Incorporating Machine Learning Techniques in MT Evaluation

Nisheeth Joshi, Iti Mathur, Hemant Darbari, and Ajai Kumar

Abstract. From a project manager's perspective, Machine Translation (MT) Evaluation is the most important activity in MT development. Using the results produced through MT Evaluation, one can assess the progress of MT development task. Traditionally, MT Evaluation is done either by human experts who have the knowledge of both source and target languages or it is done by automatic evaluation metrics. These both techniques have their pros and cons. Human evaluation is very time consuming and expensive but at the same time it provides good and accurate status of MT Engines. Automatic evaluation metrics on the other hand provides very fast results but lacks the precision provided by human judges. Thus a need is being felt for a mechanism which can produce fast results along with a good correlation with the results produced by human evaluation. In this paper, we have addressed this issue where we would be showing the implementation of machine learning techniques in MT Evaluation. Further, we would also compare the results of this evaluation with human and automatic evaluation.

1 Introduction

Since the beginning of the time, translation has been the most important activity as the traders, who travelled distant places, needed to communicate with local population in their own language. Moreover while framing strategic alliances, the

Nisheeth Joshi · Iti Mathur
Department of Computer Science, Banasthali University
e-mail: {nisheeth.joshi,mathur_iti}@rediffmail.com

Hemant Darbari
Executive Director, C-DAC, Pune
e-mail: darbari@cdac.in

Ajai Kumar
Applied Artificial Intelligence Group, C-DAC, Pune
e-mail: ajai@cdac.in

El-Sayed M. El-Alfy et al. (eds.), *Advances in Intelligent Informatics*,
Advances in Intelligent Systems and Computing 320, DOI: 10.1007/978-3-319-11218-3_20

rulers of different countries needed translators, so that they may communicate with each other. In modern times this task has become even more important as the world have shrunk down to a global village. Now, not only governments, but also multinational corporations require translations of text from one language to another. Employing human translators for the job is very expensive and time taking. Moreover, skilled human translators are also very hard to find. Since the dawn of computers, computer aided translation or machine translation has been seen as the alternate to human translation. With initial MT engines being just dictionary matchers, current state of the art MT engines have come a long way. Today MT Engines can produce good and comparable results as that of human translators.

For rapid development of an MT Engine, it is required that we get quick and precise evaluation of the outputs of an engine, so that the development process could run smoothly. A manager of an MT system needs this information, so that he man according plan the future course of action. Getting an evaluation result of an MT engine can be achieved by two different approaches: Human Evaluation or Automatic Evaluation. Human evaluation is done by human expert who has an understanding of both source as well as target languages. In this evaluation the human judge is provided with a subjective questionnaire. Based on this questionnaire the judge is required to rate the outputs of an MT engine. Figure 1 shows the working of this approach. Automatic evaluation as the name suggests uses techniques which are independent of human evaluations. In this technique we employ a computer algorithm (popularly known as an evaluation metric) to ascertain the quality of MT output. In this technique, MT output is provided to an evaluation metric which compares the output with a reference translation which has been provided by a human translator. The quality of MT output is assessed by the checking its closeness to the reference translation. Figure 2 shows the working of this approach. This approach although termed as an automatic approach, is not entirely automatic as still it requires human intervention. Until we do not have a

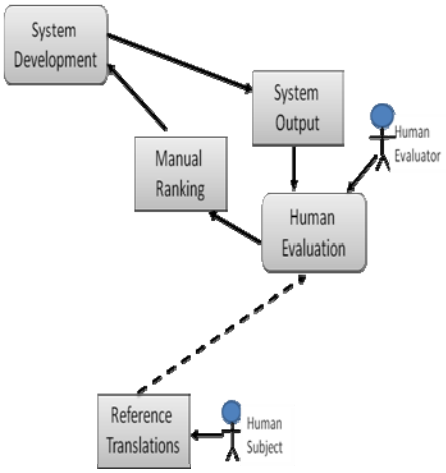

Fig. 1 Human Evaluation Process

translation for a sentence, it cannot be evaluated by an automatic evaluation metric. At times this tends to be a major bottleneck. Thus a need is being felt to have a completely automatic evaluation process. Several researchers are looking into various different techniques where they are trying to develop measures which are completely automatic and could provide evaluation results without any human intervention.

The rest of paper is organized as follows: Section 2 gives a brief review of the evaluation measures that have been used by different researchers. Section 3 shows our experimental setup for a completely automatic MT evaluation. Section 4 shows the evaluation results and its comparison with human and automatic evaluation metrics. Section 5 concludes the paper.

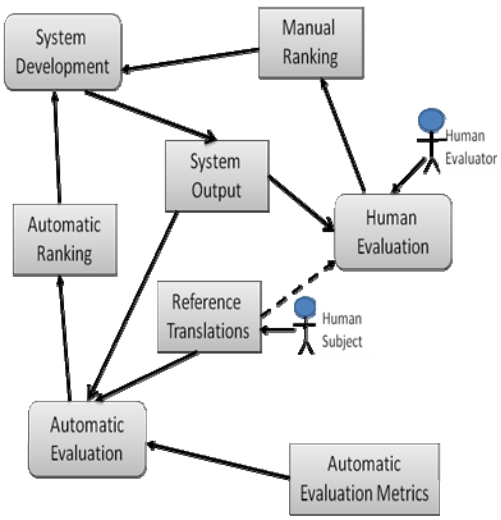

Fig. 2 Automatic Evaluation Process

2 Related Work

During the first formidable years of MT development, evaluation process was completely manual also known as human evaluation process. Miller & Beebe-Center[1] and Pfafflin[2] were the first researchers who proposed methodologies for MT evaluation. Their approach was simple, to provide MT outputs to humans and ask them questions based on the syntax and semantics of the target language. This approach was further modified by Slype[3], who evaluated SYSTRAN MT system. Instead of looking for correctness of the translation, he adjudged SYSTRAN for acceptability. He provided multiple outputs of the system to the evaluators and asked them if translation A is better than translation B. The prime objective of this evaluation was to distinguish between correct translations from incorrect ones. This evaluation, not only gave a measure to check the output of the system, but also found the cost of post editing the incorrect translations. This

evaluation changed the view of the people towards MT evaluation. Now, people started looking at the cost of post-editing translations instead of looking for correct translations. In 1980, a more detailed evaluation was carried out for English-French TAUM-Aviation MT system[4]. Here raw MT outputs and post-edited MT outputs were compared with human translations and the results were analyzed in terms of quality and cost of producing such translations.

By the dawn of the current century, automatic evaluation metrics started to emerge, providing an alternate to human evaluation. They gained popularity because they were fast and could provide repeatable results in very less time and in cost effective manner. The very first automatic evaluation metric which caught an eye of research community was BLEU (Bilingual Alignment for Evaluation Undestudy)[5]. The Basic goal of this metric was to calculate similarity between MT output and one or more human reference translations based on n-gram precision. A special measure called brevity penalty (BP) was introduced in this metric which penalized the MT outputs which were too shorter than their human counterparts. Looking at the success of BLEU, a plethora of automatic evaluation metrics followed. Some of the most successful metrics are discussed below.

National Institute of Standards Technology (NIST) came up with their own version of MT Evaluation metric[6]. This metric was named after them and incorporated slight modifications in BLEU. The main difference between BLEU and NIST was the method of their averaging n-gram scores. While BLEU relied on geometric mean, NIST performed an arithmetic mean. Another unique feature of NIST was its calculation of weights based on reliance upon n-grams which occurred less frequently, as this was an indicator of their higher informativeness.

Turian et. al.[7] proposed another metric which also claimed to perform better than BLEU. Here, Turian attempted to model movement of phrases during translation by using the maximum matching size to compute the quality of a translation. It could find the longest sequence of words that matched between human reference translation and MT output. Here precision and recall were computed as the size of the matches divided by the length of lexicons in MT output or human reference translation. Then, the harmonic mean of these two measures was computed for calculation of final GTM score. Snover et.al.[8] proposed another metric which was based on edit distance algorithm proposed by Levenshtein[9]. This metric accounted for no. to shifts along with no. of insertions, deletions and substitutions to compute the results. This metric tried to measure the amount of post editing that a human would have to perform on the machine output so that it exactly matches the reference translations.

Meteor[10] was a major breakthrough in MT evaluation research as this metric not only measured the performance of MT output on lexical level, but also looked at deeper linguistic levels (shallow syntactic and semantic). Meteor was a tunable alignment oriented metric while BLEU was simply a precision oriented metric. This metric used several stages of word matching between the system output and the reference translations.

The major drawback with all these metrics was that they all required one or more human reference translations. The very first completely automatic evaluation was performed by Gammon et.al.[11] where they applied their technique in

ascertaining MT quality and fluency. They employed SVM based classifier and used several features from syntactic and semantic levels. Specia et. al.[12] used machine learning techniques in identification of features which can be used for MT evaluation and post editing. They applied a SVM based classifier which was trained in 74 features which were from lexical, shallow syntactic and semantic levels. They concluded that given a trained model, a classifier can work well for a particular language pair and can produce better correlations with human judgments.

3 Experimental Setup

For development of our evaluation system, we used a 3,300 sentence corpus that was built during ACL 2005 workshop on Building and Using Parallel Text: Data Driven Machine Translation and Beyond[13], as the training corpus. The statistics of this corpus is shown in Table 1.

Table 1 Statistics of training corpus used

Corpus	English-Hindi Parallel Corpus	
Sentences	3,300	
	English	Hindi
Words	55,014	67,101
Unique Words	8,956	10,502

We also focused on using supervised machine learning in evaluation of MT engine outputs without using human reference translations. For this we used Decision Tree (DT) based classifier which used J48 algorithm which is a java version of C4.5 decision tree algorithm[14]. We used WEKA toolkit[15] for training this classifier. We also trained a Support Vector Machines (SVM) based classifier which was developed using LIBSVM package developed by Chang and Lin[16]. We used 27 features for training our classifiers. These features were as follows:

1. Length of the source sentence.
2. Length of the target sentence.
3. Average source token length.
4. LM probability of source sentence.
5. LM probability of target sentence.
6. Average no. of occurrences of a target word within a target sentence.
7. Average number of translations per source word in the sentence (as given by IBM 1 table threshold so that prob(t|s) > 0.2).
8. Average number of translations per source word in the sentence (as given by IBM 1 table threshold so that prob(t|s) > 0.01) weighted by the inverse frequency of each word in the source corpus.
9. Percentage of unigrams in quartile 1 of frequency (lower frequency words) in a corpus of the source language (SMT training corpus).

10.Percentage of unigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source language.
11.Percentage of bigrams in quartile 1 of frequency of source words in a corpus of the source language.
12.Percentage of bigrams in quartile 4 of frequency of source words in a corpus of the source language.
13.Percentage of trigrams in quartile 1 of frequency of source words in a corpus of the source language.
14.Percentage of trigrams in quartile 4 of frequency of source words in a corpus of the source language.
15.Percentage of unigrams in the source sentence seen in a corpus.
16.Count of punctuation marks in source sentence Count of punctuation marks in target sentence.
17.Count of punctuation marks in target sentence.
18.Count of mismatch between source and target punctuation marks.
19.Count of content words in the source sentence.
20.Count of context words in the target sentence.
21.Percentage of context words in the source sentence.
22.Percentage of context words in the target sentence.
23.Count of non-content words in the source sentence.
24.Count of non-content words in the target sentence.
25.Percentage of non-content words in the source sentence.
26.Percentage of non-content words in the target sentence.
27.LM probabilities of POS of target sentence.

To test the classifiers, we were also required to have MT Engines. Thus we trained three MT toolkits on tourism domain. These engines were

1. Moses Phrase Based Model[17] (PBM) where phrases of one language are statistically aligned and translated into another language. Moses PBM uses conditional probability with linguistic information to perform the translation.
2. Moses Syntax Based Model[18] (SBM) which implements a Tree to String Model. In this system an English sentence is parsed and its parsed output is matched with the target string and thus transfer grammar is generated which has a parsed output at one end and the string at the other.
3. An Example Based Machine Translation[19] (EBMT) model where examples in the training data which are of higher quality or are more relevant than others are produced as translations.

We registered the outputs of the training corpus against all these three MT engines and asked a human evaluator to judge the outputs. The judging criteria was same as used by Joshi et. al.[20]. All the sentences were judged on ten parameters using a scale between 0-4. Detailed discussion on use of these parameters have been discussed by Joshi et al. [21]. Table 2 shows the interpretation of these scales. The ten parameters used in evaluation were as follows:

1. Translation of Gender and Number of the Noun(s).
2. Identification of the Proper Noun(s).
3. Use of Adjectives and Adverbs corresponding to the Nouns and Verbs.
4. Selection of proper words/synonyms (Lexical Choice).
5. Sequence of phrases and clauses in the translation.
6. Use of Punctuation Marks in the translation.
7. Translation of tense in the sentence.
8. Translation of Voice in the sentence.
9. Maintaining the semantics of the source sentence in the translation.
10. Fluency of translated text and translator's proficiency.

Table 2 Interpretation of HEval on Scale 5

Score	Description
4	Ideal
3	Perfect
2	Acceptable
1	Partially Acceptable
0	Not Acceptable

Once the human evaluation of these outputs was done, we used these results along with the 27 features that were extracted from the English source sentences and Hindi MT outputs. We tested the classifiers using another corpus of 1300 sentences. Table 3 shows the statistics of this corpus.

Table 3 Statistics for Test Corpus

Corpus	English Corpus
Sentences	1,300
Words	26,724
Unique Words	3,515

These 1300 sentences were divided into 13 documents of 100 sentences each. We registered the output of the test corpus on all three MT engines and performed human evaluation on them. Further we also used some of the popular automatic evaluation metrics on the output produced and then we evaluated the results of the systems using the two classifiers. We used BLEU and Meteor metrics for the evaluation. For incorporating automatic evaluation we were also required to have reference translations so we used single human references to be used with automatic evaluation metrics. Since Meteor works on shallow syntactic and semantic levels, we were required to develop tools for this metric. For syntactic matching, we developed a Hindi stemmer based on the light weight stemming algorithm proposed by Rangnathan and Rao [22] and for semantic matching we used Hindi WordNet[23]. For generation of paraphrases we used Moses PBM's phrase table. Moreover, we also compared the results of these evaluations.

4 Results

We correlated the results of the output produced by the MT engines. We used spearman's rank correlation as it produces the unbiased results. Table 4 shows the results of correlation at document level between Human and BLEU & Meteor 1.3 which matches for exact, stem, synonym and paraphrase matches and both the classifiers (DT and SVM). In all the cases EBMT had better correlation with human judgments while Moses PBM showed the poorer results. Correlation of human evaluation with automatic metrics was very low as compared to the results of correlation with classifiers.

Table 4 Document level correlation between human and different automatic evaluation measures

	BLEU	Meteor	DT	SVM
Moses PBM	0.011	0.297	0.089	0.750
Moses SBM	0.181	0.313	0.139	0.773
EBMT	0.490	0.352	0.161	0.781

Table 5 Sentence level correlation between human and different automatic evaluation measures

	BLEU	Meteor	DT	SVM
Moses PBM	0.063	0.045	0.579	0.682
Moses SBM	0.077	0.048	0.635	0.707
EBMT	0.106	0.040	0.590	0.600

Table 6 System level average scores of the all the evaluations

	Moses PBM	Moses SBM	EBMT
Human	0.393300	0.356000	0.464000
BLEU	0.017666	0.012901	0.026553
Meteor	0.069900	0.061205	0.102117
DT	0.538077	0.505577	0.589615
SVM	0.526154	0.492500	0.589038

At sentence level this phenomenon was repeated. In all the cases except for meteor EBMT showed best results and Moses PBM showed poor results. For Meteor, Moses SBM's results were better than EBMT's results. Table 5 shows the results of this study. Here again the correlations of automatic metrics were very low as compared to classifiers. Table 6 shows the average system scores of all the evaluations incorporated. In all the cases EBMT had the best scores and Moses SBM had the poorer scores.

5 Conclusion

In this paper we have discussed the use of machine learning techniques in MT evaluation. We have also compared the results of our study with human and automatic evaluation metrics. By looking at the results we can confidently say that machine learning techniques can prove to be an alternate to human evaluation as it produces consistent results with human evaluations which at times, is not possible with automatic evaluations. Moreover, using machine learning techniques we can have minimum human intervention. In machine learning based MT evaluation, humans are involved only during the training of classifiers while in automatic evaluation human are required every time the evaluation corpus is changed as we would require human reference translations for the new corpus. Thus by looking at these points we can say that machine learning based MT evaluation can be a very good alternative for human and automatic evaluation metrics.

References

[1] Miller, G.A., Beebe-Center, J.G.: Some Psychological Methods for Evaluating the Quality of Translation. Mechanical Translations 3 (1956)
[2] Pfafflin, S.M.: Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments. Mechanical Translation and Computational Linguistics 8, 2–8 (1956)
[3] Slype, G.V.: Systran: evalaution of the 1978 version of systran English-French automatic system of the Commission of the European Communities. The Incorporated Linguist 18, 86–89 (1979)
[4] Falkedal, K.: Evaluation Methods for Machine Translation Systems. An Historical overview and Critical Account. Technical Report, ISSCO, Universite de Geneve (1991)
[5] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. RC22176 Technical Report, IBM T.J. Watson Research Center (2001)
[6] Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proceedings of the 2nd International Conference on Human Language Technology, pp. 138–145 (2002)
[7] Turian, J.P., Shen, L., Melamed, I.D.: Evaluation of Machine Translation and its Evaluation. In: Proceedings of MT SUMMIT IX (2003)
[8] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA), pp. 223–231 (2006)
[9] Levenshtein, V.I.: Binary Codes Capable of Correlating Deletions, Insertions and Reversals. Soviet Physics Doklady 8(10) (1966)
[10] Denkowski, D., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Workshop on Statistical machine Translation (2011)

[11] Gamon, M., Aue, A., Smets, M.: Sentence-level MT evaluation without reference translations: Beyond language modeling. In: Proceeding of Annual Conference of European Association of Machine Translation (2005)
[12] Specia, L., Raj, D., Turchi, M.: Machine Translaiton Evaluation and Quality Estimation. Machine Translation 24(1), 24–39 (2010)
[13] Koehn, P., Martin, J., Mihalcea, R., Monz, C., Pedersen, T.: Proceedings of the Workshop on Building and Using Parallel Texts (2005)
[14] Quinlan, J.R.: Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research 4, 77–90 (1996)
[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
[16] Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 27, 1–27 (2011)
[17] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of ACL: Demonstration Session (2007)
[18] Hoang, H., Koehn, P.: Improved Translation with Source Syntax Labels. In: Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, July 15-16, pp. 409–417 (2010)
[19] Joshi, N., Mathur, I., Mathur, S.: Translation Memory for Indian Languages: An Aid for Human Translators. In: Proceedings of 2nd International Conference and Workshop in Emerging Trends in Technology (2011)
[20] Joshi, N., Darbari, H., Mathur, I.: Human and Automatic Evaluation of English to Hindi Machine Translation Systems. In: Wyld, D.C., Zizka, J., Nagamalai, D. (eds.) Advances in Computer Science, Engineering & Applications. Advances in Soft Computing, vol. 166, pp. 423–432. Springer, Heidelberg (2012)
[21] Joshi, N., Mathur, I., Darbari, H., Kumar, A.: HEval: Yet Another Human Evaluation Metric. International Journal of Natural Language Computing 2(5), 21–36 (2013)
[22] Ramnathan, A., Rao, D.: A Lightweight Stemmer for Hindi, In Proceedings of Workshop on Computational Linguistics for South Asian Languages. In: 10th Conference of the European Chapter of Association of Computational Linguistics, pp. 42–48 (2003)
[23] Narayan, D., Chakrabarti, D., Pande, P., Bhattacharyya, P.: An Experience in Building the Indo WordNet - a WordNet for Hindi. In: Proceedings First International Conference on Global WordNet, Mysore, India (January 2002)