A HYBRID APPROACH OF ENGLISH- HINDI NAMED-ENTITY TRANSLITERATION

Vaishnavi Singhal[1]**, Neha Tyagi**[2]

[1,2]*Department of CSE &IT, Jaypee University of Information Technology, Waknaghat, H.P., (India)*

ABSTRACT

In recent years, machine transliteration has gained a center of attention for research. Both machine translation and transliteration are important for e-governance and web based online multilingual applications. As machine translation translate source language to target language which results in wrong translation for named entities. Named entities are required to be translated with preserving their phonetic properties. Thus we need named-entity transliteration is utmost required. In this paper, the main focus is on the English-to-Hindi Named-Entity transliteration and a hybrid approach is proposed. The main problem arises during the transliteration of English-Hindi is the possibility of combination of English alphabets to Hindi akshara. There is no specific rule set has been designed to convert Hindi akshara to particular English syllable yet. To solve this issue a hybrid approach has been proposed where syllabification and the uni-gram model is used. The syllabification is done based on rule-based approach. This approach is first syllabifies the English name into appropriate syllable using rule based approach which is termed as syllabification. Then syllables are matched into particular Hindi akshara on the basis of corpora that is designed on the basis of English-Hindi Name-pairs knowledgebase.

Keywords: Machine Translation, Named-Entity, Rule-Based Approach, Syllabification, Uni-Gram Model

I INTRODUCTION

Machine translation is playing an important role in research from last sixty years. But still we didn’t get any good translation which will give the desired result. One of the drawbacks of machine translation system is improper translation of named entities (NEs). Named Entities are to be translated without losing their phonetic properties. Most of the existing machine translation systems are unable to address this issue and thus provide a poor quality translation. To resolve this issue, transliterators came into existence. Transliteration is thus a conversion of text from one script to another.

1.1 Named-Entity Transliteration

Named Entity Transliteration is a process of converting an input named-entity from source language to target language. The process of translating source word to target while preserving their phonetic properties is called Transliteration. For Example: the translation of word “Honey Singh” into Hindi language will be “मधु संह whereas its transliteration will be “हनी संह”. Hence in order to convert named-entities form source language to target language, the named-entity transliteration is done.

Machine Transliteration approach is characterized into two categories [7]:

- Grapheme-based approach

In this approach, an orthographic method is follows and the source language grapheme/characters are directly mapped into target language grapheme/character.

- Phoneme-based approach

  In this approach, the phonetic process is follows and the source language phoneme is converted into target language phoneme.

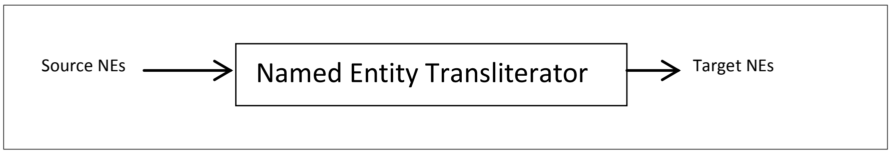

Figure 1- Named-Entity Transliterator

1.2 Named-Entities

According to [8], the entities which come under named-entities are:

- People: Individuals, fictional characters, small groups
- Organization: companies, agencies, political parties, sport teams
- Location: physical extents, mountains, lakes, seas
- Geo-Political Entities: countries, states, provinces, localities
- Facilities: bridges, buildings, airports
- Vehicles: planes, trains, automobile

1.3 English-to-Hindi Named Entity Transliteration

Transliteration is a process of converting a text string in the source writing system or orthography to another text string in the target writing system or orthography, such that the target language name is phonemically equivalent to the source name and conforms to the phonology of the target language.

There are 22 constitutionally recognized languages and 11 scripts in Indian constitution that are used in different regions spread across the country [7]. The factors which make the named-entity transliteration difficult are:

- Devanagari script is used by Hindi language which is much more difficult than Roman script used by English language.
- The characters used in English language are 26 which are much less than the aksharas used in Hindi language i.e. 52.
- We can also not recognize the named entities by the capitalized word in Hindi language because of no concept of capitalization like English or other European languages.
- Due to the Devanagari script, Hindi is highly phonetic and inflectional language than English.
- Indian place names are frequently homographic with the common words of person names, presence of exonyms and presence of endonyms.
- There can be more than one valid transliteration for a single English name into Hindi languagedue to no

  रामा

The various issues occur during English-Hindi named-entity transliteration are orthographic variations, morphological variations, lexical ambiguity, tokenization, translation divergence conflation.

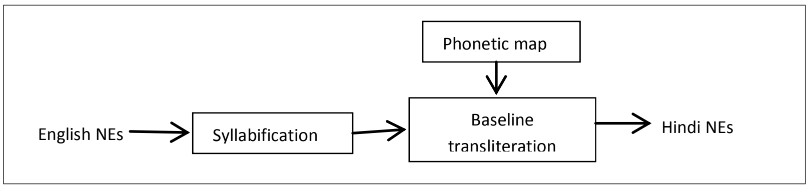

Figure 2- Architecture of Named-Entity Transliterator

The rest of the paper is structured as follows: section 2 contains the literature review required for this research. Section 3 explains the proposed methodology, its flow chart and corresponding algorithm which is explain by taking an example. The next section 4 shows the implementation result and the last section 5 concludes this research.

II RELATED STUDY

The base paper [2] which has been referred for this research used English-Hindi language pair for their experiments. First of all the rule-based approach is used in order to extract individual phonemes from the English words. Then that English phoneme is converted into corresponding Hindi phoneme using statistical approach. The accuracy gained by this approach was 83.40%. For the phonification process they used 7 different phonemes V, CV, VC, CVC, CCVC, CVCC and VCC respectively. They used probability calculation model to generate probability on English-Hindi phoneme knowledgebase for transliteration process.

Jiang et. al. [10] did translation of named-entities using transliteration with web-mining. They trained the classifier in pronunciation similarity, bilingual context and co-occurrence by using maximum-entropy based approach. A phonetic based algorithm is proposed by Joshi and Mathur [11] which created a mapping table and a set of rules for English-Hindi transliteration. Bhalla et. al. [12] translated the English-Punjabi name-entities by using Moses toolkit and clamed 88% of accuracy. .

Sharma et. al[13] trained a statistical machine translation system for successfully translating English-Hindi named entities using CRF-based approach. They showed 85.79% accuracy and showed that CRF is best suited for processing Indian languages. Similarly Manokaro et. al. [1] designed a Hindi to English transliteration of Named-entities using CRF. Ameta et. al[14] developed a transliteration system for Gujrati-Hindi language-pair.In [3] , the authors designed the translation system for English-Arabic language pairs. Whereas Wolodjaet. al. [4] builded a system which is multilingual for named entity disambiguation, translation and transliteration. In [5], [6], the authors also defined the various approaches for named entity transliteration and its improvement.

The major players in the machine transliteration of Indian Languages are C-DAC, NCST and Indictrans. C-DAC provides their technology based on ISCII in 1980 in the form of hardware based card called GIST [7]. NCST developed a phonemic code based scheme for effective processing of Indian languages in 2003 [15].

III PROPOSED METHODOLOGY

This paper focuses on the problem of English-Hindi named-entity transliteration. This is a challenging task because of the highly inflectional and phonetic characteristic of Hindi language. Named entity transliteration considers various issues like are orthographic variations, morphological variations, lexical ambiguity, tokenization, translation divergence conflation. To solve these issues we have proposed a hybrid approach for English-Hindi named-entity transliteration. For transliteration, a knowledgebase including the English-Hindi named pairs is used. The English name is firstly looked up into the knowledgebase and if it is found then the corresponding Hindi name is chosen. Otherwise the named-entity will go for Syllabification. Syllabification is a process of dividing the name into syllables i.e. C and V. Then these syllables are combined into corresponding 5 phonemes namely C, V, CV, CVC, VC. A corpora is designed for all the possible combination of English alphabets for these phonemes and their corresponding Hindi aksharas are mapped with it. From these corpora the corresponding Hindi akshara are chosen for the source name. Finally these Hindi aksharas are combined to result out the corresponding Hindi name.

3.1 System Design

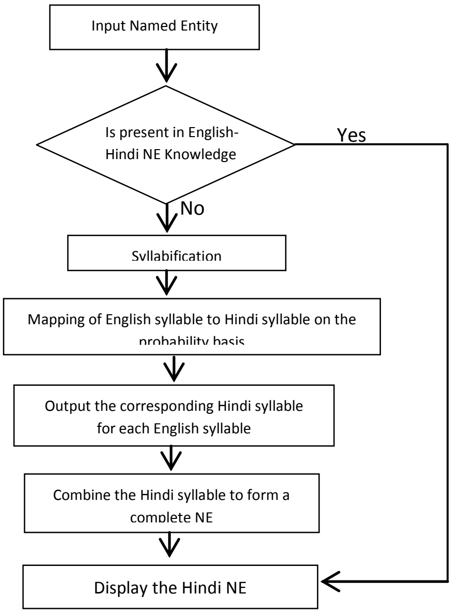
Input Named Entity
Is present in English-Hindi NE Knowledge
Yes
No
Syllabification
Mapping of English syllable to Hindi syllable on the probability basis
Output the corresponding Hindi syllable for each English syllable
Combine the Hindi syllable to form a complete NE
Display the Hindi NE

Figure 3- flow chart of proposed methodology

3.2 Proposed Algorithm

Step 1. Take the input i.e. the named entity.

Step 2. Check whether the corresponding Hindi name is present in the knowledge base.

Step 3. If yes then return back the corresponding Hindi name

Step 4. If no then Syllabification is done this will give the grouping of syllables

Step 5. Then mapping of each English syllable to Hindi syllable has been done from the English-Hindi syllable knowledge base. This will return the entire possible Hindi syllable for each English syllable.

Step 6. From the resulted Hindi syllable, the syllable which has maximum probability of matching English to Hindi syllable is chosen.

Step 7. All the resulted Hindi syllable of each English Syllable is then combined to display the final Hindi transliterate of English named entity.

3.3 Syllabification

A. Label each English character into syllable as C (consonants) or V (vowels).

B. Then grouping of the syllables into phoneme is done by defining some rules. This grouping is done in two filters.

C. For each syllable[i]

  a. If syllable[i] = V

    i. If char[i+1] = “m” or “n” then combine char[i] & char[i+1] into char[i] and label the syllable[i] as “VC”

    ii. If syllable[i+1] = “V” then combine char[i] & char[i+1] into char[i] and label the syllable[i] as “V”

    iii. Else syllable[i]= “V”

  b. Else

    i. If syllable[i+1] = “C” and syllable[i+2] = “C” then check whether this combination of char[i], char[i+1] and char[i+2] is combine to a single Hindi akshara. If yes then combine char[i], char[i+1] and char[i+2] into char[i] and label the syllable[i] as “C”. Otherwise continue

    ii. Else if syllable[i+1] = “C” then check whether this combination of char[i], char[i+1] and char[i+2] is combine to a single Hindi akshara. If yes then combine char[i], char [i+1] into char[i] and label the syllable[i] as “C”. Otherwise continue

    iii. Else continue

  c. Return syllable

For example:

Take input “Vaishnavi” convert into lowercase “vaishnavi”

Syllabification: CVVCCCVCV

Grouping of Syllabification: CVVCCCVCV

C V C VCV→CV CV CV

vaishnavi→ v ai shn a v i→ vai shna vi

Hindi mapping of each syllable from the knowledgebase: v (व) ai(ै, ाइ)

shn(ष्ण) a(◌, ा)

v(व) i(ि, ी)

Now combination is:

vai: वै, वाइ

shn: ष्ण, ष्णा

vi: वी, व

Now the corresponding Hindi syllable will be chosen which will have maximum probability in the English-Hindi pair syllable created by checking the pairs which named entity generally used from the corpora.

Suppose the probability of vai is maximum for “वै” similarly for shn is “ष्ण” but there is no data found for the combination “vi” then randomly any combination will be considered.

Finally combine all Hindi syllables: वैष्णवी will be the output from this algorithm.

IV IMPLEMENTATION AND RESULT

The proposed algorithm is implemented on the JAVA platform. Here the knowledgebase is used to calculate the probabilities of English-Hindi pairs. This knowledgebase is designed by taking the English-to-Hindi named entities from various sources like newspaper, online repositories and magazines.Some examples of English-to-Hindi name-pairs are tested on the simulation environment.

The evaluation result has been shown below.

Table 1- Evaluation Table of Proposed System

English Named Entity	Transliterated Hindi Named Entity	Process	Correct/wrong
Vaishnavi	वैष्णवी	knowledgebase	Correct
Mohak	मोह्क	Syllabification	Correct
Robin	रोबिन	Syllabification	Correct
Sanjay	सन्जय	Syllabification	Wrong
Archana	अर्चाना	Syllabification	Wrong
Jhalak	झलक	Syllabification	Correct
Dev	डव	Syllabification	Wrong
Neha	नेहा	Syllabification with unigram	Correct
Rahul	राहुल	Syllabification	Correct
Rakesh	रा कश	Syllabification	Wrong
Shanti	शान्ति	Syllabification	Correct
Khushi	खु श	Syllabification	Wrong
Nikita	नि कता	Syllabification	Correct
Aradhya	आराध्या	Knowledgebase	Correct
Lokendra	लो कन्द्रा	Syllabification	Wrong
Ram	राम	Syllabification	Correct
Ankur	अंकुर	Syllabification	Correct
Amitabh	अ मताभ	Syllabification	Correct
Naksh	नाक्ष	Syllabification	Wrong
Rashi	रा श	Syllabification	Correct
Anshul	अंशुल	Syllabification	Correct
Aashish	आ शश	Syllabification	Wrong

The accuracy is calculated with the help of precision and recall. It is observed that this system is 84.23% accurate for transliterating English-Hindi named-entities.

V CONCLUSION

In this project we've developed a system for the Hindi-English named-entity transliteration. First of all syllabification is done. Here we've used only 5 phonemes for the syllabification process i.e. C, V, CV, VC, and CVC as compared to 7 phonemes used in [2]. After that mapping of each English syllable with the all possible Hindi syllable is done by looking to the English-Hindi pair knowledgebase. From all the combinations of Hindi syllable, one syllable is chosen by using unigram statistical model. After that the combination of the all the syllable is done so that the corresponding Hindi transliteration of English named-entity will be resulted. This approach gives accuracy of 84.23% which is improvement over the approach given in [2].

As it is mentioned earlier that named-entity transliteration plays an important role in e-governance and web-based online multilingual applications. Thus we can use this approach to various applications where Hindi-to-English named entity transliteration is required.

This system is only designed and tested for the person name-entity only. This can be extended for the other name-entities too.Sharma et.al. in [13] mentioned that CRF-based approaches are better for the Indian languages. Thus, we will try to combine our approach with CRF-based approach in future and try to improve the accuracy of the system.

REFERENCES

[1] Manokrao L. Dhore, Shantanu K Dixit, Tushar D Sonwalkar, Hindi to English Machine transliteration of Named Entities using Conditional Random Fields, International Journal of Computer Applications, 48(23), June 2012

[2] Shruti Mathur, Varun Prakash Saxena, Improving the Quality of English-Hindi Name Entity Translation,International Journal of Computer Applications, 96(25), June 2014

[3] Nasreen Abdul Jaleel and Leah S. Larkey, Statistical Transliteration for English-Arabic Cross Language Information Retrieval", Proc. Of CIKM'03, Nov. 3-8, 2003

[4] Wolodja Wentland, Johannes Knopp, Carina Silberer, Matthias Hartung,Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration

[5] Sriniasan C. Janarthanam, Sethuramalingam S, Udhyakumar Nallasamy,Named Entity Transliteration for Cross-Language Information Retrieval using Compressed Word Format Mapping algorithm, Proc. Of 2nd International ACM Workshop on Improving Non-English Web Searching (iNEWS08), CIKM-2008

[6] Dan Goldwasser, Dan Roth,Active Sample Selection for Named Entity Transliteration, Proc. Of ACL-08: HLT, Short Papers (Companion Volume), pages 53-56, June 2008

[7] M L Dhore, R M Dhore, P H Rathod, Transliteration by Orthography or Phonology for Hindi and Marathi to English: Case Study, International Journal on Natural Language Computing (IJNLC) 2(5), October 2013.

[8] Daniel Jurafsky and James H. Martin,Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition (Pearson publication)

[9] Darvinder kaur, Vishal Gupta, A survey of named entity recognition in English and other Indian languages, IJCSI International Journal of Computer Science Issues, 7(6), 2010, 239-245.

[10] L. Jiang, M. Zhou, L. Chein, and C. Niu, Name Entity Translation with Web Mining and Transliteration, Proc. of 20th International Joint Conference on Artificial Intelligence, 2007, 1629-1634.

[11] N. Joshi and I. Mathur,Input Scheme for Hindi Using Phonetic Mapping, Proc. of the National Conference on ICT: Theory, Practice and Applications, 2010.

[12] D. Bhalla, N. Joshi and I. Mathur,Rule Based Transliteration Scheme for English to Punjabi, International Journal of Natural Language Computing,2(2), 2013, 67-73.

[13] S. Sharma, N. Bora and M. Halder,English-Hindi Transliteration Using StatisticalMachine Translation in Different Notation, 2012.

[14] J. Ameta, N. Joshi and I. Mathur, Improving the Quality of Gujarati-Hindi MachineTranslation through Part-of-Speech Tagging and Stemmer Assisted Transliteration, International Journal of Natural Language Computing, 2(3), 2013,49-54.

[15] Joshi R K, Shroff Keyur and Mudur S P, A Phonemic code based scheme for Effective processing of Indian languages, National Centre for Software Technology, Mumbai, 23rd Internationalization and Unicode Conference, Prague, Czech Republic, 2003, 1-17.