International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 10, October - 2013

Key Issues in Machine Translation Evaluation of English-Indian Languages

Pooja Malik [1], Abhilasha Gupta [2], Anurag Baghel [3]
Noida International University[1, 2], Gautam Buddha University[3]

Abstract

Basically, the word Machine Translation (MT) refers to the use of a machine or computer for performing translation task that converts text or speech from one Natural Language (NL) into another Natural Language (NL). MT Evaluation is not an easy task because there may exist many perfect translations of a given source sentence. The evaluation of the sentences or a corpus of sentences translated with the help of machines or computers by comparing it with the sentences or a corpus of sentences translated by humans is known as Machine Translation Evaluation. The basic necessity of evaluation is for the comparison of the performance of different MT Engines or to improve the performance of a specific MT Engine. In past years, the MT Evaluation task is used to be performed by human beings. Human evaluation of Machine Translations is extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. Now-a-days, automatic evaluation methods are becoming popular. A large number of metrics have been developed for the automatic evaluation of Machine Translation systems. Most of them are based on n-gram metric evaluation. In this paper, authors are discussing some of the metrics developed by various researchers that are presently used for the automatic evaluation of Machine Translation and various issues in the automatic evaluation of English-Indian languages MT because all these techniques can not be applied as it is in evaluating English-Indian language MT systems due to the structural differences in the languages involved in the MT language pair.

1. Introduction

Machine Translation (MT) is the sub-field of computational linguistics concerned with the translation of text or speech from one Natural Language (NL) to another with the help of machines. It is the process in which a text from one Natural Language (such as English) is translated into another (such as Hindi). MT means automatic translation of text by computer from one Natural Language into another Natural Language [3].

Basically, Machine Translation is a two-step process. The first step involves the decoding of the source text and the second step involves the re-encoding of this meaning in the target language. The first step shows that the translator must interpret and analyze all the features of the text, a process that requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc. of the source language, as well as, the culture of its speakers.

Work on Machine Translation started in the 1950s after the second world war. The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The experiment was a great success and ushered in an era of machine translation research. Today there are many software available for translating natural languages between themselves [1, 2].

To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restored in the target language, i.e. the translation. While on the surface this seems straightforward, it is far more complex. Translation is not a mere word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know how each word may influence another.

India has a linguistically rich area-It has 18 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. English is very widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. In such a situation, there is a big market for translation between English and the various Indian languages [4].
Machine Translation systems are a powerful tool and are very important as they offer low-quality translations in situations where low quality translation is better than no translation at all, or where a rough translation of a large document delivered in seconds or

minutes is more useful than a good translation delivered in three weeks' time.

.

2. Machine Translation Evaluation

Basically, the word 'Evaluation' means to assess or to check the correctness of anything. Every time when a new technology is developed, it has to be tested or evaluated on some grounds. In the same way round, the need for the evaluation of the Machine Translation arises. Evaluation of a MT System is as important as the MT itself, answering the questions about the accuracy, fluency and acceptability of the translation and thus artifying the underlying MT algorithm.

Evaluation of Machine Translation (MT) has historically proven to be a very difficult exercise. The difficulty stems primarily from the fact that translation is more an art than a science; most sentences can be translated in many acceptable ways. Consequently, there is no gold standard against which a translation can be evaluated.

Traditionally, MT evaluation has been performed by human judges. This process, however, is time-consuming and highly subjective. The investment in MT research and development being what it is, the need for quick, objective, and reusable methods of evaluation can hardly be over-emphasized. To this end, several methods for automatic evaluation have been proposed in recent years, some of which have been accepted readily by the MT community. Especially popular is BLEU (Bi Lingual Evaluation Understudy), a metric that is now being used in MT evaluation forums to compare various MT systems and also to demonstrate improvements in translation quality due to specific changes made to systems. BLEU is an n-gram co-occurrence based measure-by this we mean that the intrinsic quality of MT output is judged by comparing its n-grams with reference translations by humans.

On a broader term, the Machine Translation Evaluation can be performed at two levels:

2.1 At Sentence Level

When the evaluation of the machine translated text is done sentence by sentence means each and every sentence is evaluated separately, it is known as the sentence level evaluation.

2.2 At Corpus Level

In general, corpus can be defined as a collection of sentences. So at corpus level evaluation, the evaluation of a large machine translated document file is done at once.

A large number of methods have been developed for the evaluation of the machine translation. Most of the methods focus on the evaluation of the output of machine translation, rather than on performance. One of the typical ways for lay people to assess the quality of the output of a machine translation engine is through translating from a source language into a target language, and then back to the source language using the same engine. This type of evaluation technique is known as Round-Trip translation or (back translation). But this methodology is deficient for any serious study of the quality of MT output.

Traditionally, MT evaluation has been performed by human judges. This process, however, is time-consuming and highly subjective. However, in recent times, automatic evaluation methods have become popular.

In the following paragraph, we will discuss the two major methods used for the evaluation of the Machine Translation systems. These are the Human Evaluation and the Automatic Evaluation.

2.3 Human Evaluation

When the evaluation of the translated sentences is performed by the human-beings, it is known as the Human Evaluation. For this purpose, we need human translator who should be a native speaker of the language-pairs involved in the translation process. But we cannot use the human evaluation every time due to a large number of reasons.

Human evaluations of Machine Translation (MT) weigh many aspects of translation, including *adequacy*, *fidelity*, and *fluency* of the translation. For the most part, these various human evaluation approaches are quite expensive. Moreover, they can take weeks or months to finish. This is a big problem because developers of machine translation systems need to monitor the effect of daily changes to their systems in order to weed out bad ideas from good ideas.

Because humans are the golden standard for using language, obviously human evaluation is the holy grail for evaluation for machine translations but still it has some drawbacks. The major issues with human evaluation are : It is very expensive, very time consuming and therefore not always an option, when using human evaluation one should take care for maintaining objectivity, for a more statistically significant result and elimination of subjective

evaluation, human evaluation of each MT output needs to be done by more than one evaluator.

Due to the above mentioned factors, the need for machine or automatic evaluation arose.

2.4 Automatic Evaluation

When the evaluation of the translated text is done with the help of machines, it is known as Machine Evaluation.

In automatic or machine evaluation, evaluation is done at two levels – at sentence level and at corpus level. At sentence level, the scores are calculated by the metric or the algorithm for a set of translated sentences, and then correlated against human judgment for the same sentences while at the corpus level, the scores over the sentences are aggregated for both human judgments and metric judgments, and then aggregate scores are then correlated [8].

The big advantage of using Machine Evaluation is that the scoring is objective, while human evaluation /scoring will often differ not only from time to time but much bigger from human to human.

3. Metrics For The Automatic Evaluation

A metric for the evaluation of machine translation output is nothing but simply a measurement of the quality of the output of the Machine Translation engine. The quality of a translation is inherently subjective, there is no objective or quantifiable "good". Therefore, the task for any metric is to assign scores for quality in such a way that they correlate with human judgment of quality. That is, a metric should give high scores to those translations which humans give high scores to, and give low scores to those which humans give low scores to.
Now the problem is that how does one measure translation performance? "The closer a machine translation is to a professional human translation, the better it is."[7] This is the central idea behind the evaluation of the machine translation. To judge the quality of a machine translation output, one measures its closeness to one or more reference human translations according to a numerical metric.

There are a large number of algorithms for the evaluation of machine translation systems. All of them are based on different concepts and have a different way of evaluation.

A brief overview of some of them is as below.

3.1 BLEU Metric

The BLEU metric is an IBM-developed metric and is probably the best known Machine Evaluation Metric. The central idea is that the closer a machine translation is to a professional human translation, the better it is [9]. To check how close a candidate translation or a machine translation is to a reference translation or a human translation, a n-gram comparison is done between both translations.

Typically, there are many "perfect" translations of a given source sentence. These translations may vary in word choice or in word order even when they use the same words. And yet humans can clearly distinguish a good translation from a bad one.

The BLEU metric works on n-gram concept. BLEU compares candidate (machine translation) with reference(s) (human translation(s)). This comparison is done by performing '*n-gram*' checking, where n varies from 1 to length of candidate. In this, 'n' consecutive words of a sentence pair are compared. 1-gram is called unigram, 2-gram → bigram, 3-gram → trigram, etc.

The BLEU score is based on the geometric mean of n-gram precision. The score is given by:

$$\text{BLEU} = \text{BP} * \exp\left[\sum_{n=1}^{N}\left\{\left(\frac{1}{N}\right) * \log(\text{Pn})\right\}\right]$$

Where N is the maximum n-gram size and n=1 to N.

3.2 NIST

NIST is an NIST (National Institute of Standards and Technology) developed metric. It is based on the same ideas as the BLEU metric of IBM, and it can be seen as an upgrade to this metric [10].

It is also an n-gram counting metric, but the idea is to fix two problems with the BLEU metric: Firstly, BLEU usage a geometric means of n-grams. But it employs the arithmetic average of n-gram counts rather than a geometric average.

Secondly, BLEU treats all n-grams equally. That means that n-grams which occur often and have little information have as much impact on overall precision as information rich n-grams. But in NIST, the n-grams are weighted according to their information contribution. The score represents the average

information per word, given by the n-grams in the translation that match an n-gram of a reference in the reference set.

3.3 F-Measure

It is a metric developed in the New York University. It uses 'maximum matching' from Graph Theory, subset of co-occurences in the candidate and reference text are counted in such a way that a token is never counted twice. On the matching value a Recall and Precision is defined where Recall is the amount of counted tokens which also appear in the candidate text and Precision the amount which also appear in the reference text.

Recall(Candidate|Reference) =
MMS(Candidate, Reference)/ |Reference|

Precision(Candidate|Reference)=
MMS(Candidate, Reference)/|Candidate|
A reward for longer matches is introduced as the square root of the squares of the different lengths. This reward is bigger when larger matches are found, this takes care for the "fluent" measure of the translations [8].

The final F-measure is calculated as the harmonic mean of both the Precision and Recall which is defined as (2*Precision*Recall) / (Precision + Recall).

3.4 METEOR

Meteor is a Metric for Evaluation of Translation with Explicit Ordering and was developed at Carnegie Mellon University. It uses one-gram overlaps and uses Wordnet to calculate the use of synonym from the reference text. It has a separate module to address ordering which explains why higher n-grams are not used. A penalty for reordering is calculated on how many chunks in the produced text need to be moved around to get the reference text.

Another feature in Meteor is that stemmed words can be used so that when a produced translation choses a slightly different grammatical structure, the metric still spots the same words are used. Most older metrics, like Bleu, expect that these problems, synonyms/stemming, is resolved by using enough reference texts, in practice however it is very hard to get hold on enough reference texts, so that a lot of evaluation is done on 1 or 2 reference texts [11].

4. Some Issues in The Automatic Evaluation of English-Indian Languages Machine Translation

Though BLEU is probably the best-known evaluation metric among all the metrics developed till date. But according to the research study of IIIT Mumbai and CDAC Mumbai, BLEU has some major drawbacks while evaluating the Machine Translation of English-Indian languages.

The authors are discussing only some of the major criticisms of BLEU in automatic evaluation of English-Indian languages machine translation by taking some specific examples.

The notations used in the examples are as :

ES : English Language Sentence(source)

C : Candidate Sentence(translated sentence)

R1, R2, R3 : Reference Sentences

BLEU : BLEU Score

HES : Human Evaluation Score,

The major issues are discussed as under :

4.1 Meaningless Score

The first criticism of BLEU is that the score that it provides is not meaningful in itself. The scores for words are equally weighted so missing out on content bearing material brings no additional penalty.

Example :

ES: It was rainning when we left for Goa. C : जब हम गोआ के लिए निकले यह बारिश हो रही थी। R1 : जब हम गोआ के लिए निकले बारिश हो रही थी।
BLEU Score= 0.4647, HES = 0.6430

In the above example, even on missing out a simple word "यह" will give a low BLEU Score.

4.2 Considers Synonyms as different words.

Languages allow a great deal of variety in choice of vocabulary. BLEU treats synonyms as different words. Word choice is captured only to a limited extent even if multiple references are used.

Example :

ES: Daman and Dui offers you refreshing holiday. C : दमन एवं द्वीप आपको ताज़गीभरी छुट्टियाँ देता है । R1 : दमन एवं द्वीप आपके अवकाश ताज़गीभरे बना देता है । R2 : दमन एवं द्वीप तुम्हारे अवकाश ताज़गीभरे बना देता है । R3 : दमन एवं द्वीप तुम्हारे अवकाश ताज़गीभरे बनाता है ।
BLEU Score= 0.3097, HES = 1.000

In the above example, the two words "छुट्टियाँ" and "अवकाश" have been considered different words, though the two words are the synonyms of each other, and hence will give a low BLEU Score.

4.3 Considers the same words written in two different forms as different.

We use Unicode to encode Hindi characters. There is a separate code for encoding each separate Hindi character. It is well-known that in Hindi language, the same word can be written in more than one form.

Example :

ES : That temple is very beautiful. C : वह मंदिर बहुत सुंदर हैं। R1 : वह मन्दिर बहुत सुन्दर हैं।
BLEU Score= 0.4099, HES = 1.0000

In the above example, the words "मंदिर" and "सुंदर" have been written in two different forms and hence have been considered as different words thereby reducing the final BLEU Score.

Example :

ES: Daman and Dui offers you refreshing holiday. C : ताज़गीभरी देता अवकाश द्वीप एवं दमन है आपको । R1 : दमन एवं द्वीप आपके अवकाश ताज़गीभरे बना देता है । R2 : दमन एवं द्वीप तुम्हारे अवकाश ताज़गीभरे बना देता है । R3 : दमन एवं द्वीप तुम्हारे अवकाश ताज़गीभरे बनाता है ।
BLEU Score = 1.0000, HES = 0.1020

In the above example, though the sentence is semantically incorrect but even though the BLEU provides a high score.

4.4 Better Score does not indicate better translation.

Another criticism is that the n-gram matching technique allows too much variation. There are typically thousands of variations on a hypothesis translation – a vast majority of them both semantically and syntactically incorrect – that receive the same BLEU score. That means even when a sentence is semantically and syntactically incorrect, it receives good BLEU score. Thus higher BLEU score is not necessarily indication of better translation.

4.5 Requires a number of reference sentences for the evaluation purpose

The BLEU metric requires a large number of reference human sentences for the evaluation of machine translated sentences. The availability of such references is not that much easy task.

4.6 Poor correlation with human judgments.

BLEU scores generally do not correlate with human judgments. According to Turin et al.(2003) report, with larger corpora the correlation between BLEU and human judgments is poor.

The main point that comes out of these criticisms is that BLEU needs to be used with caution; there is a need for greater understanding of which uses of BLEU are appropriate, and which are not.

5. Conclusion

In this paper, we have reviewed various automatic evaluation tools/metrics available for the evaluation of various MT engines or to improve the performance of a specific MT engine. We know that a metric that only works for text in a specific domain is useful, but less useful than one that works across many domains – because creating a new metric for every new evaluation or domain is undesirable.

The various automatic evaluation metrics are working well in their respective domains and BLEU, the best-known evaluation metric, has far been accepted by many of the researchers in the field of Machine Translation Evaluation. But still BLEU do not correlate with human judgment to the degree that it is currently believed to do for English-Hindi pair.

We have reviewed existing criticisms of BLEU and concluded that BLEU score is not sufficient to reflect a genuine improvement in translation quality.

Though it has been believed that we have achieved a lot in this field but still there is a long way to go when it comes to the automatic evaluation of English-Indian languages machine translation.

6. References

[1] James Allen: Natural Language Understanding,(Benjamin/Cummings Series in Computer

Science) Menlo Park: Benjamin/Cummings Publishing Company, 1987.
[2] S Nuremburg, J Carbon ell, M Tomita, K Goodman: Machine Translation: A Knowledge-Based Approach, Morgan Kaufmann Publishers Inc., San Francisco, CA,USA 1994.
[3] Durgesh Rao, Machine Translation in India: A Brief Survey, NCST(CDAC) Mumbai, available at http://www.elda.org/en/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf
[4] Sudip Naskar, Siviji Bandyopadhyay, Use of Machine Translation in India: Current Status, Jadavpur University, Kolkata, India.
[5] Arnold, D.,Balkan, L., Meijer, S., Humphreys, R. and Sadler, L., (1994) Machine Translation: an Introductory Guide, Black wells-NCC, London.
[6] Salil Badodekar, Translation Resources, Services and Tools for Indian Languages, Computer Science and Engineering Department, IIT Mumbai, available at http://www.cfilt.iitbac.in/Translation-survey/survey.pdf.
[7] Joseph P. Turian, Luke Shen, and I. Dan Mela, Evaluation of Machine Translation and its evaluation Proceedings of MT Summit IX; New Orleans, USA, 23-28 September 2003
[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation". In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL):311-318. Philadelphia.
[9] G. Doddington (2002) "Automatic Evaluation of Machine Translation Quality Using-gram Co-Occurrence Statistics". In Human Language Technology :Notebook Proceedings:128-132. San Diego.
[10] Banerjee Satanjeev, Lavie Alon, "METEOR: An Automatic Correlation with Human Judgments", http://www.cs.cmu.edu/~alavie/papers/BanerjeeLavie2005-final.pdf.
[11] Chris Callison-Burch, Miles Osborne, Phillipp Koehn(2006)"Re-evaluating the Role of BLEU in Machine Translation Research". School on Informatics. University of Edinburgh. Edinburgh. EH8 9LW.
[12] Ananthakrishnan R, Pushpak Bhattacharyya, M. Sasikumar, Ritesh Shah, Some Issues in Automatic Evaluation of English-Hindi MT: more blues for BLEU in proceeding of 5th International Conference on Natural Language Processing(ICON-07), Hyderabad, India.