ISSN 2007-9737

Using Linguistic Knowledge for Machine Translation Evaluation with Hindi as a Target Language

Samiksha Tripathi, Vineet Kansal

AKTU, Lucknow,
India

samiksha.tripathi@gmail.com

Abstract. Several proposed metrics of MT Evaluation like BLEU have been criticized for their poor performance in evaluating machine translations. Languages like Hindi which have relatively free word-order and are morphologically rich pose additional problems in such evaluation. We attempt here to make use of linguistic knowledge to evaluate machine translations with Hindi as a target language. We formulate the problem of MT Evaluation as minimum cost assignment problem between test and reference translations with cost function based on linguistic knowledge.

Keywords. Machine translation evaluation, linguistic knowledge, word group matching, cost function.

1 Introduction

In recent years, there has been much work in machine translation keeping in mind the need of automatic tools to translate text in one language to the equivalent text in another language. As existing systems for Machine translation are not competent to the human translators, there is huge scope for improvement. Evaluating the performance of these systems is of key concern to their development.

One of the major issues in the performance evaluation of machine translation systems is that of measuring the quality of translated text. The problem of Machine Translation Evaluation is therefore central to the Machine translation systems.

The problem of machine translation evaluation is stated as the problem of evaluating how close the machine translation is to the human translations. Evaluating machine translations by humans is inherently slow and costly process. Considering the growth of machine translation research, tools for automatic MT evaluation are of prime concern to NLP and specifically to MT research.

There have been several attempts to design a metric for automatic MT Evaluation. The proposed metrics try to measure the similarity or semantic closeness between candidate machine translation and the actual reference translation. In (Papineni et al., 2002), a metric, BiLingual Evaluation Understudy (BLEU) based on n-gram matching is proposed. However, it does not take into account recall i.e. what fraction of reference translation the candidate translation covers. A variation of BLEU with frequency weights is proposed in (Babych and Hartley, 2004). It tries to improve BLEU by giving appropriate weights to different n-grams in the sentence which BLEU treats equally.

In (Culy and Riehemann, 2003) limitations of n-gram matching based methods are pointed out in general. The limitations apply to BLEU as well. Another metric, METEOR, proposed in (Banerjee and Lavie, 2005), tries to overcome some of the limitations of BLEU by taking into account precision and recall. It also uses fragmentation penalty to account for wrong word-order. It also incorporates three levels of matching n-grams viz. Exact, stem and WN Synonymy match. In (Lavie and Agarwal, 2007), the performance of METEOR is improved by refining the parameters for specific languages. (Lavie and Denkowski, 2009) discusses the development of METEOR.

MAXSIM proposed in (Chan and Ng, 2008), also uses POS information along with lemma and

doi: 10.13053/CyS-21-4-2869

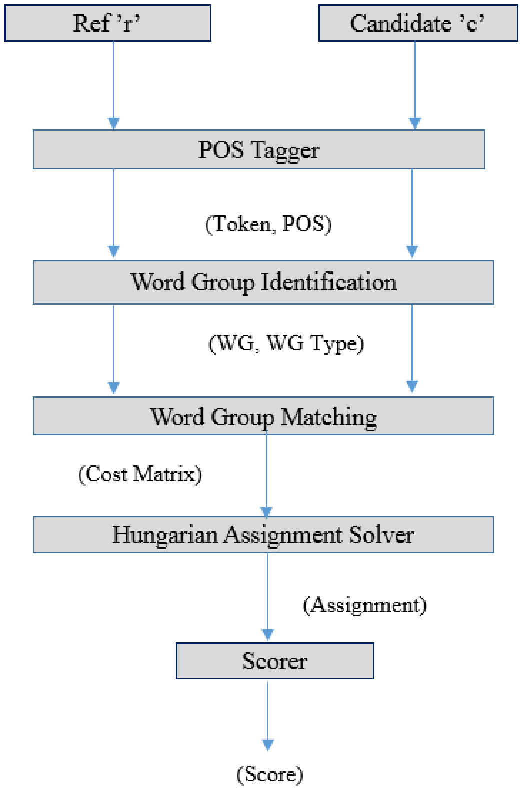

Fig. 1. Framework of Evaluation Metric

Table 1. Word Groups: Types, Heads Regular Expressions with Examples

Word group (Head)	Regular Expression (Example)
NN+PSP (NN)	NN(PSP)+ e.g. paSuoN ke lie (For animals)
ADJ+NN (NN)	(ADJ)+NN e.g. sunDar jagah (Beautiful place)
ADJ+VM+VAUX (VM)	(ADJ)*VM(V AUX)* e.g. kiyA gayA hai (has been done)

synonym information. It introduces the concept of Bipartite graph match to find the maximum matching based on Synonym match scores. It also suggests the use of dependency relations to improve the performance.

Most of the work with respect to MT Evaluation is done for Western languages. However, very little work is attempted for Indian languages. In (Aanthkrishnan et al., 2007), the author has pointed out the problems with BLEU in the evaluation of English- Hindi indicative translations. The paper also describes the divergences between English and Hindi causing BLEU to fail.

In (Chatterjee et al., 2007), the authors suggest the use of Word-Group Identification and word-group matching instead of n-gram matching as the way to improve the performance of BLEU in case of Hindi. The word-group refers to the sequence of words with fixed internal order. This takes care of free word order among the word-groups to some extent. However, they calculate the final score in the same way as does the METEOR.

Considering all these issues with the use of BLEU, METEOR in Hindi translation evaluations, we attempt to design the new metric for MT Evaluation especially for Hindi which we describe in the next section.

2 Framework of Evaluation Metric

In the proposed metric, we attempt to make use of linguistic knowledge of Hindi at various levels. First, we tokenize both the reference and test sentences into words. The POS tags are obtained using standard Hindi POS tagger (IITB). Using this POS information we identify the word-groups as discussed in subsection 2.1.

We formulate the problem of evaluation as the minimum cost assignment on a bipartite graph where the two sets of nodes represent the word groups in reference and test translations.

The weights are assigned to edges between each node in one set to every other node in the other set of this bipartite graph so as to represent the semantic dissimilarity between the reference word group and the test word group.

Special nodes are added to make the no. of nodes in the two sets equal and the weights of edges associated with such nodes are set to 1. The weights are assigned as described in subsection 2b and 2c. The overall framework is described in Section 2 (this section). Each of the component of the framework is described in detail in the following sections.

Table 2. Word matching: Different levels and scoring scheme

Match level	Match Score
Exact + POS	1.00
Lemma + POS	0.75
Syn + POS	0.50
Exact only	0.75
Stem only	0.50
Syn only	0.25
Otherwise	0.00

Table 3. PSP matching: scoring scheme

PSP Match	Match score
Exact	1.0
Strong Equivalence	0.8
Weak Equivalence	0.6
No Equivalence	0.0

Table 4. PSP Equivalence: Strong and weak equivalence of post positions

PSP Group	Strongly Equivalent	Weekly Equivalent
ke kAraN (because of)	kI vajah se	ke falswarup
ko (to)		ke lie (for)
ne	-	-

2.1 Identification of Word Groups

Our notion of word group is limited to the group of words having strict internal order.

This notion of word groups is described in (Chatterjee et al., 2007). In the current work, we identify the word groups based on the POS information. Specifically, we consider three different types of word groups viz. Noun-Post-positions, Adjective- Noun and Verb groups. Verb group includes just verbs, verbs with auxiliary verbs and verbs with associated adjectives. In the recent work (Gupta et al., 2010), METEOR is modified to handle Noun & Post-position group. However, we are dealing with a richer set of word groups. We use a simple FSM detecting simple regular expressions describing these word groups. Table 1 summarizes the different types of word groups along with their matching regular expressions, head part and representative example.

a. Matching Words

During word matching, we try to match the two given words at three different levels: Exact match, Lemma match, Synonym match and each such match receives a score depending on the type of match.

For finding Lemma and Synonym match we make use of Hindi WordNet (Jha et al., 2001). The two words are considered to be matched as synonyms if they have at least one common sense in the word net. For each of these levels we consider two sublevels: one where POS tags match and the other where they don't match. The Table 2 shows the scores for each type of match.

As shown in Table 2, we give increasing penalty as the level of match goes down. We give highest preference to exact match, then to Lemma match and then to Synonym match. Only POS matching doesn't make sense when the words do not have exact, lemma or synonym match. However, there may be changes in POS tags of a word depending on its usage. So, we consider the three types of match without POS match by giving a little penalty (0.25).

b. Word Group Matching

To match word groups, we try to match each word of the word group with each word in the other group. We match the word group only if the head part of the word groups match at least at some level of word matching.

Also, the matching of head part carries more weight in the overall score assigned to word group.

Let wg_r be the reference word group and wg_t be a test word group. Let, $wg_r(i)$ be the *ith* word of wg_r and $wg_t(j)$, be *jth* word of wg_t. Let, $wg_r(h)$ and $wg_t(h)$ be the head words of the two word groups. Then the match score of the two word groups is calculated as follows:

1. Each word in wg_t , $wg_t(i)$ is matched with every word $wg_r(j)$ of wg_r and a score

$(w(wg_t(i); wg_r(j))$ is obtained. The word score for $wg_t(i)$ w.r.t wg_r is given as:

$$S_w(wg_t(i); wg_r) = max_j S\big(wg_t(i); wg_r(j)\big). \quad (1)$$

2. If $S_w\big(Wg_t(h)\big) > 0$ then the score of word group Wg_t w.r.t. word group Wg_r is given as:

$$S_{wg}(wg_t; wg_r) = \frac{\big(\alpha * S_w(wg_t(h), wg_r) + \sum_{i \neq h} S_w(wg_t(i), wg_r)\big)}{\alpha + n_t - 1}, \quad (2)$$

where n_t the no. of words in the word group Wg_t and α is decides the importance of head part in the word-group matching. Otherwise, $S_{wg}(wg_t; wg_r)$ is regarded as 0.

$S_{wg}(wg_t; wg_r)$ takes care of additive errors whereas $S_{wg}(wg_t; wg_r)$ takes care of deletion errors. The former will have smaller value when there are elements in wg_t which don't match with any of the elements in wg_r, whereas the later one will have smaller value if there are elements in wg_r which do not match with any of the elements in wg_t.

3. The final match score of wg_t and wg_r is calculated as:

$$MS_{wg}(wg_{t,} wg_r) = \frac{S_{wg}(wg_r, wg_t) + S_{wg}(wg_t, wg_r)}{2}. \quad (3)$$

Averaging the two gives equal weight to the types of errors captured by each of them.

Note that S_{wg} is asymmetric whereas MS_{wg} is symmetric.

c. PSP Equivalence

Post-Positions play an important semantic role in Hindi. Some Post-positions are almost replaceable by each other. We call such post-positions as strongly equivalent. Similarly, some postpositions can sometimes be replaced by some other positions. These are called as weakly equivalent. So, a post-position Pi is either strongly equivalent, weakly equivalent or non-equivalent to other postposition Pj. This equivalence is not symmetric. We make use of this PSP equivalence information to add one more level of matching for NN+PSP word groups. The following Table 3 shows the scoring scheme for PSP matching.

To account for the PSP equivalence, we modify the step 2 in word-group matching as follows:

$$S_{wg}(wg_t; wg_r) = \frac{(\alpha * S_w(wg_t(h), wg_r) + S_{psp}(wg_t(psp), wg_r))}{\alpha + 1}. \quad (4)$$

Here $S_{psp}(wg_t(psp), wg_r)$is given according to Table 3 depending on whether wg_r contains strongly equivalent, weakly equivalent or nonequivalent PSP for $wg_t(psp)$.

Table 4 shows the strongly equivalent and weakly equivalent Post positions for some of Hindi post positions. In Hindi, the postposition *ke kAraN* (meaning "because of") is almost always replaceable by another postposition *kI vajah se* without affecting the meaning. However, it can be replaced by another postposition *ke falswarup* (meaning "as a result of") in some of the situations but not all. Similarly, *ko* doesn't have any strongly equivalent postposition but is weakly replaceable by *ke lie*.

d. MT Evaluation as Minimum Cost Assignment

As described in the beginning of section 2, we formulate the problem of MT Evaluation as the minimum cost assignment problem on a bipartite graph. The two sets of nodes represent the word groups in the reference and test sentences respectively. The weight assigned to the edge joining the reference word group wg_r to the test word group wg_t is given as:

$$w(wg_t, wg_r) = 1 - MS_{wg}(wg_t, wg_r). \quad (5)$$

We solve the problem using Hungarian Assignment Solver. The final evaluation score of reference sentence t is obtained as:

$$eval\_score(t, r) = 1 - ass\_cost(G(t, r), w), \quad (6)$$

where

$$\text{ass\_cost} = \frac{\sum_{wgt} w(wg_{t,} assigned(wg_t))}{\max(N_{t,} N_r)}, \quad (7)$$

where $assigned(wg_t)$ represents the referenceword group wg_r to which wg_t is assigned in the minimum cost assignment by

Table 5. Cost-Matrix: Example of Cost Matrix for given pair of reference and test translations

Ref WG **(Ref Gloss)**	**kripayA** **(Please)**	**nimnaliKit nirdeSo.N kA** **(following instructions of)**	**pAlan** **kare.N** **(follow)**
Test WG (Test Gloss)			
Kripiya (Please)	0	1	1
ke (in)	1	1	1
rUp (the form)	1	1	1
me.N (of)	1	1	1
die gaE (given)	1	1	1
nirdeSo.N kA (of instructions)	1	0.125	1
pAlan kare.N (follow)	1	1	0

Table 6. Human Evaluation: Scoring Scheme

Quality of Translation	**Score**
Ideal	4
Good	3
Acceptable	2
Not Acceptable	1
Non-Sense	0

Hungarian solver and $N_{t,}$ and N_r are the no. of nodes in the two sets of bipartite graph $G(t,r)$ defined over.

e. Overall Penalty Factors

It is observed that whenever PSP part of Noun+PSP group in reference sentence is replaced by other non-equivalent PSP or is missing in the test sentence, the meaning of the sentence is affected and hence such sentences receive very low scores during human judgment even though the rest of the sentence considerably matches with the reference. To address this issue, we additionally penalize such test sentence by multiplying its evaluation score by overall penalty factor between 0 to 1. Currently, the overall penalty factor is set to 0.75.

3 Results

To understand how our metric evaluates a sentence, consider the cost matrix shown in Table 5 for a given pair of reference and test translations.

For the sake of simplicity, dummy nodes (added to make cardinality of two sets equal) are not included in the cost matrix as all entries corresponding to these nodes are 1. In Table 5, the columns correspond to Word-groups in reference translations whereas rows correspond to Word-groups in test translations. Note that, the entry for *kripaya* in ref WG and *kripaya* in testWG is 0 as they two match exactly. Similarly, the entry 0.125 captures the penalty for missing element *nimnaliKit* in the reference translation.

To test the performance of our metric, we arranged some English sentences, their test translations and reference human translations. We also obtained human evaluations for the gathered test translations. Subsection 3.1 describes the process of human evaluation in detail.

We test our metric through three different experiments. In the first experiment we use around 50 simple English sentences and their Hindi translations. The sentences were not pertaining to any specific domain. We obtained evaluation from three different native Hindi speakers. We also obtained the Hindi translations of these English sentences from a native Hindi speaker other than the three involved in the evaluation. These translations are treated as reference translations.

For the second experiment, we collected around 150 English sentences from various domains such as Parliament, Agriculture, Insurance, and Government along with their standard Hindi reference translations.

In both these experiments, we use test translations, Hindi translations of English sentences obtained from freely available Google Translate (Google, https://translate.google.com) English-Hindi translation engine. We use human translations of English sentences as reference translations for both of these experiments.

In the third experiment, we take some of the reference translations from second experiment and generate some test translations per reference translations. The sentences are generated by making certain kinds of changes in the reference translations. These changes include addition of words, deletion of words, replacement of a word by its morphological variation, replacement of a word by its synonymous word, valid word-reordering, invalid reordering, replacing PSP's by its strongly equivalent, weakly equivalent and non-equivalent PSPs. We use these artificially generated sentences as test sentences. This experiment was aimed at analyzing the ability of the proposed metric to capture the typical errors in the machine translation systems.

3.1 Human Evaluation

We presented English sentence and their corresponding Hindi translation to human subjects and asked them to evaluate it on 5-point scale (0-4) as shown in Table 6. The reference translations used for automatic evaluation were kept hidden from the subjects. This ensures that the human judgment is not biased towards a single reference. We gathered such evaluations from three different subjects who are the native Hindi speakers.
The average score given by the subjects is considered as the overall human judgment.

a. Evaluation Results

We correlate the scores of the proposed metric with human scores. We also compare the results with METEOR with only exact match. Table 7, 8 displaying correlations1 and correlations2 with the summary of system level scores and correlations of the metric scores with human judgment for the two experiments we conducted.

As shown by Table 7, the proposed new metric performs extremely Good. The correlation of this metric with human judgment is almost double the correlation of METEOR. However, as shown in Table 8, we get only marginal improvement over METEOR for complex set of sentences. As our metric is based on other sources of information like Hindi WordNet and POS Tagger, the success of metric largely depends on accuracy of these sources. Also, set of complex sentences include many technical terms which are transliterated

Table 7. Exp. 1 Correlation Results: System level scores and Correlation with Human judgments for different metrics for Experiment 1

Metric	Scores	Coeff. Of Corr.
METEOR	0.5503	0.3576
Our Metric	0.5441	0.6833
Human	0.4620	-

Table 8. Exp. 2 Correlation Results: System level scores and Correlation with Human judgments for different metrics

Metric	Scores	Coeff. Of Corr.
METEOR	0.4419	0.3135
Our Metric	0.4036	0.3374
Human	0.5363	-

Table 9. Artificial Tests: Results of artificially generated test sentences

Test Case	METEOR Score	Our Metric Score
Invalid movement of Word- Group	0.9993	0.9375
Invalid movement within word-group	0.9960	0.8604
Valid movement of word-group	0.9993	1.0000

during automatic translation but have appropriate Hindi translations in reference translation. Also, the sentences being complex their translations were not so good and mostly received low scores during Human judgments.

The results of the third experiment are shown in Table 9. The analysis of these results shows that the proposed metric is capable of differentiating between valid and invalid word orders in Hindi. METEOR penalizes both of them equally considering that both are equally invalid. However, the proposed metric penalizes only invalid reorder but favors the valid reorder by performing matching at word-group level. Also, our metric favors the translation when a post-position is substituted by another equivalent post-position. METEOR doesn’t take into account such equivalence at all though paraphrase matching is proposed (Lavie and Denkowski, 2010).

4 Conclusion and Future Work

We have attempted here to make use of linguistic knowledge at various levels for MT Evaluation with Hindi as target language. Specifically, the proposed metric uses the knowledge about Hindi through POS and Synonymy information, word group identification and PSP equivalence. We claim that use of such linguistic information in addition to statistical one helps in evaluating translations better by showing higher correlation of our metric with human judgments. Though the approach makes the metric language specific, the use of similar linguistic knowledge in other languages is expected to help the evaluating translations in that language. However, the form of knowledge that can be used may differ from language to language e.g. PSP equivalence which is useful for Hindi may not be useful for other languages. Though framing the problem of evaluation as minimum cost assignment problem makes it computationally slower, the complexity can be reduced by making use of some heuristics.

In the current work, we have not experimented with various parameters such as specific scores to be assigned and the specific weight to be given for matching head of the word-group. Also, the weight of the head is likely to change from word group to word-group. The weight for NN in NN+PSP may be different for weight for NN in ADJ+NN. As these scoring functions are crucial to the performance of our metric, experimenting with these parameters is the main focus of future work.

It would be nice to study how human evaluation process works and model the automatic process likewise. Such approach would definitely take MT Evaluation to the level of human evaluation. Also, we would like to extend the idea of MT evaluation to answer Church-Turing hypothesis where we want to distinguish whether a given test translation is human translated or automatically translated by a machine.

References

1. **Bogdan, B. & Hartley, A. (2004).** Extending the BLEU MT evaluation method with frequency weightings. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL ’04,* Stroudsburg, PA, USA. Association for Computational Linguistics. DOI: 10.3115/1218955.1219034.
2. **Satanjeev, B. & Lavie, A. (2005).** METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, Association for Computational Linguistics.
3. **Yee-Seng, Ch. & Hwee-Tou Ng (2008).** MAXSIM: A maximum similarity metric for machine translation evaluation. *Proceedings of ACL: HLT*, pp. 55–62, Association for Computational Linguistics.
4. **Niladri, Ch., Johnson, A., & Krishna, M. (2007).** Some improvements over the BLEU metric for measuring translation quality for Hindi. *ICCTA*, pp. 485–490.
5. **Culy, C. & Riehemann, S. Z. (2003).** The limits of N-Gram Translation Evaluation Metrics. *Proceedings of MT Summit IX.*
6. **Google Translate. (2017).**
7. **Gupta, A.S., Venkatpathy, R., & Rajeev, S. (2010).** Meteor-Hindi: Automatic MT evaluation metric for Hindi as a target language. *Proceedings of 8th International Conference on NLP*, pp. 178–187.
8. **IITB, CFILT.** *CRF-based Hindi POS tagger*. http://www.cfilt.iitb.ac.in.
9. **Jha-Narayan, S.D., Pande, P., & Bhattacharya, P. (2001).** *A Wordnet for Hindi*. http://www.cfilt. iitb.ac.in
10. **Alon, L. & Agarwal, A. (2007).** Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. *Second Workshop on*

Statistical Machine Translation at ACL'07, pp. 228–231.

11. **Alon, L. & Denkowski, M. (2009).** Meteor metric for automatic evaluation of machine translation. *Machine Translation*, Vol. 23, pp. 105–115. DOI: 10.1007/s10590-009-9059-4.
12. **Alon, L. & Denkowski, M. (2010).** Extending the Meteor machine translation evaluation metric to the phrase level. *HLT-NAACL*, pp. 250–253.
13. **Kishore, P., Roukos, S., Ward, T., & Wei-Jing, Z. (2002).** Bleu: a method for automatic evaluation of machine translation. *ACL*, pp. 311–318.
14. **Ananthkrishnan, R., Pushpak, B., ShashiKumar, M., & Ritesh, M.S. (2007).** Some issues in automatic evaluation of English-Hindi MT: more blues for blue.
15. **Kalyani, Aditi, et al. (2014*).*** Assessing the Quality of MT Systems for Hindi to English Translation*. arXiv preprint* arXiv:1404.3992.
16. **Bharati, A., Chaitanya, V., & Sangal, R. (1991).** Local word grouping and its relevance to Indian languages. **Bhatkar V.P. & Rege, K.M. (eds.)** *Frontiers in Knowledge Based Computing (KBCS90),* Narosa Publishing House, New Delhi, pp. 277–296.

Article received on 28/12/2016; accepted on 21/02/2017. Corresponding author is Samiksha Tripathi.