Evaluation Metric for Machine Translation

DISA Report

Ayush Chaurasia
Arjya Das
Kishan Kumar

IIT Delhi
India
Summer 2018

Contents

1 Introduction

The main objective of this project is to allow the evaluation of the translations of sentences pertaining to Indian legal domain. The reports of the cases from district courts of various states, are primarily available in the local languages of the respective states. When these cases move up to high courts and supreme court, they need to be translated to the official language, English. Now, being of the legal domain, one can afford little to no margin of error for translating these sentences. Hence, a proper evaluation needs to be done post translation to ensure correctness .To evaluate these translations, there are two possible ways -

- Manual evaluation
- Automatic evaluation

Not only is automatic evaluation very fast, it is economically feasible as well, as it eliminates the need to hire bilingual linguists to manually carry out the evaluation.

2 Evaluation Metrics

We have worked on five of the most popular metrics for machine translation evaluation, namely BLEU, METEOR, AMBER, WER and GTM. The reason behind choosing these metrics is that they evaluate a given instance of translation on the basis of different parameters.

- BLEU : Precision
- WER : Levenshtein distance
- METEOR : F-Measure of BLEU 1-Gram and Recall 1-Gram
- AMBER : Precision and Recall
- GTM : Precision and Recall

Now, although these metrics are available online, there is no existing app that readily displays the scores pertaining to these different metrics. Hence, we developed an app with the vision of decreasing the discrepancy and ambiguity in these algorithms, by defining them clearly and henceforth allowing even a non tech-savvy person to automatically generate the scores for evaluating translation quality.

2.1 BLEU

BLEU metric is based on precision. It returns the geometric mean of 1-Gram, 2-Gram, 3-Gram and 4-Gram based precision as the score. In case any of the n-gram score is 0(zero), then it is neglected and the geometric mean for the left n-grams is taken as the score. Also, a penalty by the name of brevity penalty is used in the case number of words in candidate sentence is less than the number of words in reference sentence. A penalty called Brevity Penalty is calculated so as to prevent shorter sentences from having the favor of high score as they are more likely to have less mismatch with the reference as compared to longer counterparts. In the case that the candidate is longer than the reference, the brevity penalty is taken to be 1. The evaluation of 8735 Hindi to English translations was carried out using 2 automatic translators - Google translate and Bing translate. The algorithm for BLEU is given as -

2.1.1 PSEUDOCODE

INPUT ref[1], trans
OUTPUT Bleu Score

Step 1 Set b1 = Bleu1gram(ref,trans)
Set b2 = Bleu2gram(ref,trans)
Set b3 = Bleu3gram(ref,trans)
Set b4 = Bleu4gram(ref,trans)
Set bp = $e^{1-(referenceLength/translationLength)}$

Step 2 If b4 = 0 then
If b3 = 0 then
If b2 = 0 then Set b=b1
else Set b = (b1*b2)$^{\frac{1}{2}}$
else Set b = (b1*b2*b3)$^{\frac{1}{3}}$
else Set b = (b1*b2*b3*b4)$^{\frac{1}{4}}$

Step 3 OUTPUT: Blue score = b*bp
STOP

[1]Refer APPENDIX 5.1

2.1.2 Example

Reference : He had witnessed the incident at a distance of about 7-8 feet.
Candidate : He saw the incident at a distance of 7-8 feet.
b1=$\frac{9}{10}$=0.9
b2=$\frac{6}{9}$=0.667
b3=$\frac{4}{8}$=0.5
b4=$\frac{3}{7}$=0.428
bp=$e^{1-(12/10)}$
bp=0.8187
b=0.598
bleuscore=b*bp =.49

2.1.3 Statistics for BLEU Score

The data from the evaluation has the following properties -

Attribute	Google Translate	Bing Translate
Max	1.0	1.0
Min	0	0
Range	1.0	1.0
Mean	0.26	0.26
Variance	0.0161	0.0141
Standard Deviation	0.1268	0.1189

2.2 GTM

The GTM score is calculated by taking the harmonic Mean of BLEU 1-Gram and Recall 1-Gram scores, where these scores are defined as follows - **BLEU 1-Gram score (a)**: Number of words matching in the candidate and reference sentence divided by the number of words in candidate sentence. **Recall 1-Gram score (b)**: Number of words matching in the candidate and reference sentence divided by the number of words in reference sentence.

$$\text{score} = \frac{2ab}{a+b}$$

2.2.1 PSEUDOCODE

INPUT ref[2], trans
OUTPUT GTM Score

Step 1 Set a = Bleu1gram(ref,trans)
Set b = Recall(ref,trans)

Step 2 If a = 0 and b = 0 then set GTM Score = 0
else set GTM Score = $\frac{2ab}{a+b}$

Step 3 OUTPUT: GTM Score
STOP

2.2.2 Example

Reference : He had witnessed the incident at a distance of about 7-8 feet.
Candidate : He saw the incident at a distance of 7-8 feet.
a=$\frac{9}{10}$=0.90
b=$\frac{9}{12}$=0.75

gtm score= $\frac{2*0.9*0.75}{0.9+0.75}$ = 0.81

2.2.3 Statistics for GTM Score

The data from the evaluation of 8735 Hindi to English translations has the following properties for the GTM metric -

Attribute	Google Translate	Bing Translate
Max	1.0	1.0
Min	0	0
Range	1.0	1.0
Mean	0.56	0.59
Variance	0.0183	0.0177
Standard Deviation	0.1353	0.1330

[2]Refer APPENDIX 5.1

2.3 WER

The word error rate is based on the Levenshtein distance, but it works at the word level instead of the character level. For calculating the Word error rate score, we first determine the number of insertions, substitutions and deletions of words required to get from the reference to the hypothesis. The score is then calculated using the formula -

$$\text{WER} = \frac{Insertions+Substitution+Deletions}{Numberofwordsinreference}$$

The data from the evaluation of 8735 Hindi to English translations has the following properties for the WER metric -

2.3.1 Example

2.3.2 Statistics for WER Score

Attribute	Google Translate	Bing Translate
Max	1.46	1.33
Min	0	0
Range	1.46	1.33
Mean	0.62	0.67
Variance	0.0359	0.0355
Standard Deviation	0.1896	0.1885

2.4 METEOR

The score is calculated by taking F-Measure of BLEU 1-Gram and Recall 1-Gram. F-Measure=10*BLEU 1-Gram*Recall 1-Gram/((9*BLEU 1-Gram)+Recall 1-Gram) .If Recall 1-Gram and BLEU 1-Gram both are equal to zero, then the score is set to zero. The penalty is calculated in METEOR using fragments and words.First w is calculated as the number of matching words between reference and translated sentence.Number of fragments is the number of parts that have the words in maximum continuous chain.

Assumption If number of total words in the fragments is equal to 1 then fragmentation is taken to be equal to 1 which is used to calculate the discounting factor.

2.4.1 PSEUDOCODE

INPUT ref[3], trans
OUTPUT METEOR Score

Step 1 Set a = Bleu1gram(ref,trans)
Set b = Recall(ref,trans)

Step 2 If a = 0 and b = 0 METEOR Score = 0
else Set fm = $\frac{10ab}{9a+b}$
Set fr = frag(ref,trans)
Set df = $\frac{1}{2}$(fr)^3
Set METEOR Score = (1-df)*fm

Step 3 OUTPUT: METEOR Score
STOP

2.4.2 Example

Reference : He had witnessed the incident at a distance of about 7-8 feet.
Candidate : He saw the incident at a distance of 7-8 feet.
a = $\frac{9}{10}$=0.9
b = $\frac{9}{12}$=0.75
fmean=$\frac{10*0.9*0.75}{(9*0.9)+0.75}$ =0.762
w = 9
fragment=3
fr=$\frac{3-1}{9-1}$=0.25
df=0.5*$(0.25)^3$ = 0.0078125
Meteor Score =(1-0.0078125)*0.762 = 0.756

2.4.3 Statistics for METEOR Score

The data from the evaluation of 8735 Hindi to English translations has the following properties for the Meteor metric -

[3]Refer APPENDIX 5.1

Attribute	Google Translate	Bing Translate
Max	1	1
Min	0	0
Range	1	1
Mean	0.53	0.56
Variance	0.0201	0.0226
Standard Deviation	0.1401	0.1503

2.5 AMBER

Amber is a modified BLEU enhanced ranking metric.It incorporates recall, extra penalties and some text processing variants. We take into account geometric mean of n gram precisions(AvgP),F-measure derived from the arithmetic averages of precision and recall (Fmean), arithmetic average of Fmeasure of precision and recall for each n-gram (AvgF) The effective amber score is calculated by taking the product of the score and the penalties. That is,

$$AMBER = score * netpenalty \tag{1}$$

$$\text{p(n)} = \frac{ngrams(T \cap R)}{ngrams(T)}$$
$$\text{r(n)} = \frac{ngrams(T \cap R)}{ngrams(R)}$$

$$AvgP(N) = (\prod_{i=1}^{n} p(n))^{\frac{1}{N}} \tag{2}$$

$$P(N) = \frac{1}{N} \sum_{n=1}^{n} p(n) \tag{3}$$

$$R(M) = \frac{1}{M} \sum_{n=1}^{m} r(n) \tag{4}$$

$$Fmean(N, M, \alpha) = \frac{P(N)R(M)}{\alpha P(N) + (1-\alpha)R(M)} \tag{5}$$

$$AvgF(N, \alpha) = \frac{1}{N} \sum_{n=1}^{N} \frac{p(n)r(n)}{\alpha p(n) + (1-\alpha)r(n)} \tag{6}$$

$$Score(N) = \theta_1 AvgP(N) + \theta_2 Fmean(N, M, \alpha) + (1-\theta_1-\theta_2) AvgF(N, \alpha) \tag{7}$$

The parameters are set as N = 4, M = 1, $\alpha = 0.9$, $\theta_1 = 0.3$ and $\theta_2 = 0.5$

$$\text{Net Penalty} = \prod_{i=1}^{p} pen_i^{w_i}$$

Name of penalty	Weight Value
SBP	0.30
SRP	0.10
CSBP	0.15
CSRP	0.05
SWDP	0.10
LWDP	0.20
CKP	1.00
CTP	0.80
NSCP	0.50
NKCP	2.00

2.5.1 Example

Reference : He had witnessed the incident at a distance of about 7-8 feet.
Candidate : He saw the incident at a distance of 7-8 feet.

$R(1) = \frac{9}{12} = 0.75$ $P(1) = \frac{9}{10} = 0.90$
$R(2) = \frac{6}{11} = 0.545$ $P(2) = \frac{6}{9} = 0.667$
$R(3) = \frac{4}{10} = 0.40$ $P(3) = \frac{4}{8} = 0.500$
$R(4) = \frac{3}{9} = 0.33$ $P(4) = \frac{3}{7} = 0.428$

AvgP(N)=$(0.90*0.667*0.50*0.428)^{0.25} = 0.598$
P(N)=$\frac{0.9+0.667+.5+.428}{4} = 0.623$
R(M)= 0.75
FMean =10*0.623*.75/(9*.623)+.75= 0.735
FMean1 = 0.762
Fmean2 = 0.555
Fmean3 = 0.408
Fmean4 = 0.34
AvgF = $\frac{0.762+0.555+0.408+0.34}{4} = 0.516$
Score = 0.3*0.598 + 0.5*0.735 +0.2*0.516 = 0.65

2.5.2 Statistics for AMBER Score

The data from the evaluation of 8735 Hindi to English translations has the following properties for the AMBER metric -

Attribute	Google Translate	Bing Translate
Max	0.93	0.96
Min	0	0
Range	0.93	0.96
Mean	0.24	0.25
Variance	0.0161	0.0153
Standard Deviation	0.1270	0.1238

2.6 Issues

Although the above-mentioned metrics to some extent do help in determining the quality of translation, they are not consistent with respect to each other in terms of the scores they assign to a translation.
The following examples illustrate this inconsistency :-

Hindi Sentence : जहाँ तक कि स्थायी विकलांगता और जीवन की सुख - सुविधाओं की हानि का संबंध है , ट्रिब्यूनल ने २० , ००० / - रुपये देने का निर्णय दिया है|
Reference sentence: As regards permanent disablement and loss of amenities of life , the tribunal has awarded Rs . 20 , 000 / - .
Bing candidate : As far as permanent disability and the pleasure of life-loss of facilities are concerned , The Tribunal has 20 , 000/ -The Bucks have decided to give.
Google candidate : As far as permanent disability and the loss of life's amenities are concerned, the tribunal has decided to give Rs 20,000 / -.

Metric	Google Translate	Bing Translate
BLEU	0.14	0.12
GTM	0.47	0.46
WER	0.70	0.87
METEOR	0.46	0.49
AMBER	0.24	0.19

As can be observed from the above table, the metrics do not correlate well with each other, with the scores assigned to the translation varying from 0.14 to 0.70 in case of Google and from 0.12 to 0.87 in case of Bing. The table below states the correlation between the five metrics.

.	BLEU	METEOR	GTM	AMBER	WER
BLEU	1	0.60	0.65	0.69	-0.6
GTM	0.61	1	0.96	0.59	-0.58
WER	0.65	0.96	1	0.62	-0.63
METEOR	0.69	0.59	0.62	1	-0.75
AMBER	-0.6	-0.58	-0.63	-0.75	1

Some observations from the above table

- As can be seen from the table, GTM and METEOR are very highly correlated. This is due to the fact that their underlying algorithms are based on recall and precision and are very similar.
- WER is negatively correlated to the other 4 metrics. This is due to the fact that, for an accurate translation, WER scores it closer to 0, whereas the other 4 metrics score it closer to 1, i.e. accuracy is inversely proportional to WER score and is directly proportional to the BLEU, METEOR, GTM, AMBER and WER scores.

3 APP Description

During the course of the project, an app was developed with the motive of aiding in simple evaluation of sentences using the above metrics. The development was done in Python, using the module Tkinter for GUI(Graphical User Interface) and Openpyxl for accessing and editing Excel Sheets. The functioning of the app is mentioned below using screenshots.

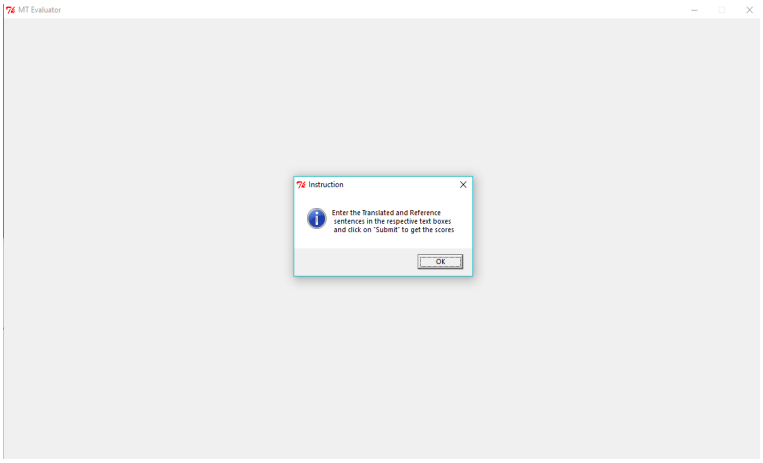

Figure 1: On running the app, a message-box pops up with instructions

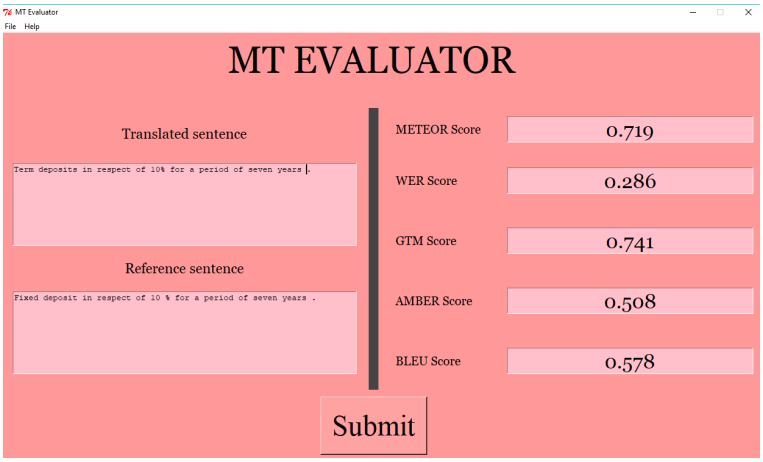

Figure 2: A sample of evaluation of a translation

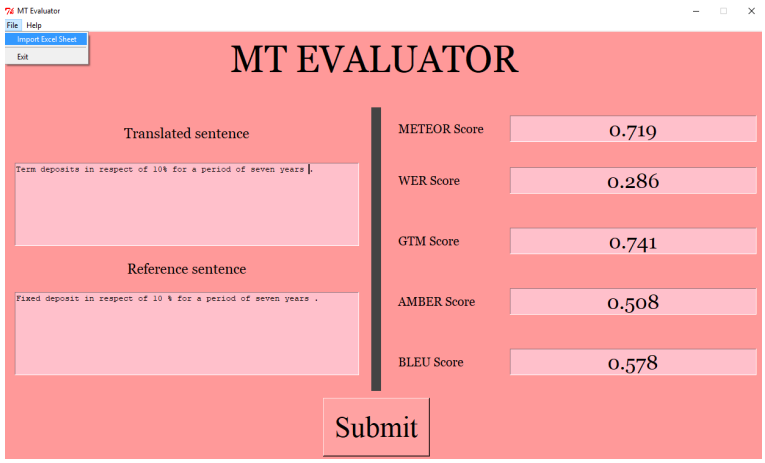

Figure 3: Option in the menu to import an excel file for evaluation

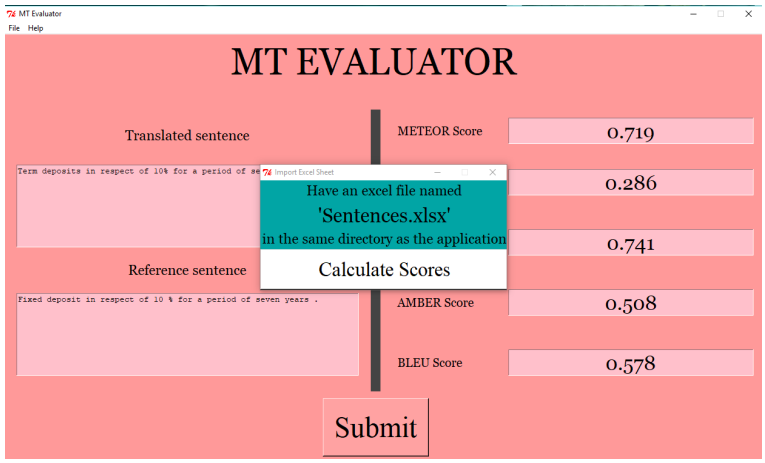

Figure 4: Pop-up window for calculation of score from an excel file

4 Appendix A

Formulas

- ref: Reference sentence
- trans: Translated/Candidate sentence
- refl: Length of Reference sentence
- transl: Length of Translated/Candidate sentence
- Bleu1gram: no. of words matched in ref and trans divided by number of words in translated sentence.
- Bleu2gram: no. of matching consecutive word pairs in ref and trans divided by number of consecutive word pairs in translated sentence.
- Bleu3gram: no. of matching consecutive word triplets in ref and trans divided by number of consecutive word triplets in translated sentence.
- Bleu4gram: no. of matching consecutive word quadruplets of in ref and trans divided by number of consecutive word quadruplets in translated sentence.
- Recall-1Gram: no. of words matching in the trans and ref sentence divided by refl.
- Fragments: Longest continuous chunks of words in sentences common to both the ref and trans so that these chunks cannot be increased further in length.