Improving the Quality of English-Hindi Name Entity Translation

Shruti Mathur
Department of Computer Engineering
Government Women Engineering College
Ajmer, India

Varun Prakash Saxena
Department of Computer Engineering
Government Women Engineering College
Ajmer, India

ABSTRACT

Machine Translation is a novel step in overcoming the language barrier. The results of the research in this area has started to show its results with some good machine translators being available. In the context of English-Hindi language pair, due to bad name entity translations, the quality of translation deteriorates. In this paper we have addressed this issue where we show that only a single approach is not effective in dealing with this issue. We need to devise mechanism and address this problem in a multi pronged approach.

General Terms

Natural Language Processing, Machine Translation

Keywords

Name Entity, Machine Transliteration, Entity Translation, Phoneme Identification

1. INTRODUCTION

In the past sixty years, there has been a good progress in the area of machine translation; though getting a good translation is not possible. One of the many reasons is poor translation or mis-handling of name entities. Most of the MT engines are not able to properly translate name entities. They either try to leave it as it is or try to transliterate the text. The first case is not acceptable as the user who wishes to see the translation in Hindi would not appreciate the output, either in part or in full, coming in English. In the second case, the output would not always be correct, because some of the same entities cannot be transliterated, they are required to be translated. For example, let us consider the sentences in table 1.

In sentence 1, Ram and Malaviya National Institute of Technology are two name entities. The first one is the person name while the second one is a collected of five words and is an organization name. Person name needs to be transliterated and in the translation it was rightly produced, but the second name entity, organization name would provide good results if it is transliterated, thus it is transliterated and the output produced is even messier. Out of the five words, none were translated properly. The proper translation of this name entity would have been “मालवीय राष्ट्रीय प्रौद्योगिकी संस्थान”. The case is same with sentence 2, where “Indian Institute of Technology” was not properly translated. The third sentence produced the correct translation/transliteration for both the name entities. In Sentence 4, again “Jawaharlal Nehru Marg” was not properly transliterated. In sentence 5, “Indian Institute of Information Technology” was not properly translated. Though some of the words in the string were, but their sequence got jumbled up. Name entities of Sentence 6 were properly translated/transliterated.

Table 1: Sample English Sentences and Their Translations

S.No.	English Sentence	Hindi Translation
1.	Ram is studying in Malviya National Institute of Technology.	राम तकनीक के मल्विय नेशनल इंस्टिट्यूट में अध्ययन करते है ।
2.	Ram is studying in Indian Institute of Technology.	राम आय आय टी में अध्ययन करते है ।
3.	Ram is studying in Jawaharlal Nehru Medical College.	राम जवाहरलाल नेहरू मेडिकल कॉलेज में अध्ययन करते है ।
4.	Ram is going to Jawaharlal Nehru Marg.	राम जवाहरलाल नेहरू छोटी करने के लिए जा रहा है।
5.	Ram is studying in Indian Institute of Information Technology.	राम की सूचना तकनीकी इंडियन इंस्टिट्युट में अध्ययन करते है ।
6.	Ram is studying in Jawaharlal Nehru University.	राम जवाहरलाल नेहरू विश्वविद्यालय में पढ़ने है ।

From this illustration, we can clearly see that at times name entities which should be identified in a group and should be translated are not handled properly. At times some name entities are not properly transliterated as well.

The rest of the paper is organized as: Section 2 describes literature survey. Section 3 describes our proposed approach, it shows our experimental setup and the methodology adopted. Section 4 shows the evaluation of our model and section 6 concludes the our work.

2. LITERATURE SURVEY

Babych and Hartley [1] implemented an automatic name entity recognition system which was implemented on the outputs generated by five different machine translation systems. They incorporated GATE’s information extraction module in their systems concluded that combining IE technology with machine translation has a great potential for improving the overall output quality. Al-Onaizan and Knight [2] developed an algorithm for translation Arabic-English name entity phrases. They used both monolingual and bilingual resources and compared their results with the results produced by human translators and some commercial MT systems. They showed that their system had better correlation with human translators than any other system. They achieved an accuracy of 84%. Hassan et al. [3] performed a similar study which was done on extracted translation pairs. They showed that by using their approach the performance of a named entity translation system improves. Jiang et al. [4] used transliteration with web mining in translation of name entities. They used a maximum entropy based approach to train a classifier on pronunciation similarity, bilingual context and co-occurrence. This classifier was used to rank the candidate translations produced. Yeh et al. [5] proposed a pattern matching method for finding name entity's translation online. They developed an algorithm which automatically generated and weighted pattern which were used to search for name entities from bilingual corpus.

In an Indian context, Joshi and Mathur [6] proposed a phonetic mapping based algorithm for English-Hindi transliteration system which created a mapping table and a set of rules for transliteration of text. Joshi et al. [7] also proposed a predictive approach of for English-Hindi transliteration. Here instead of generated a single output they provided a list of possible text that can be selected by the user for correct transliteration. They looked at the partial text and tried to provide possible complete list as the suggestive list. Bhalla et al. [8] who used these two approaches for transliterating person and location name entities. Sharma et al. [9] trained a statistical machine translation system which could successfully translate English-Hindi name entities. They used Moses and Phrasal for this purpose. Moore [10] trained a classifier for English-Hindi transliteration using CRF based approach. They showed that using this approach we can successfully translate name entities with 85.79% accuracy and concluded that CRFs are best suited for processing Indian languages. Kharpa et al. [11] proposed a compositional machine transliteration where several transliteration approaches were combined to improve the accuracy. Their experiments showed the benefits of compositional methodology using some state of the art machine transliteration approaches. Agrawal and Singla [12] used three pronged approach in translating name entities. They used an aligner which generated English equivalents for Chinese name entities, a language model which improved the readability and a ranker which selected the best weighted translations. Ameta et al. [13] developed a transliteration system for Guajarati-Hindi language pair and used it in their Gujarati-Hindi translation engine which could effectively translate Gujarati name entities into Hindi. Bhalla et al. [14] used the Moses toolkit for generating translations for English-Punjabi name entities and claimed an accuracy of 88%.

3. PROPOSED SYSTEM

3.1 Experimental Setup

In order to implement a name entity translation system, we first collected 12,000 sentences from various English news sites. This was our training corpus onto which we have build our model. Next we used Stanford's NER [15] tool to extract name entities from these sentences. In all 51,583 name entities were extracted. Table 2 shows the statistics of this extraction.

Table 2: Statistics of Name Entities Extracted

S.No.	Name Entity	Count
1.	Person	13,482
2.	Location	11,926
3.	Organization	14,000
4.	Date	4,926
5.	Time	3,889
6.	Misc.	3,630
Total		**51,583**

Once the extraction of name entities was complete, we then extracted the phonemes from them. For this we developed an algorithm. We identified that all the words in English can be captured using seven different combinations of vowels (V) and consonants (C). These were V, CV, VC, CVC, CCVC, CVCC, VCC. After generating all the phonemes of the English name entities, we employed a human annotator to transcribe their Hindi equivalents. Thus, we have a list of English phonemes and their Hindi equivalent phonemes. Next we generated the three frequencies which are as follows:

1. English Phonemes
2. Hindi Phonemes
3. Combination of English-Hindi Phonemes

After generation of frequencies of these, we generated the probabilities of these three variants. For generating probabilities of English and Hindi phonemes separately, we used equation 1 and for generating probabilities of combination of English-Hindi phonemes, we used equation 2.

$$Prob(Phoneme) = \frac{Freq(Phoneme)}{|Vocab|} \quad (1)$$

$$Prob(English, Hindi) = \frac{Freq(English,Hindi)}{Freq(English)} \quad (2)$$

Here, Prob(Phoneme) is the probability of a particular language of phonemes which is calculated by the frequency of that particular phoneme divided by the total number of unique phonemes available in the frequency table. Prob(English,Hindi) was the probability calculation of the combination of English and Hindi phonemes. This was calculated using the combined frequency count of English and Hindi phonemes which occur together divided by the frequency count of English phonemes. Table 3,4 and 5 shows the snapshot of probability table of English, probability table of Hindi and probability table of the combination of English and Hindi.

Table 3: Probability of English Phonemes

English Phoneme	Probability
a	0.76214
bh	0.54325
i	0.08765
ra	0.09321
ro	0.05432
shi	0.02312

Table 4: Probability of Hindi Phonemes

English Phoneme	Probability
अ	0.43421
भ	0.14211
ई	0.09321
रा	0.07453
रो	0.04324
शि	0.02123

Table 5: Probability of English-Hindi Phonemes

English Phoneme	Hindi Phoneme	Probability
A	अ	0.4532
Bh	भ	0.3218
I	ई	0.4312
Ra	रा	0.6532
Ro	रो	0.4533
Shi	शि	0.6788

Table 5: Snapshot of Translation Knowledgebase

English	Hindi
Indian Institute of Technology	भारतीय प्रोद्योगिकी संस्थान
World Health Organization	विश्व स्वास्थ संगठन
United Nations	संयुक्त राष्ट्र
Gulf Countries	खाड़ी राष्ट्र
National Institute of Technology	राष्ट्रीय प्रोद्योगिकी संस्थान

Besides this, we also used a knowledge base to store the names of the organizations in India and some of the popular organizations of the world. This was done because at times we need to translate the some name entities. Most of the time these name entities are organization names. So we created this knowledge base. Table 6 shows the snapshot of this knowledgebase.

3.2 Methodology

Since our main goal was to improve the quality of machine translation by properly translating name entities, we applied both translation and transliteration on the extracted name entities. We first extracted the name entities from the input string and checked if the name entity was organization or not. If it was organization, then we checked it in a knowledge base that we have created and translated it, as it is. This was the same approach as that applied in example based machine translation. If the name entity was not an organization name or if the particular name was not available with us then sent it to the phonification algorithm which extracted the phonemes from the text. For each name entity which was to be translated we recursively transliterated the English phonemes to their individual Hindi phonemes. When there were two Hindi phonemes for a English phoneme, then the priority was given to the one which had the highest probability. If there was a case when no corresponding Hindi phoneme was found, then the English phoneme was left as it is. This complete process is shown in the following algorithm.

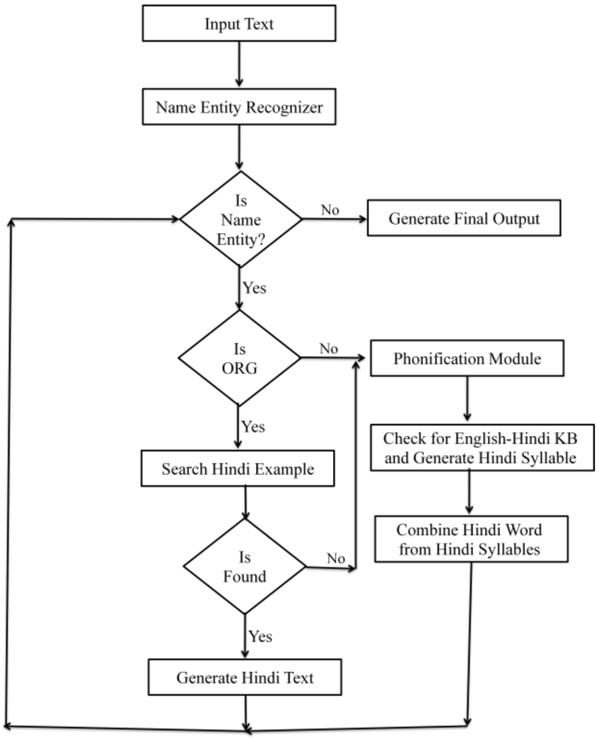

Figure 1: Complete Name Entity Translation System

Input: English Phoneme List
Output: Hindi Word
Conversion Algorithm

```
1.  Input the English Phoneme List as phoneme.
2.  Read English-Hindi Probability KB as KB
3.  phlen = phoneme.length
4.  count = 1
5.  repeat steps 5 to 8 till count <= phlen
6.  generate list of English-Hindi phonemes for
    phoneme[count] with their respectively probability
7.  hinpho[count] = max_prob(phoneme[count])
8.  count += 1
9.  combine hinpho to form Hindi word as hword
10. return hword
```

The entire working of the system is shown in figure 1, which shows how a input string is checked and if a name entity is found, then it is checked for an organization name. If it is not found so then the system generated the statistical transliteration of the text. Otherwise it generates an example based translation of the text.

4. EVALUATION

To evaluate the system we again collected 1000 sentences from the English news sites and extracted the name entities from them. This test corpus was separate from the training corpus. This corpus had 9234 name entities. Table 6 shows the statistics of this data.

Table 6: Statistics of the training corpus

S.No.	Name Entity	Count
1.	Person	5,263
2.	Location	2,770
3.	Organization	1,108
4.	Date	13
5.	Time	27
6.	Misc	53
Total		**9,234**

This test corpus was given as an input to our system which generated the results. WE also asked a human annotator to provide us the results of the name entities extracted. We calculated the quality of the translation using standard precision, recall and f-measure scores. The computation of these is shown in equation 3,4 and 5 respectively.

$$Precision\ (P) = \frac{correct}{System\ Output} \qquad (3)$$

$$Recall\ (R) = \frac{correct}{Reference\ Output} \qquad (4)$$

$$F - Measure = \frac{2 \times P \times R}{P+R} \qquad (5)$$

Here, the system generated outputs of name entities which were matched with human annotators outputs were deemed as correct. Thus precision was calculated by dividing correct matching divided by all the name entities that the system was able to generate. Recall was calculated by dividing correct matching divided by all the human annotators output and f-measure was the combined computation of the two.

Thus, out of the 9,234 name entities, our system was able to provide the output for 9,230 name entities out of which 8,483 were correct. Thus the system attained the accuracy of 91.89%. Table 7 depicts the summary of this evaluation and figure 2 the summary of this data.

Table 7: Summary of Evaluation

Total Name Entities	9,234
System Generated Name Entities	9,230
Human Generated Name Entities	9,234
Correct Name Entities	8,483
Precision	0.9187
Recall	0.9191
F-Measure	0.9189

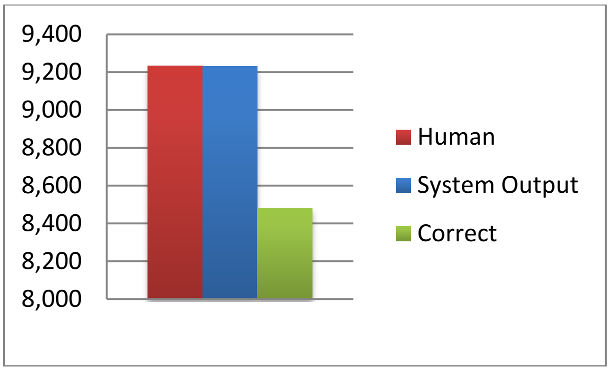

Figure 2: Overall Score

Table 8 shows the entity-wise result of the evaluation and is summarized in figure 3. We got this low score as we were not able to translate person, location and organization name entities.

Table 9: Entity Wise Analysis of Evaluation

S.No.	Name Entity	Count	System Output	Correct
1.	Person	5,263	5,263	4,893
2.	Location	2,770	2,770	2,603
3.	Organization	1,108	1,107	897
4.	Date	13	13	13
5.	Time	27	27	27
6.	Misc	53	50	50
Total			9,234	9,230

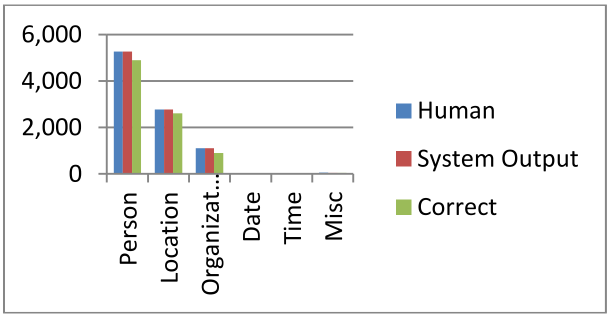

Figure 3: Entity-wise Results

5. CONCLUSION

In this paper we have shown the implementation of a Name Entity translation system for English-Hindi language pair. The system was implemented with a partial rule based approach where we generated rules for phonifications and created a knowledge base for organization names and with partial statistical approach where we generated the probabilities fo phonemes of English and Hindi. The system did fairly well with all name entities. Our system gave an accuracy of 91.89%. Thus our immediate efforts would be improve the accuracy of the system.

6. REFERENCES

[1] 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, 1-8, 2003/4/13.

[2] 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 (ACL), Philadelphia, pp. 400-408, 2002.

[3] A. Hassan, H. Fahmy and H. Hassan, “Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora” Preoceedings of AMML07, 2007.

[4] L. Jiang, M. Zhou, L. Chein, and C. Niu, “Name Entity Translation with Web Mining and Transliteration”, Proceedings of 20th International Joint Conference on Artificail Intelligence, pp. 1629-1634, Morgan Kaufmann, 2007.

[5] A. Yeh, A. Morgan, M. Colosimo, and L. Hirchman, “BioCreAtIvE Task 1A: gene mention finding evaluation” BMC Bioinformatics, 2005.

[6] N. Joshi and I. Mathur, “Input Scheme for Hindi Using Phonetic Mapping.” In Proceedings of the National Conference on ICT: Theory, Practice and Applications, 2010.

[7] 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, ACM, pp 690-693, 2010.

[8] D. Bhalla, N. Joshi and I. Mathur, “Rule Based Transliteration Scheme for English to Punjabi.” International Journal of Natural Language Computing, Vol 2, No. 2, pp 67-73, 2013.

[9] S. Sharma, N. Bora and M. Halder, “English-Hindi Transliteration Using Statistical Machine Translation in Different Notation.” 2012.

[10] R. Moore, “Learning Translations of Name Entity Phrases from Parallel Corpus.” Proceedings of EACL, 2003.

[11] M Khapra, P Bhattacharya, A Kumaran, “Compositional Machine Transliteration.” ACM Transactions of Asian Language Information Processing, Vol 9, 2010

[12] N. Agrawal and A. Singla, “Using named entity recognition to improve machine translation.” Technical report, Standford University, Natural Language Processing. 2012.

[13] J. Ameta, N. Joshi and I. Mathur, “Improving the Quality of Gujarati-Hindi Machine Translation Through Part-of-Speech Tagging and Stemmer Assisted Transliteration”, International Journal of Natural Language Computing, Vol. 2, No. 3, pp. 49-54, 2013.

[14] D. Bhalla, N. Joshi and I. Mathur, “Imporving the Quality of MT Output Using Novel Name Entity Translation Scheme.” Proceedings of 2013 International Conference on Advances in Computing, Communications and Informatics, 2013.

[15] J.R. Finkel, T. Grenger 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 Linguistics (ACL 2005), pp. 363-370, 2005.