਾਵਾਲੇ

TABLE VI.
INCORRECT OUTPUT

Input Word	*Correct Output*	*Our System*
Arora	ਅਰੋੜਾ	ਅਰੇਰਾ
Kaushal	ਕੌਸ਼ਲ	ਕੋਊਸ਼ਲ
chintalpelly	ਚਿਨਤਾਲਪੱਲੀ	ਚੀਨਤਾਲਪਲਲੀ
Poddar	ਪੋਦਦਾਰ	ਪੋੜਦਾਰ
Vikranth	ਵਿਕ੍ਰਾਂਥ	ਵਿਕਰਾਂਤ

VI. Conclusion

In this paper we have emphasized on improving the quality of machine translation through appropriate transliteration of name entities. This paper gives a brief introduction of syllabification i.e. syllable extraction. We have shown the recognition of name entities from the input text, extracted syllables from our input text and performed transliteration of name entities in source language to target language through probability calculation. This transliteration of name entities will prove to be very beneficial in the field of machine translation. Through our experiment we attain an accuracy of 89.76%.

References

[1] P. Singh, G. S. Lehal, "Corpus Based Statistical Analysis of Punjabi Syllables for Preparation of Punjabi Speech Database", *International Journal of Intelligent Computing Research,* Volume 1(3), 2010.

[2] V. Gupta, G. S. Lehal, "Named Entity Recognition for Punjabi Language Text Summarization", *International Journal of Computer Applications,* Vol33(3), 2011.

[3] K. Deep, V. Goyal, "Development of a Punjabi to English transliteration system", *International Journal of Computer Science and Communication Vol. 2(2), 2011*, pp. 521-526.

[4] B. Babych, A. Hartley," Improving Machine Translation Quality with Automatic Named Entity Recognition", *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools,* 2003, pp 1-8.

[5] Y. AL-Onaizan, K. Knight, "Translating Name Entities Using Monolingual and Bilingual Resources", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 2002, pp. 400-408.

[6] A. Hassan, H. Fahmy and H. Hassan, "Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora", *AMML07, 2007*.

[7] H. Ondrej, H alek, R. Rudolf, A. Tamchyna, and O. Bojar, "Named Entities from Wikipedia for Machine Translation", *Proceedings of the Conference on Theory and Practice of Information Technologies,* 2011.

[8] K. Deep, V. Goyal, "Hybrid Approach for Punjabi to English Transliteration System", *International Journal of Computer Applications,*Vol28(1), 2011.

[9] D. Kaur, V. Gupta, "A survey of Named Entity Recognition in English and other Indian Languages"

[10] International Journal of Computer Science Issues, Vol. 7(6), 2010.
[11] N. Joshi., I. Mathur, and S. Mathur, "Frequency based predictive input system for Hindi", *In Proceedings of the International Conference and Workshop on Emerging Trends in Technology*, pp. 690-693. ACM, 2010.
[12] N. Joshi, I. Mathur," Input Scheme for Hindi Using Phonetic Mapping, Proceedings of National Conference on ICT: Theory, Practice and Applications", 2010.
[13] Deepti Bhalla, Nisheeth Joshi, Iti Mathur, 2013, "Rule-based Transliteration scheme for English to Punjabi", *International Journal on Natural Language Computing*, pp 67-73, Vol 2 (2).
[14] Juhi Ameta, Nisheeth Joshi, Iti Mathur, 2013, "Improving the Quality of Gujarati-Hindi Machine Translation Through Part of Speech Tagging and Stemmer Assisted Transliteration", *International Journal on Natural Language Computing*, pp 49-54, Vol 2 (3).
[15] Deepti Bhalla, Nisheeth Joshi, Iti Mathur, 2013, "Improving the quality of MT output using novel name entity translation scheme" *Proceeding of Second International Symposium on Natural Language Processing (NLP'13). Published in ICACCI 2013, MYSORE. Proceedings which will be available through* IEEE Xplore®.
[16] J. R. Finkel, T. Grenager, and C. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistic*, 2005, pp. 363-370.
[17] D. Jurafsky, J. H. Martin, Speech and Language processing An Introduction to speech Reccognition, Natural Language Processing, and Computational Linguistics, 2nd Edition, 2010, Pearson.