ISSN (e): 2250 – 3005 || Volume, 09 || Issue, 6 || June– 2019 ||
International Journal of Computational Engineering Research (IJCER)

Performance Evaluation of Rule Based Token Mapping Translation System for Indian Sibling Languages Hindi-Gujarati

Dr. Kalyani Patel[1], Dr. Jyoti Pareek[2]
[1]*Assistant Professor,*
M.Sc (CA & IT),
K.S. School of Business Management, Gujarat University, Ahmedabad 380009,Gujarat ,India.
[2] *Professor,*
Department of Computer Science, Gujarat University, Ahmedabad 380009, Gujarat , India.

ABSTRACT*: The paper presents evaluation of rule based token mapping translation system for Indian Sibling languages Hindi-Gujarati. The system was evaluated on the test bed obtained from FIRE 2010, literature on Gandhiji and ELRA-W0037. For establishing relevance of the model 'into-Gujarati' BLEU, PER and METEOR score have been calculated.*
General Terms: *Machine Translation, Evaluation*
Additional Keywords and Phrases*: Sibling language, Gujarati, Hindi, GH-MAP*

Date of Submission: 11-06-2019 Date of acceptance:28-06-2019

I. INTRODUCTION

Machine translation means automatic translation of text by computer from one natural language into another natural language. The GH-MAP [Patel K. and Pareek J, 2013] is a rule based token mapping system for translation between sibling language pair Gujarati and Hindi. As Gujarati and Hindi are structurally similar languages, GH-MAP system generates target language sentence retaining the flavor of the source language. It should be noted that translation is not done in the sense of linguistics; instead word-by-word translation has been performed. Implementation of the GH-MAP system can be considered a success only if the quality of the translation produced by the system is of acceptable. The paper presents evaluation of the GH-MAP system using Hindi sentences extracted from FIRE 2010[1], literature on Gandhiji [Gandhiji, 1999; 2007] and ELRA-W0037[2]. The quality of the system has been measured using prototype developed by us for automatic calculation of evaluation metrics such as Position-independent word Error Rate (PER) [Tillmann et al., 1997], BiLingual Evaluation Understudy (BLEU) [Papineni et al, 2000] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [Banerjee and Lavie, 2005].

II. MACHINE TRANSLATION EVALUATION TECHNIQUES

Evaluation is needed to identify limitations, errors and deficiencies, which may be corrected or may be improved. In such situation human evaluation is the best option but it is impractical and costly. Evaluating a machine translation system using automatic metrics is much faster, easier and cheaper compared to human evaluations. The intuition behind metrics is that machine translation would be considered good if it resembles closely to human translation of the same sentence [Papineni et al. 2002].For our experiments, we have selected the most widespread automatic evaluation metrics based on:

- **Levenshtein-Based Measures:**

Position-independent word Error Rate (PER) [Tillmann et al.,1997] computes the Levenshtein distance without taking the word order into account.

- **N-Gram-Based Measures:**

Bilingual Evaluation Understudy (BLEU) [Papineni et al, 2000] score, measures translation quality based on precision; it compares n-gram matches between candidate translation and a reference translation.

- **The Importance of Recall:**

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [Banerjee and Lavie, 2005] put more weight on recall than on precision in the harmonic mean to measure translation quality.

[1] Forum for Information Retrieval Evaluation (FIRE,2010)
[2] ELRA : Evaluation and Language Resources Distribution Agency, France, The EMILLE/CIIL corpus from www.elda.org

III. GH-MAP SYSTEM

An effective system called the GH-MAP [Patel K. and Pareek J, 2013] has been designed for rule based token mapping for the pair of Gujarati and Hindi language. The system has been designed considering fascinating factor of closely related languages. The GH-MAP performs effective word-for-word translation with minimum efforts and resources. As a result, all the efforts which have been consumed into machine translation applications developed for Hindi to other Indian language / English language and vice-versa till now, can be used with minimum efforts along with the limited memory for Gujarati language. It is a very time consuming to develop machine translation between Gujarati and Telugu or Gujarati and English, but through our model we can translate Gujarati to Hindi with minimum efforts. Then considering Hindi as Interlingua we can translate Hindi to English or to Telugu with existing machine translation system.

3.1 Translation Mechanism

The GH-MAP system takes a sentence in source language and search for phrases in the source sentence. If the source language sentence contains phrase then the target language phrase replaces the source language phrase. The sentence is then divided into tokens. Translate token using empirical rules which depend on grammatical properties of predecessor tokens. If the source token does not get translated using empirical rule then the tokens are translated by using token mapping engine.

Token Mapping Engine performs word-for-word translation. The effective translation has been achieved by simple and computationally inexpensive methods such as

- Substring substitution [Mark´o et al., 2005].

Hindi and Gujarati languages have many words with small difference in substring. To take the advantage of this particular feature of language pair, we have developed a method for substring substitution. By substring substitution, a new word of the target language can be derived from the word of source language.

- Generating and mapping inflected noun/verb/adjective forms

Though Hindi and Gujarati are inflectionally rich languages, most of the inflected form of words can be reduced to a common stem by the operation of suffix removal. To translate compound words and inflected words, we have developed a method for stem-suffix substitution and mapping. By stem-suffix substitution and mapping, a new inflected word of the target language can be derived from the inflected word of the source language.

- Dictionary word mapping.

Hindi and Gujarati languages have some domain specific words (which have different typologies), phrases (which include compound words and idioms), invariable (uninflected) karaks, pronouns, adjectives and adverbs. To translate such words, we have created bilingual dictionary.

- Empirical Rules

For the language dependent grammatical properties, (such as differences in gender / number / person properties,) many-to-many or one-to-many token mapping is required in certain cases of post position markers, pronouns, adjectives and adverbs. For one-to-many and many-to-many token mapping an exclusive empirical rule layer has been designed on the token mapping layer in the GH-MAP system. The empirical rules resolve syntactic, semantic and structural divergence issues and improve translation quality of the GH-MAP system.

IV. EVALUATION OF GH-MAP SYSTEM

We have used two reference translations to evaluate candidate translation (output from the GH-MAP system).

- Reference_1 translation

We have asked the language expert to translate the documents word-for-word (i.e. word-based), retaining the flavour of the source language. Such a translation is feasible because Gujarati and Hindi are structurally similar languages.

- Reference_2 translation

The language model based translation is obtained from a bilingual corpus. The translation for each test sentence is extracted from bilingual corpus FIRE 2010, literature on Gandhiji and ELRA-W0037.

Automatic evaluation software has been designed to measure translation quality of the GH-MAP output (candidate translation).The software accepts file1 (the target language document generated by the GH-MAP, i.e. the candidate translation) and file2 (a reference translation) as an input file and gives the translation quality in terms of three scores, viz. PER, BLEU and METEOR. For example,

Input: *अहिंसा के द्वारा आर्थिक समानता कैसे लायी जा सकती है इसका विचार करें।*

Candidate translation[3]

Output: *અહિંસા વડે આર્થિક સમાનતા કઈ રીતે લાવી શકાય છે એનો વિચાર કરીએ.*

[3] Output of GH-MAP system

Sentence from Reference$_1$[4]

અહિંસા વડે આર્થિક સમાનતા કઈ રીતે લાવી શકાય છે એનો વિચાર કરીએ.

Sentence from Reference$_2$[5]

અહિંસા વડે આર્થિક સમાનતા કેમ લાવી શકાય એ વિચારીએ.

Table 4 shows that the candidate translation of an input sentence through GH-MAP is same as sentence of reference$_1$.

Table 4: Evaluation Score for Example

Hindi to Gujarati Translation			
	BLEU	PER	METEOR
Reference$_1$	1	0	1
Reference$_2$	0.29	0.89	0.74

4.1 Corpus level BLEU, METEOR and PER score with respect to Reference$_1$

It is observed that BLEU, METEOR and PER score are used for approximate assessment at a corpus level. It would perform badly if used to evaluate the quality of individual sentence. Considering this as a base we have divided test data into blocks where each block consists of 50 sentences. As shown in Table 4; we have calculated score with respect to the (word-based translation from human translators) reference$_1$ for individual sentence then we have calculated an average per block. Table 5 shows average score per block with respect to reference$_1$ translation and Figure 6 shows overall quality exploration.

Table 5: Evaluation Results for Hindi-Gujarati Machine Translation

Hindi to Gujarati Translation			
Candidate & Reference$_1$ translation			
Blocks of 50 Sentences	BLEU	PER	METEOR
1	0.93	0.06	0.97
2	0.88	0.11	0.95
3	0.79	0.15	0.92
4	0.95	0.05	0.98
5	0.93	0.06	0.97
6	0.92	0.06	0.98
7	0.77	0.14	0.92
8	0.77	0.14	0.92
9	0.85	0.14	0.94
10	0.75	0.16	0.92
11	0.85	0.12	0.94
12	0.94	0.04	0.98
13	0.97	0.05	0.98
14	0.92	0.04	0.96
15	0.96	0.04	0.99
16	0.88	0.11	0.95
17	0.77	0.14	0.92
18	0.93	0.06	0.97
19	0.95	0.05	0.98
20	0.92	0.06	0.98
Average Score	0.88	0.09	0.96

The above results of an average METEOR score 0.96 and PER score 0.09 represent that the candidate translation is reasonably close to the referece$_1$ translation. The BLEU is an accuracy measure and the average BLEU score is 0.88 because longer n-grams dominate over shorter n-grams in the BLEU score calculation.

[4] Word-based translation from human translators to produce the additional references

[5] Get from a parallel corpus, Language model based translation

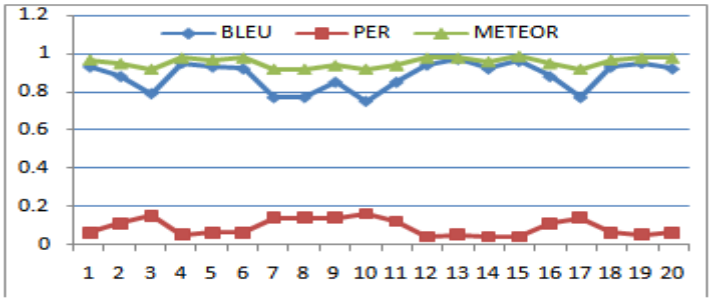

Figure 6 Evaluation Results for Hindi-Gujarati Machine Translation (Reference$_1$ and Candidate Translation)

4.2 Corpus level BLEU, METEOR and PER score with respect to Reference$_2$

Table 6 shows average score per blocks with respect to reference$_2$ translation and Figure 7 shows overall quality exploration

Table 6: Evaluation Results for Hindi-Gujarati Machine Translation

Hindi to Gujarati Translation			
Candidate & Reference$_2$ translation			
Blocks of 50 sentences	**BLUE**	**PER**	**METEOR**
1	0.02	1.19	0.28
2	0.19	1.18	0.5
3	0.15	1.15	0.5
4	0.11	1.1	0.4
5	0.2	1	0.5
6	0.12	1.13	0.4
7	0.09	1.21	0.32
8	0.11	1.15	0.42
9	0.16	0.98	0.5
10	0.18	0.99	0.51
11	0.24	0.93	0.54
12	0.22	0.86	0.63
13	0.31	0.61	0.7
14	0.12	0.94	0.6
15	0.25	0.72	0.69
16	0.19	1.18	0.5
17	0.09	1.21	0.32
18	0.2	1	0.5
19	0.11	1.1	0.4
20	0.12	1.13	0.4
Average Score	0.159	1.038	0.4805

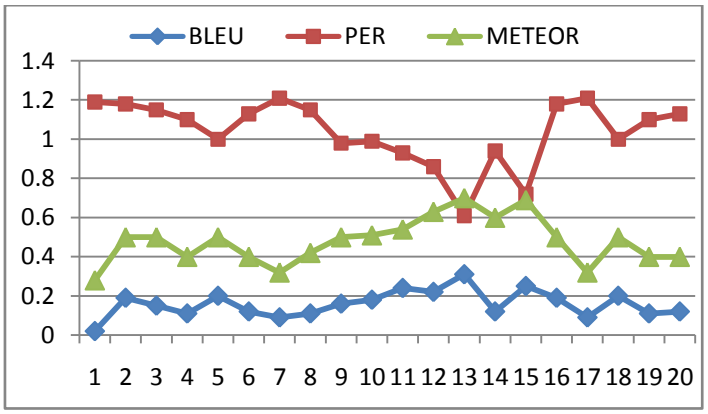

Figure 7 Evaluation Results for Hindi-Gujarati Machine Translation (Reference$_2$ and Candidate Translation)

As stated in [Ananthakrishnan et al. 2007], the BLEU is not the suitable evaluation method for "into-Hindi". Similarly METEOR and PER are also not suitable evaluation method for "into-Hindi" and "into-Gujarati". We have observed very poor score for reference$_2$ as shown in table 6 and figure 7. The prime reason for such poor score is 'into-Gujarati' language characteristics, such as free word-order, synonyms, case-markers, and morphological richness of Gujarati language. Basically, the 'into-Gujarati' language characteristics affect:
1) n-gram matching strategy of BLEU,
2) word-to-word matching of METEOR and
3) PER score due to variance in syntax.

4.2.1 Corpus level 'into-Gujarati' BLEU, METEOR and PER score with respect to Reference$_2$

From the error analysis of above results, we have performed following steps to incorporate 'into-Gujarati' characteristics on candidate translation and reference$_2$ translation before calculating BLEU, PER and METEOR score.

- Substitute synonyms
- Remove suffixes to generate stem
- Normalize token[6]
- Remove stop words
- Perform the word reordering

Table 7 and Figure 8 show quality exploration after performing above mentioned steps on 1000 sentences of the candidate translation and the reference$_2$ translation.

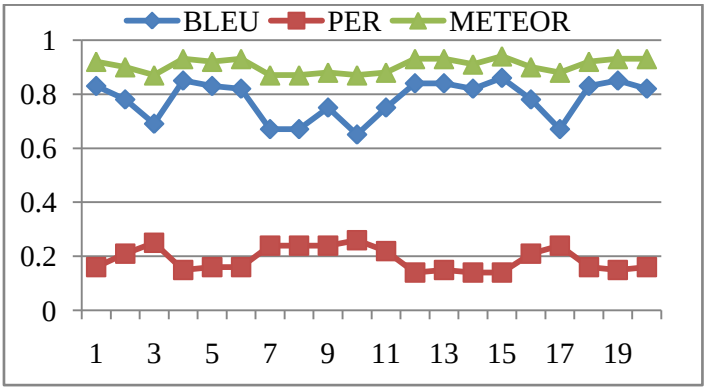

Figure 8 Evaluation Results for Hindi-Gujarati Machine Translation (Refined Reference$_2$ and Candidate Translation)

[6] Token normalization is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens.

Table 7: Refined Results of Hindi-Gujarati Machine Translation

Hindi to Gujarati Translation			
Candidate & Reference$_2$ (Keyword + Synonyms + Morpheme +Word reorder + Token Normalization)			
Blocks of 50 sentences	**BLEU**	**PER**	**METEOR**
1	0.83	0.16	0.92
2	0.78	0.21	0.9
3	0.69	0.25	0.87
4	0.85	0.15	0.93
5	0.83	0.16	0.92
6	0.82	0.16	0.93
7	0.67	0.24	0.87
8	0.67	0.24	0.87
9	0.75	0.24	0.88
10	0.65	0.26	0.87
11	0.75	0.22	0.88
12	0.84	0.14	0.93
13	0.84	0.15	0.93
14	0.82	0.14	0.91
15	0.86	0.14	0.94
16	0.78	0.21	0.9
17	0.67	0.24	0.88
18	0.83	0.16	0.92
19	0.85	0.15	0.93
20	0.82	0.16	0.93
average score	0.78	0.19	0.91

The above results resemble the results for the referece$_2$ translation with reference$_1$.

V. CONCLUSIONS

Paper presents experimental result to demonstrate the potential advantage and accuracy of the mentioned approach. It provides evidence for limited linguistic effort and tools for achieving the said goal [Patel K. and Pareek J, 2013]. Through experimental observation, it can be concluded that for the given test bed, significant result of BLEU, METEOR and PER score are obtained which proves the effectiveness of translation.

REFERENCES

[1]. Ananthakrishnan R, Bhattacharyya P, Sasikumar M, Shah R. 2007. Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU, In Proceedings of International Conference on Natural Language Processing.

[2]. Banerjee S.,Lavie A.2005.METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Association for Computational Linguistics.

[3]. Gandhiji M. 1999. Maaraa Svapna nu Bhaarat. Navjivan Prakashan Mandir.

[4]. Gandhiji M. 2007. Mangalprabhaat. Navjivan Prakashan Mandir.

[5]. Mark´o K, Schulz S, Medelyan O, Hahn U. 2005. Bootstrapping Dictionaries for cross-Language Information Retrieval, SIGIR.

[6]. Papineni K., Roukos S., Ward T, Zhu W.,2000. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL),pp.311-318.Philadelphia.

[7]. Patel K., Pareek J.2013. GH-MAP: translation system for sibling language pair Gujarati-Hindi, CSI Transactions on ICT, vol.1, issue 2, pp183-192

[8]. Tillmann C., Vogel S., Ney H., Zubiaga A., Sawaf H.1997. Accelerated DP-based search for statistical translation. In Proceedings of the European Conference on Speech Communication and Technology, page 2667-2670. Rhodes, Greece.