Chapter-3

3. Machine Translation from English to Hindi

3 Machine Translation from English to Hindi

A suitable machine translation engine will be helpful up to the certain extent. Translate the source content, when the native speaker wants to know the meaning of the source content written in another foreign language. The suitability of an engine may be identified by the language pair, and the domain of a source e-content supported by the engine.

3.1 Scope of English to Hindi MT Engine

As per the VIIIth schedule of Indian Constitution, there is no uniform national language of India. However, 23 languages are recognized as official languages. Hindi is recognized as the first official-language and the English is an associate language, narrated in NPTEL online course of language and society, (NPTEL, 2015).

Hindi is one of the best ranking languages internationally. It is on 4th rank as per the use of the languages by the native speakers (Wikipedia, 2010).

The quality translations are required by the learners in Hindi, which communicates the appropriate meaning of the source e-content. When such online translation engines are freely available for the native users than it is a real achievement of the digital world for the mankind. This study is targeted for the Indian learner. Those are conversant with Hindi language and want to learn technical contents, especially for computer science domain, present on the web.

3.2 Existing English to Hindi Engines

During reviews of resources, some freely available machine translation engines were found, which satisfy the need of native and global perspective users. The findings are classified under two categories. Engines those are developed in India and other in foreign countries. Further in each category engine are classified as general and domain specific.

Developed in India:

- **Mantra:** Designed for translation of Government appointment letters. It uses the 'Tree Adjoining Grammar' approach. Available in multiple and also in general domain.
- **Anuvadaksha:** This engine provides English to Hindi language translation for certain domains.
- **Angla Bharati:** This engine provides English to Hindi language translation for certain domains and also in general domain. An approach developed by IIT-Kanpur.
- **UNL based MT:** Developed by IIT-Mumbai.
- **Shiva:** An Example Based Machine Translation (EBMT) Developed by IIIT-Hyderabad and IISc-Bangalore.
- **Shakti:** An English-Hindi MT System developed by IIIT-Hyderabad.

Developed in Foreign Countries:

- **MyMemory:** Available in multiple domains.
- **Google-Translate:** General domain. Developed by Google.
- **MS-Bing:** General domain. Developed by Microsoft.
- **Worldlingo:** General purpose engine.
- **BabelFish:** General domain. Developed by Yahoo.
- **ImTranslator:** General purpose engine.
- **Babylon:** General purpose engine.

3.3 Translation Quality expectation

Quality of machine translated test is concerned with the native-ness of the words used as the output, and the sentence should convey the exactly same meaning as in the source text. The ultimate assessment is judged by those humans those who are the recipient of the target text. Various automated MT metrics are evolved over past years. However, the main challenge is how to compare the translated-output with reference text. In each comparison, we need at least one reference text on that basis; we may assess the quality of translation.

It is a very tedious task for any automatic evaluation method that how to get the reference text to judge the translated text. It is also obvious that if the machine translated texts are perfect, and then every word covers the native-ness.

3.4 Improvements Required in MT Quality

It has been observed by thorough reviews on existing MT engines. Why the translation quality produced by an individual engine is not appropriate?

- The target translation is lacking by fluency, sometime just because of word sequence.
- The translation does not have the adequate word lattice, which may be appreciated by the native learners.
- The change of meaning for an entire sentence is merely caused by the presence of context-less words.
- Domain specific large amount of corpus is required.

3.5 Metrics for MT Quality evaluation

Researchers have been developed many automatic evaluation methods for assessment of translation quality and discovered metrics to judge the quality of translation in different parameters. Machine translated e-content has to be fluent in reading and convey the exact meaning in the source content. It should be in same length of words as the source sentence. When the domain like computer science is the concern, it should not change the technical term and numerals in the translation and remain its phonetics in Hindi, for example, boot: बूट , CD-ROM: सीडी-रोम. Out of the list of evaluation metrics, few of them have been discussed here to provide the glimpse of metrics and adopt the most feasible metrics to evaluate our designed framework.

- BLEU (BiLingual Evaluation Understudy)
- F&A (Fluency & Adequacy)
- METEOR (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **Or**dering)
- NIST
- F-Measures
- TER (Translate Error Rate)

3.5.1 About BLEU

An automated evaluation metrics proposed by Papineni et al. (2002) is Bi Lingual Evaluation Understudy (BLEU). This is cost-effective and language independent method of machine translation evaluation. It is preferred when quick or frequent evaluation of translation is required with less support of skilled human evaluators.

The performance of machine translation is generally measured through its proximity with the translation done by a professional translator (Papineni et al., 2002).

The Bi Lingual Evaluation Understudy (BLEU) is considering the n-gram precisions of the hypothesis content translated by MT engine versus the reference content translated by the bilingual human evaluator/user of the system. The multiplication of the geometric mean (GM) of the cumulative n-gram precisions and the brevity penalty (BP) finally generates the BLEU score.

The process has been initiated by computing the geometrical average of the n-gram precisions p_n by using n-grams up to length of sentence N and positive weight w_n and summing to one where assumption as c is the length of the candidate translation and r is the effective reference length of the participating corpus. It is then followed by computing the brevity penalty BP (Papineni et al., 2002).

The BLEU score is also considered aspects like avoid unfair credits to the nonsensical hypothesis text. The counting of any word within the hypothesis text is not occurring or repeating in the reference text. A brevity penalty (BP) is the check for too short hypothesis text from receiving high-precision scores. The smoothing precision is monitoring that any sentence not get a zero score in the calculations. Precision and recall are computed by an equation-3.5.1.1 & 3.5.1.2.

Precision is the percentage of retrieved documents that are in fact relevant to the query. (Correct response)

$$Precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \qquad (3.5.1.1)$$

Recall is the percentage of documents that are relevant to the query and were in fact retrieved.

$$Recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|} \qquad (3.5.1.2)$$

Equation (3.5.1.3) and (3.5.1.4) illustrate that first brevity penalty is calculated, and after that the BLEU score is derived (Papineni et al., 2002).

$$BP= \left\{ \begin{array}{ll} 1 & \text{if } c > r \\ e^{(1-\ r/c)} & \text{if } c \leq\ r \end{array} \right\} \qquad (3.5.1.3)$$

$$BLEU = BP.Exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (3.5.1.4)$$

Where

n-gram precisions p_n = The number of (uni-grams), candidate translation words, which occur in any reference translation then divided by the total number of words of the candidate translation.

Positive weight w_n=1/N

An example is illustrated that how to calculate BLEU score and its required component in table 3.5.1.1.

Table 3.5.1.1 Illustration of BLEU score

Source E-content	Moses is a statistical machine translation system that allows you to automatically train translation models for any language pair.	N=19
Human Translation **(Reference)**	**मोसेस** एक **सांख्यिकीय** मशीन अनुवाद प्रणाली है जो **आपको स्वचालित रूप से** किसी भी भाषा **जोड़ी** के लिए अनुवाद मॉडलों **को प्रशिक्षित करने की अनुमति देता है।**	r=27
Translated by Anuvadaksha Engine (Candidate)	**मूसा** एक **स्टटिस्टिकाल** मशीन अनुवाद प्रणाली है जो किसी भी भाषा **जोड़ा** के लिए अनुवाद मॉडलों **स्वतः प्रशिक्षण देने के लिए आपको प्रदान करता है**	c=25

c= 25 Candidate Sentence (Translated Sentence)

r=27 Reference Translation (Human translation)

In above example r >c hence brevity penalty:

$\mathrm{BP}= e^{(1-\ r\,/\,c)}$

$\mathrm{BP}= e^{(1-\ 27/25)}$

$\mathrm{BP}= e^{(1-\ 1.08)}$

$\mathrm{BP}= e^{(-\ 0.08)}$

BP=0.923

p_n=13/25=0.52 (Refer table 3.5.1, candidate content)

log p_n= -0.284

w_n = 1/N=1/19=0.053

w_n log p_n= 0.053*(-0.284)=-0.231

This is for uni-gram calculation, and similar calculation has to be done for n=1…...N, up to n-gram.

The extension of mathematical expression of BLEU is explained as interactive BLEU (iBLEU) in the following section. Which is resulting, the score for above example is 0.21.

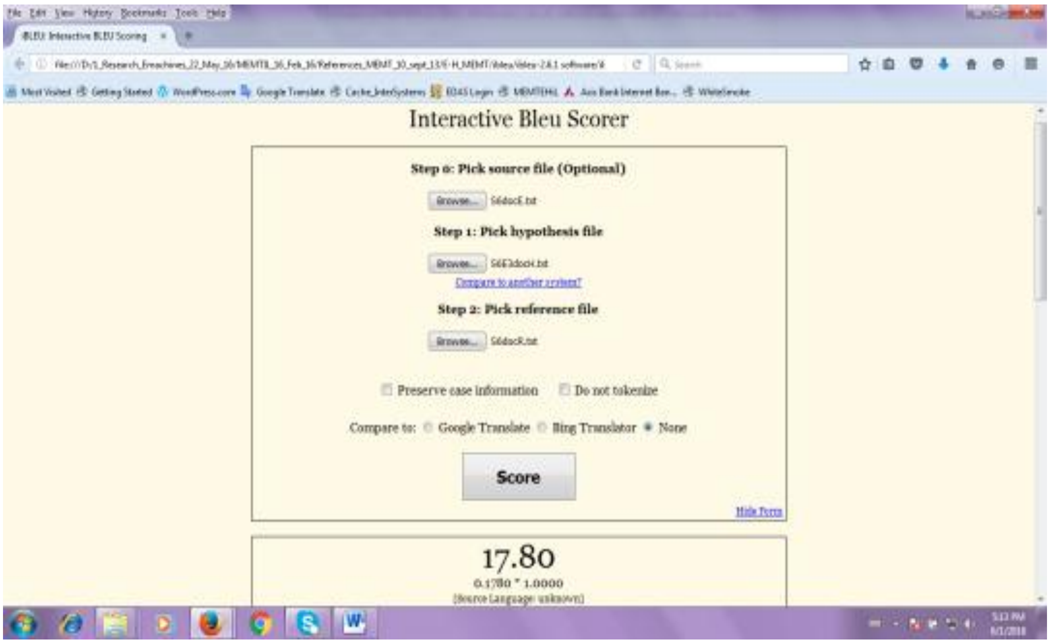
3.5.1.1 About iBLEU

An Interactive BLEU (iBLEU) is an Internet browser based mathematical tool, which performs the qualitative evaluation and comparison of machine translation in the form of numeric score.

Madnani (2011) developed a novel tool (iBLEU), which is a quick BLEU scoring tool in GUI environment. This tool is effective where system performance is poor (p. 213). It has been developed in JavaScript and automates the BLEU metric.

The best iBLEU score for the perfect translation should be 1.0. The hypothesis translation should exactly match with the reference text supplied to the iBLEU scorer. In this study, the threshold value of iBLEU is considered as 0.4, which is an average value computed during study based on human evaluator acceptability.

The interactive BLEU scorer for one of the randomly picked source e-content is displayed in Figure 3.5.1.1.1 (Madnani, 2011) where translation by "MS-Bing" MT engine is compared with the reference content.



Interactive Bleu Scorer

Step 0: Pick source file (Optional)

Browse... 56docE.txt

Step 1: Pick hypothesis file

Browse... 56E3doc4.txt

Compare to another system?

Step 2: Pick reference file

Browse... 56docR.txt

□ Preserve case information □ Do not tokenize

Compare to: ○ Google Translate ○ Bing Translator ● None

Score

Hide Form

17.80

0.1780 * 1.0000

(Source Language: unknown)

Figure 3.5.1.1.1 A Screen Shot of iBLEU Scorer. Adapted from Madnani, 2011.

Scoring Mechanism of Interactive BLEU (iBLEU)

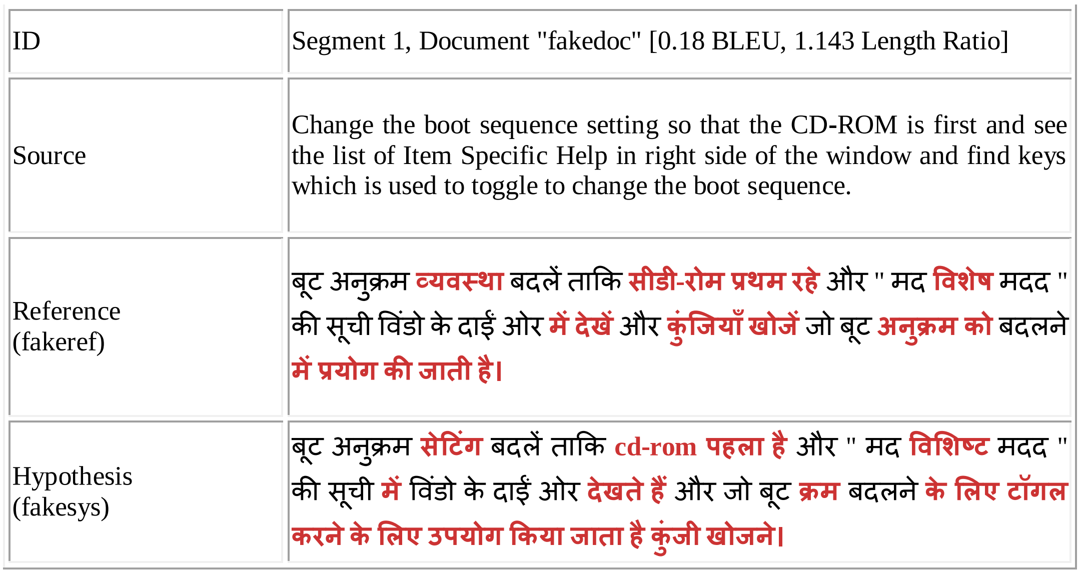
The following pre-requisites text files in (.txt Format) are required before starting the evaluation of any e-content from iBLEU interface:

- Source e-content (.txt) file in English language.
- Hypothesis (.txt) file in Hindi language, which is translated by MT Engine for which testing is being performed.
- Reference (.txt) file in Hindi language, generated by the system.

Once all input files are ready the iBLEU scorer interface picks files as required by the interface, and finally it generates the BLEU score. The best iBLEU score for perfect matching text would be 1.0.

An extended screen shot of an iBLEU scorer shown in Table-3.5.1.1.1, shows BLEU score of hypothesis content. Source, reference and hypothesis e-contents are also displayed. A contributor MT engine "MS-Bing" is demonstrated in this test case.

Table- 3.5.1.1.1 iBLEU results with MS-Bing Engine



ID	Segment 1, Document "fakedoc" [0.18 BLEU, 1.143 Length Ratio]
Source	Change the boot sequence setting so that the CD-ROM is first and see the list of Item Specific Help in right side of the window and find keys which is used to toggle to change the boot sequence.
Reference (fakeref)	बूट अनुक्रम **व्यवस्था** बदलें ताकि **सीडी-रोम प्रथम रहे** और " मद **विशेष** मदद " की सूची विंडो के दाईं ओर **में देखें** और **कुंजियाँ खोजें** जो बूट **अनुक्रम को** बदलने **में प्रयोग की जाती है।**
Hypothesis (fakesys)	बूट अनुक्रम **सेटिंग** बदलें ताकि **cd-rom पहला है** और " मद **विशिष्ट** मदद " की सूची **में** विंडो के दाईं ओर **देखते हैं** और जो बूट **क्रम** बदलने **के लिए टॉगल करने के लिए उपयोग किया जाता है कुंजी खोजने।**

3.5.2 Fluency and Adequacy (F&A)

Translation quality is often described in terms of fluency and adequacy (F&A). According to (Ma & Cieri, 2006), Fluency reflects the 'native-ness' of the translation while adequacy indicates how well a translation captures the meaning of the original text. A fluent sentence is one that is grammatically framed, contains correct spellings, strictly use of common terms, titles and names. Furthermore, intuitively acceptable and can be sensibly interpreted by a native speaker of Hindi.

The objective of fluency evaluation is to determine how much like "good fluent" a translation appears to be, without taking into account the correctness of the information. In other ways, fluency refers to the grammatical accuracy of the translated text. The evaluation has done segment-by segment on a 0-5 scale without referring to any reference-text. The objective of the adequacy is to determine the extent to which a text is conveyed its meaning regardless of the quality of the language in the translation. Simply say the adequacy is defined as the degree to which the reference sentence is conveyed in the translation. The evaluation has done segment-by-segment on 0-5 point scale.

As mentioned by (Chris Callison-Burch et al., 2001), the translation industries are following the fluency, adequacy and translation-error-classification criteria, to identify the quality of machine translation. A different aspect is defined by (Papineni et al. 2002), while evaluating the translated text (candidate text) using the same words (1-grams) as in the reference text tends to satisfy the adequacy. The longer n-gram matches accounts for fluency of the translated-text.

In statistical machine translation (SMT), the quality of translation is based on availability of qualified corpora. Snover & Madnani (2009) explained that the translation is fluent, it meaning it reads well and therefore, meets user expectations. In RBMT dictionary based, customization can improve quality but translation results may lack the fluency reader's expectation. An automatic MT metrics measured varying aspects of MT performance, which cover the different types of human judgments such as Fluency, Adequacy (Snover et al., 2009).

Scaling of fluency can be measures in the 5-point as per the following:

5:	Flawless
4:	Good
3:	Non native target language
2:	Disfluent
1:	Incomprehensible

Adequacy refers to the degree to which information present in the original is also communicated in the translation.

5:	All
4:	Most
3:	Much
2:	Little
1:	None

The human translators of native language have been preferred to judge the correctness of translation in Hindi. The selected translators were well versed with English-Hindi Language pair. The experimental reference for evaluation of multi engine system is influenced from (Fuji Ren & Hongchi Shi et al., 2001). The best fluent and adequate translation by human expert has been taken into the consideration as reference text. Translators, whose mother tongue is Hindi, were used for judgment about the correctness of the machine engine generated translations. The correct translation accounted in the testing if more than two human translators judged the same correctness.

3.5.3 METEOR (**M**etric for **E**valuation of **T**ranslation with **E**xplicit **Or**dering)

METEOR is an automatic metric for machine translation evaluation as explained by Banerjee & Lavie (2005). It is based on a generalized concept of unigram matching between the machine produced translation and human produced reference translations. It can be matched based on their surface forms, stemmed forms, and meanings. Furthermore, METEOR can be easily extended to include more advanced matching strategies. Once all generalized unigram matches between the two strings have been found. METEOR computes a score for this matching using a combination of unigram-precision and recall. The measure of fragmentation that is designed as how well-ordered the matched words in the machine translation are in relation to the reference. METEOR was designed to explicitly address the weaknesses in BLEU. It evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation. If more than one reference translation is available, the given translation is scored, against each reference independently, and then best is captured.

(Denkowski & Lavie, 2010), was experimented with a machine translation hypothesis and a reference translation. The traditional METEOR metric calculates a lexical similarity score based on a word to word alignment between the two strings (Banerjee & Lavie, 2005). When multiple references are available, the hypothesis is scored against each, and the highest value reference is used.

The final alignment is defined as the union of all stage of alignments. Once an alignment has been constructed, the total number of uni-gram matches (m), the number of words in the hypothesis (t), and the number of words in the reference (r) are used to calculate precision (P = m/t) and recall (R = m/r). Then, parameterize harmonic mean of P and R is calculated. Fmean is calculated by using Equation-3.5.3.1 (C. Van Rijsbergen, 1979).

$$\text{Fmean} = \frac{10PR}{R + 9P} \tag{3.5.3.1}$$

Precision, recall and Fmean are based on unigram matches. To take into account longer matches, METEOR computes a penalty for a given alignment by Equation-3.5.3.2 (Banerjee & Lavie, 2005).

$$Penalty = 0.5 * \left(\frac{\#\ Chunks}{\#\ Unigrams\_matched}\right) \qquad (3.5.3.2)$$

The final METEOR score is then calculated by Equation-3.5.3.3.

$$\text{Score} = \text{Fmean} * (1 - \text{Penalty}) \qquad (3.5.3.3)$$

Metric for Evaluation of Translation with Explicit Ordering (METEOR) as explain by (Kalyani et al., 2014). It is based on a word-to-word alignment between the machine-generated translation and the reference translation. Every uni-gram in the test translation should map to zero or one unigram in the reference sentence. If there are two alignments with the same number of mapping, the alignment is chosen with a lesser number of intersections of the two mappings. The score is equal to the harmonic mean of unigram precision and recall. METEOR consists of:

1) Exact Match mapping words that are exactly same;
2) Stem Match links words that share the same stem;
3) Synonym Match mapping unigrams that are synonyms of each other.

METEOR-Hindi includes following additional modules to make more efficient for Hindi (Gupta, Venkatapathy & Sangal, 2010):

1) **The Local Word Group (LWG)** consisting of a content word and its associated function words.

2) **Clause Match -** Clause is defined as a phrase containing at least a verb and a subject.

3) **POS matcher** compute the number of matching words with same POS tag.

3.5.4 NIST

Doddington was established NIST in 2002, which is similar to BLEU. However, it assigns a weight to each uni-gram depending upon its uniqueness or how informative is the n-gram. It uses Arithmetic mean rather than the geometric mean. The formula for NIST is presented in equation 3.5.4.1 (Doddington, 2002).

$$Score = \sum_{i=1}^{n} \left\{ \sum_{all\ w_i \dots\ w_n\ that\ co-occur} \frac{p\ (W_i \dots\ W_n\ )}{\sum_{all\ w_i \dots\ w_n\ in\ system\ output}\ (1\ )} \right\} . exp \left\{ \beta\ log^2 \left[min \left(\frac{L_{sys}}{Lref}, 1 \right) \right] \right\}$$

(3.5.4.1)

The information weight is computed by equation 3.5.4.2.

$$Info\ (w_i \dots .\ w_n\ ) = log_2 \left(\frac{the\ \#\ of\ occurences\ of\ \ w_1 \dots .\ w_{n-1}}{the\ \#\ of\ occurences\ of\ w_i \dots .\ w_n} \right)$$

(3.5.4.2)

3.5.5 F-Measure

Turian et al. (2003) were established the GTM (General Text Matcher). It uses the sharing of matched words between MT output and reference. The evaluation score is obtained, and it is not only based on precision and recall but also on the harmonic mean of precision and recall. It is known as F-measure proposed by Turian et al. (2003) and calculated by an equation-3.5.5.1.

$$F\ measure = \frac{2PR}{P + R}$$

(3.5.5.1)

3.5.6 TER

TER (Translate Error Rate) was proposed by Snover et al. (2006). It represents the number of edits necessarily required to transform the machine output as compare to the reference translation. That is normalized to the length of the reference-text (Snover et al., 2006). The TER is computed by an equation-3.5.6.1.

$$TER = \frac{Number\ of\ edits}{Average\ of\ reference\ words} \qquad (3.5.6.1)$$

3.6 Conclusion

The native learner of India wants to translate the E-contents which is generally available in foreign language. The Hindi language is widely used as first language by the Indians. Therefore, the need of translation from English to Hindi is very high. The accessibility of the freely available engine is the path lighting tool for the learner of English to Hindi. The quality of translation which is actually expected by the learners is far away for the general domain. Whereas, in the computer science domain, there is no specific engine for English to Hindi is available that satisfy the translation need of native learners. Even though it is tedious to judge the translated contents based on its technical correctness. Certain evaluation metrics has been selected, that will be useful for the evaluation of our designed machine translation system. We have earmarked a combination of human and automated metrics. Fluency & Adequacy (F&A) as human evaluated and iBLEU as automated-metrics has been selected. Now as a conclusion, we found few machine translation engines; those were performed better in other languages have been selected for English to Hindi translation as component engines for our proposed framework.