CLASSIFICATION OF MT-OUTPUT USING HYBRID MT-EVALUATION METRICS FOR POST-EDITING AND THEIR COMPARATIVE STUDY

Kuldeep Yogi and Chandra Kumar Jha

Banasthali University, Rajasthan, India

ABSTRACT

Machine translation industry is working well but they have been facing problem in postediting. MT-outputs do not correct and fluent so minor or major changes need for publishing them. Postediting performs manually by linguists, which is expensive and time consuming. So we should select good translation for postediting among all translations. Various MT-evaluation metrics can be used for filter the good translations for postediting. We have shown the use of various MT-evolution metrics for selection of good translation and their comparative study.

INDEX TERMS

Machine translation, MT-Evaluation, Postediting, MT-Engine Good Translation, BLEU, METEOR, FMEASURE

1. INTRODUCTION

Language has their different syntax and rules. A human can take an appropriate meaning of a double meaning word based on flow of sentence or meaning of previous sentence. Some authors write in continuation so a machine can't split the sentence accordingly and confuses in many situations to take right decision.

MT-Engines suffer from many problems like ambiguity, Gender identification, subject reorganization etc. Although, MT-Engines give 10-70% correct translation but it needs 100% for publishing. So it is require that translation should check manually by linguists systematically. Linguist will make some changes for remove linguistic errors(Knight Kevin, 1994). Translations shall read one by one by linguist for errors and will makes chances according to the need (Michel Simard,2007).

If we post-edit all translations then it will not be convenient and expensive also. Translations having minor/little modification suitable for post-editing but how do we decide that a translation has minor changes? To do so, we used various automatic-machine-translation-evaluation metrics and their combination. We can filter-out good translations for post-editing using these evaluation metrics.

DOI:10.5121/acii.2015.2403

2. RELATED WORK

2.1 MT Human Evaluation

MT evaluation task was originated in 1956 by ALPAC. 144 Russian book's sentences were taken for experiment. Some sentences were translated in English by human translators and some from machine. The evaluation result of these sentences found that MT outputs are not fluent and accurate as human translation.

Another effort had taken by Slype in 1978 for MT evaluation. He evaluated SYSTRAN system. The objective of evaluation was not set to correctness of translation where as the aim was to find-out cost of corrections. This task has changed the views of peoples about MT evaluation. They were started to think that MT output can be make useful with some minor changes (post-edit). One more experiment performed by Falkedal, 1991 for English-French MT system. MT output and post-edited MT-output compared with human translators and result shown that MT post-edited output were better than MT-outputs.

Church and Hovy tried a different trick in 1993 for MT-evaluation. They gave some MT-outputs to human evaluator and asked him some multiple choice questions related to MT-output' information. If the human evaluator is able to answering all questions it means the translation is correct otherwise it is less correct.

If a translation is expressing the right meaning of source sentence then it is good translation. This idea followed by Eck and Hori, 2005. They ranked a translation between 0-4 scales. If translation is expressing 100% right meaning then it should be rank 4. Following ranking criteria had taken for evaluation:-

Scale	Particulars
0	Totally different meaning
1	Partially the same meaning but misleading information is introduced
2	Partially the same meaning and no new information
3	Almost the same meaning
4	Exactly the same meaning

2.2 AUTOMATIC MT EVALUATION

Language translation as well as evaluation task performed by human mostly accurate but it has slow and expensive so it requires that we should use machine for MT-evaluation but how can we do this? The idea came in our mind from the task done by Falkedal, 1991. He compared MT-output with human translation. It means we can evaluate MT-output by comparing machine translation with human translation.

The comparison can be lexicon, syntactic or semantic. Tillmann developed a metric in 1997 named WER(Word Error Rate). It was a lexicon based metric in which word mismatches tried to find between MT-output and human references. A Translation can be too long so find the match is better then find mismatches.

The extension or we can say improvement of WER innovated in 2006 by Snover named TER(Translation Edit Rate). We can compute the score of TER by finding the edits need to make MT-output as human translation. Edit includes number of shifting, deletion, substation and insertion of words in a translation.

BLEU MT-evaluation metric developed in 2002 by Papineni. It is much closer to human evaluation. It is also a lexicon based evaluation metric. Evaluation performed by calculating N-Grams precision between MT-output and one or more human references. Short machine translation can get high score by only calculating N-grams precision score so a new parameter included in this metric that is BP(Brevity Penalty). A very short machine translation behind human translation will be penalized. It was mile-stone in automatic MT-evaluation area so that it has been very popular and used.

After a great success of BLEU metric Banerjee and Lavie gave improved idea for MT-evaluation in 2005 named METEOR. They have shown that word matching can be performed on 3 levels and the score of machining calculates as precision and recall. In the first level, we look exact word match between MT and human reference. If some words left in first level due to mismatch so we first extract the root word by removing the stemmed characters for all remaining words and again look for match. If some words left without match in 2 step then look their synonyms in the database and match again remaining words. METEOR gives weight to word position. Number of crosses calculates in matching process for all human references.

Denkowski and Lavie modified METEOR metric in 2010 and 2011. They included one more step in word matching. They found that sentences consisted in some paraphrases so we can collect some similar group of words(Phrase) in a database and replace these phrased in word matching.

2.3 POST-EDITING

Post-editing mostly used when MT-output is not enough accurate so it is better to retranslate it some academician and industrialist suggested that it should be post-edited in place of retranslate because it save 40% time (Masselot 2010).

The task of post-Editing stared from a conference which held in London, 1982. V. Lawson gave the idea for contribution of translators to improvement of an MT system. A meeting of SYSTRAN users held in California, April 1981. The aim of this meeting was to co-ordinate the feedback from users.

3. EXPERIMENTAL SETUP

In brief, the general idea of the work presented in this paper is filter the sentences in two different categories. First category related to those sentences, those can be post-edited and second, includes those sentences, those are fully meaningless and need to retranslation. For do the same, we did according to figure1:-

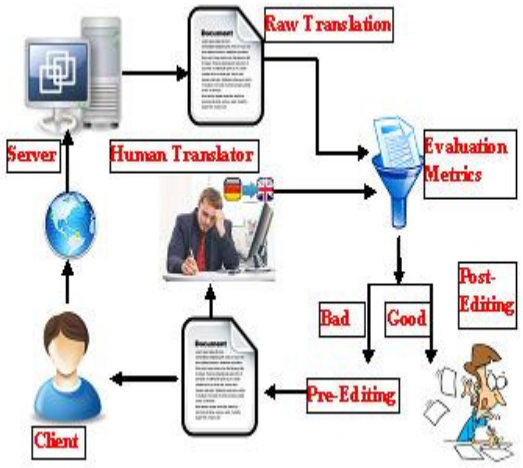

Fig. 1.Overall System flow diagram

3.1 CORPUS

Translation task is not a simple task. It varies language to language. Western languages and Indian languages have their different sentence-structure. Sentence's form in western languages is Subject(S) + Verb (V) + Object (O) where as in India languages it is SOV. We use same word for translation in different situations so we have set our particular domain for translation. We have selected Hindi language and tourism domain for the same. We collected 16200 different structures English language-sentences from various web-sites, magazines, news-papers related to tourism and stored them in 162 files. Every file has 100 sentences.

1.2 MT-ENGINES

We have selected 1300 English sentences for our experiment. We have used four MT-engines for comparative study. These MT-engines are freely available online. All 1300 sentences translated by following MT-engines:-

1. Babylon MT-Engine
2. Microsoft MT-Engine
3. Google MT-Engine
4. EBMT MT-Engine

3.3 REFERENCES

All automatic MT evaluation metrics match MT-output against correct translation for measure the accuracy and completeness of sentence. These translations are built-up by language experts or linguists. Linguists use different words for same meaning and MT-engines also do the same. So we use more than one human translation for matching. These translations are called here references. We have translated our corpus files from 3 different human translators. One is assistant professor (English Literature) and two translators are M.Tech(CS) students.

3.4 HYBRID MT-EVALUATION METRIC

We have been using various automatic evaluation metrics for score the MT since few year. The most used and reliable metric is said BLEU which works on ngram matching the exact words and calculate geometric mean of precision(papinani, 2002) but it does not support stemmed and synonyms words. This facility is given in METEOR MT-evaluation metric (Banerjee and Lavie, 2005). Meteor matches words position-wise and bleu matches Ngram in continuation.

Both techniques work fine but bleu gives good correlation with human. So we have adopted BLEU's Ngram matching technique and METEOR stemmer and synonym feature in our new metric.

We have also computed the score using F-Measure, BLEU and Meteor metrics and compared these score with our new hybrid metric.

3.5 CLASSIFICATION

Our aim of experiment is estimate the quality of translation for post-editing so that we can filter-out good and bad translations using their evaluation scores. We calculated a threshold for making decision that translation is good or bad by following formula:-

$$T = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} e_{score}}{n+m} \quad (1)$$

Where as :-

If e_{score} >=T

Good Translation

else

Bad translation (2)

Overall process of our experiment can understand by taking the following example:-

English Sentence:-

Tripura is presently open to tourists and special permits are no longer required.

Translation 1(Babylon)

वर्तमान में खुले है त्रिपुरा को पर्यटकों और विशेष परमिट आवश्यक नहीं रह रहे हैं । **Translation**

2(Microsoft)

त्रिपुरा पर्यटकों के लिए वर्तमान में खुला है और विशेष परमिट अब आवश्यक नहीं हैं। **Translation**

3(Google)

Tripura पर्यटकों का खुला हुआ इन दिनों है और असाधारण अनूज्ञा अब नहीं जरूरी है गये हैं

Translation 4(EBMT)

त्रिपुरा खुला है और पर्यटकों को विशेष हैं की आवश्यकता नहीं पड़ता है ।

	Babylon	Microsoft	Google	EBMT	Avg.(T)
Language Expert	0.63	0.6	0.4	0.375	0.5012
Hybrid Metric	0.3265	0.3010	0.20	0.1921	0.2549
Meteor	0.3197	0.4682	0.2447	0.3453	0.344475
Bleu	0.2	0.1875	0.105	0.1432	0.158925
F-Measure	0.1155	0.2494	0.1398	0.2269	0.1829
Threshold					**0.2352**

Table 2. Manual and automatic evaluation score of four MT-Engines

According to above result, we can decide that which translation is good and need little post-editing using following comparison:-

Is 0.3265(Babylon score) is greater than or equal to 0.2352(threshold)

Good Translation

Else

Bad Translation

As above described strategy, we have computed evaluation scores using all metrics and compared these scores with computed threshold. We have achieved good success. Our results of classification received from this process mentioned in Table3.

As above described strategy, we have computed evaluation scores using all metrics and compared these scores with computed threshold. We have achieved good success. Our results of classification received from this process mentioned in Table3.

Evaluation Metrics	Machine Translation Engines							
	Babylon		Microsoft		Google		EBMT	
	Good	Bad	Good	Bad	Good	Bad	Good	Bad
Hybrid Metrics	650	650	680	620	630	670	670	630
BLEU	606	694	669	631	611	689	663	637
METEOR	645	655	759	541	591	709	658	642
F-Measure	691	609	802	498	690	610	792	508
Human evaluation	665	635	657	643	676	624	682	618

Table 3. MT-Engines output's quality estimated scores of all evaluation metrics

If we see in Fig 2. Bleu and hybrid metrics are matching with human judgment comparatively.

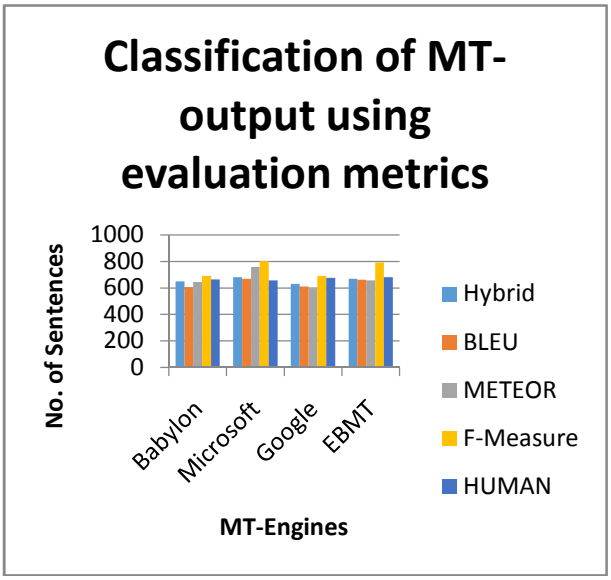

Fig. 2. Comparison of different Evaluation metrics for post-editing.

5. CONCLUSION

Machine translation does not fully accurate and fluent it needs minor checks. To save human effort and cost of post-editing, we need to select only those translations out of all translations those can make fully accurate with little modification. We found that calculated result of geometric mean of Ngram precision is much more nearest to human judgment. BLEU metric follow this computation. We can't evaluate MT-output without references and reference can have multiple words of a single meaning so evaluation metric should use some linguistic tools like stemmer, synonyms matcher and phrase-matcher like METEOR metrics. Use of combine features of BLEU and METEOR gives better results. We can use this mechanism for MT-output quality

estimation and can filter-out good scored translation for post-editing. It effort will save the time and money for post-editing.

REFERENCES

[1] Knight Kevin & Ishwar Chander (1994). Automated post-editing of documents. In Proceedings of the twelfth national.

[2] Simard, M., Goutte, C., & Isabelle, P. (2007, April).Statistical Phrase-based Post-editing. Proceedings of NAACL HLT 2007, ACL , 508-515.

[3] Pierce and Caroll's ALPAC report (1956) persuaded the U.S. government to stop funding (statistical) machine translation (MT) research

[4] VAN SLYPE, Second evaluation of the English-French SYSTRAN machine translation system of the Commission of the European Communities.-Luxembourg, CEC, Final report, November 1978, 179 p.

[5] Falkedal, Kirsten. 1991. Evaluation Methods for Machine Translation Systems. An historical overview and a critical account. ISSCO. University of Geneva. Draft Report

[6] Church, K. W., and E. H. Hovy. (1993). Good applications for crummy machine translation. Machine Translation 8: 239–258

[7] Sanjika Hewavitharana, Bing Zhao, Almut Silja Hildebrand, Matthias Eck, Chiori Hori, Stephan Vogel and Alex Waibel. The CMU Statistical Machine Translation System for IWSLT 2005. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[8] Christoph Tillmann, Stephan Vogel, Hermann Ney, Alex Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In European Conf. on Speech Communication and Technology, pages 2667–2670, Rhodes, Greece, September

[9] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of Association for Machine Translation in the Americas

[10] Papineni, K., Roukos, S., Ward, R. T, Zhu, W.-Z. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, PA.

[11] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan, June.

[12] Michael Denkowski and Alon Lavie. 2010b. METEORNEXT and the METEOR Paraphrase Tables: Improve Evaluation Support for Five Target Languages. In Proc. of ACL WMT/MetricsMATR 2010.

[13] "Plitt, Mirko and Francois Masselot (2010): A Productivity Test of Statistical Machine Translation Post-Editing in A Typical Localisation Context, Prague Bulletin of Mathematical Linguistics, 93:7–16."

[14] V. Lawson, Machine translation and people, ASLIB Proc. 1981

AUTHOR

Kuldeep Kumar Yogi: Assistant Professor, Dept. of Computer Sciecne, Banasthali University, Rajasthan, India

Chandra Kumar Jha: Head Of Dept.(Comp. Sc.), Banasthali University, Rajasthan, India