Machine Translation Evaluation: Manual Vs Automatic - A Comparative Study

Kaushal Kumar Maurya   Renjith P. Ravindran

Ch. Ram Anirudh   K. Narayana Murthy

ICDECT-2019

School of Computer and Information Sciences
University of Hyderabad

15-03-2019

Table of Contents

Machine Translation and MT Evaluation

Definition

Machine Translation(MT) deals with the conversion of natural language texts from one language to another using computers.[1]

Definition

Machine Translation Evaluation deals with judging how good an MT system is[7].

- The evaluation of machine translation is a fundamentally hard problem, since it relates to the unresolved problem of semantic equivalence[7]

Manual Vs Automatic

Manual	**Automatic**
Done by a human well versed in both source and target language	Done by comparing the MT output with reference translations
Humans judge whether meaning is preserved or not directly	Do not attempt to judge meaning directly[4]
Expensive, time consuming and subjective	Inexpensive and quick; useful for tracking progress of an MT systems on fixed data set; for comparing different MT systems
Scores are reliable	Scores may not be meaningful
Metrics: *Adequacy, Fluency , Intelligibility, Fidelity[11]*	Metrics: BLEU[8], NIST[6], METEOR[1], WER[10] & TER[9]

Manual Evaluation Metrics

Metric	Underlying Idea
Adequacy	How well the meaning is captured in translation (TL)
Fluency	How fluent translation is in TL
Intelligibility	How understandable the text is in TL
Fidelity or Accuracy	How much information is retained in the TL
Task-oriented[12]	Judge whether an MT system is suitable for tasks like comprehension, extraction, etc.
Segment ranking[3]	Ranking outputs from various MT systems

Automatic Evaluation Metrics

Metric	Underlying Idea
BLEU	Geometric mean of modified n-gram precision with brevity penalty
NIST	Variant of BLEU with weighted n-gram precision and modified brevity penalty
METEOR	Harmonic mean of Precision and Recall of uni-gram as well as approximate matches (stem, synonyms etc.), using linguistic resources like steamers, Word-net, etc.
METEOR-Hindi	Modified METEOR metric which uses Hindi related resources
WER	Min number of edit operations required to transfer a MT output into a reference translation
TER	Same as WER with additional shift edit

Questions We Wanted To Ask

1. How well do the automatic scores correlate with manual scores?
2. What is the distribution of manual scores for a given interval of automatic scores?
3. Can we estimate the manual metric score for a given automatic metric score?

Choice of Metrics

Manual Metrics

- Checking if meaning is preserved or not is more important
- Therefore, we chose **Adequacy** over **Fluency**

Adequacy: how well translated sentence convey same meaning as input sentence? is phrase or part of text is distorted, added or lost?[7]

Scores	Adequacy
5	all meaning is preserved
4	most meaning is preserved
3	much meaning is preserved
2	little meaning is preserved
1	none of the meaning is preserved

Table: Manual Metric: Adequacy

Automatic Metrics

- BLEU, NIST, METEOR, WER and TER

Data and MT Systems Detail

- Translation direction: English to Hindi
- WMT14[2] published 2507 test sentences with reference translations - we randomly selected 450 sentences from this dataset
- Translation outputs considered from 3 different systems:Online-B[*][1], IIT-BOMBAY[10] and MANAWI-RMOOVE(MR)[11][2]
- Data: 450 x 3 = 1350 <source, reference, system-output> triples

[1] [*]. No exact citation is found for this system because translation outputs are collected by WMT14 organizing committee

[2] ranked 1, 5 and 9 respectively in the shared task WMT14 for English Hindi

Manual Evaluation

- Done by 9 bilingual annotators
- Each annotator evaluates 300 sentences in two rounds: 150 sentences per round
- Each will get equal proportions from all 3 MT systems
- Every system-output will be annotated by exactly 2 annotators (for getting inter-annotator agreement)
- Average of scores from two annotators is considered for further experiments

Inter Annotator Agreement - Kappa Coefficient (k)

Kappa coefficient (k)[5]

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

Where,
P(A): proportion of times the annotators agree
P(E): proportion of times they would agree by chance

Kappa	**Agreement**
< 0	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement
1	Perfect agreement

MT System	**#Sentences**	**k-Values**
Online-B	450	0.2366
IIT-Bombay	450	0.2327
MANAWI-RMOOVE	450	0.2821
All Systems	1350	0.2884

Table: Kappa coefficient interpretation and K-values for inter annotator agreement

Our results of inter annotator agreement are similar to WMT14.

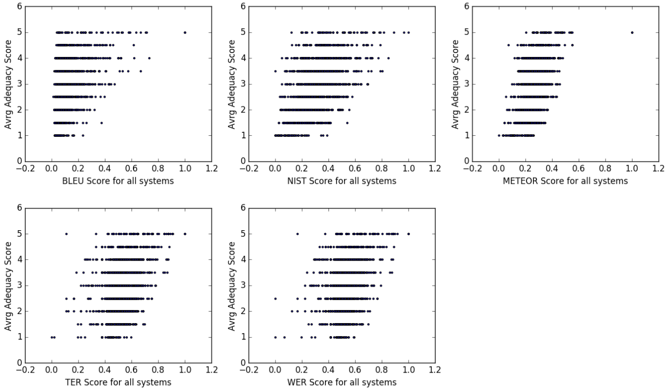
Automatic Evaluation

- Automatic metric scores are computed for all 1350 (450X3) system outputs
- Scores are obtained using open source tools[3] [4] [5]

[3] BLEU and NIST: https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl
[4] METEOR: http://www.cs.cmu.edu/ alavie/METEOR/
[5] TER and WER: http://www.cs.umd.edu/ snover/tercom/

Correlation:Best Automatic Metric

- We find the best automatic metric using correlation scores between average human judgment(adequacy score) and automatic metric scores.
- higher the correlation score better the metric is.



Pearson's correlation coefficient(ρ)[13]

$$\rho = \frac{\Sigma_{i=1}^{n}(H_i - \bar{H})(M_i - \bar{M})}{\sqrt{\Sigma_{i=1}^{n}(H_i - \bar{H})^2}\sqrt{\Sigma_{i=1}^{n}(M_i - \bar{M})^2}}$$

where,
H_i: manual evaluation score of segment i
M_i: automatic evaluation score of segment i
$\bar{H}$: average of manual scores
$\bar{M}$: average of automatic scores

Correlation	Negative	Positive
small	-0.29 to -0.10	0.10 to 0.29
medium	-0.49 to -0.30	0.30 to 0.49
large	-1.00 to -0.50	0.50 to 1.00

Metrics	ρ-Value
BLEU	0.401
NIST	0.481
METEOR	0.513
TER	0.384
WER	0.345

Table: Interpretation of Pearson's correlation coefficient and scores for different metrics

- Highest correlation score of METEOR indicates it as the best automatic metric

Kendall's tau(τ) rank correlation[14]

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

where

$n_0 = n(n-1)/2$
$n =$ number of segments
$n_1 = \sum_i t_i(t_i - 1)/2$
$n_2 = \sum_j u_j(u_j - 1)/2$
$n_c =$ Number of concordant pairs
$n_d =$ Number of discordant pairs
$t_i =$ Number of tied values in the i^{th} group of ties for the first quantity
$t_j =$ Number of tied values in the j^{th} group of ties for the second quantity

Given a set of manual and automatic score pairs:
$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$,
any pair of scores, $(x_i, y_i), (x_j, y_j) : i \neq j$ are:
Concordant if $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$
Discordant if $x_i < x_j$ and $y_i > y_j$; or if $x_i > x_j$ and $y_i < y_j$

Metrics	τ-Value
BLEU	0.287
NIST	0.336
METEOR	0.361
TER	0.269
WER	0.219

Table: Kendall's τ correlation scores for different metrics

- Above Score also indicate that best automatic metric for English-to-Hindi translation pair is **METEOR**
- Automatic scores has weak correlation with manual scores

Distribution: Manual Vs. Automatic

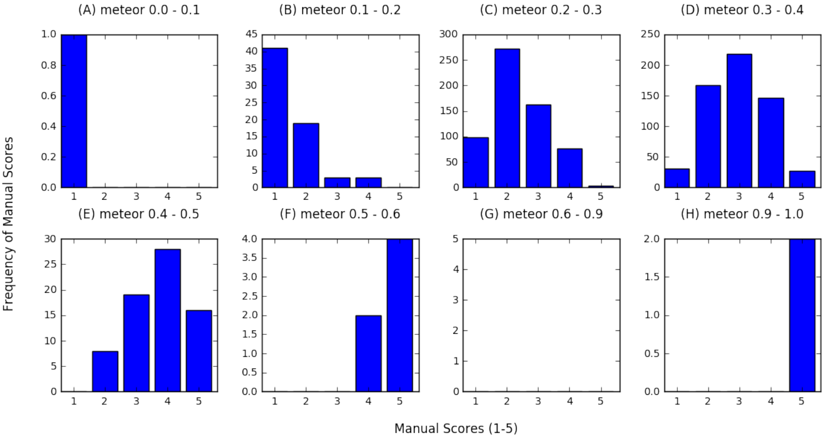


Figure: Distribution of Manual Scores for each interval of Meteor Scores

Meteor Scores	Manual scores
0.0 - 0.1	NA
0.1 - 0.2	1.48 - 1.88
0.2 - 0.3	2.52 - 2.66
0.3 - 0.4	3.11 - 3.26
0.4 - 0.5	3.73 - 4.12
0.5 - 0.6	4.56 - 5.0
0.6 - 0.9	NA
0.9 - 1.0	5.0 - 5.0

Table: 95% Confidence Interval of Manual Scores for Each interval of Meteor Scores

- Automatic scores have a weak correlation with manual scores
- METEOR correlates best with *Adequacy*
- Quality of MT can be estimated from METEOR scores in certain ranges

References I

1. W John Hutchins. *Machine translation: A brief history. Concise history of the language sciences: from the Sumerians to the cognitivists,* pages 431-445, 1995.

2. Philipp Koehn. *Statistical machine translation. Cambridge University Press*, 2009.

3. Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. *In Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12-58. Association for Computational Linguistics Baltimore, MD, USA, 2014.

4. Kenneth W Church and Eduard H Hovy. *Good applications for crummy machine translation. Machine Translation*, 8(4):239-258, 1993.

5. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311-318, 2002.

6. George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Proceedings of the second international conference on Human Language Technology Research*, pages 138-145. Morgan Kaufmann Publishers Inc., 2002.

7. Satanjeev Banerjee and Alon Lavie. 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*, volume 29, pages 65-72, 2005.

8. Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. *In Proceedings of association for machine translation in the Americas*, volume 200, 2006.

References II

9. Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. *In Proceedings of the 14th conference on Computational linguistics-Volume 2. Association for Computational Linguistics*, 433-439.

10. Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, and Pushpak Bhattacharyya. 2014. *The IIT Bombay Hindi English Translation System at WMT 2014.* ACL 2014 (2014), 90.

11. Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation.* 201-206.

12. Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychoogical measurement*, 20(1):37-46, 1960.

13. Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* 50(302):157-175, 1900.

14. Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81-93, 1938.

15. Christian Federmann. Appraise: An open-source toolkit for manual evaluation of machine translation output. The Prague Bulletin of Mathematical Linguistics, 98:25âĂŞ35, September 2012.

16. Stanford NLP Group. Machine Translation. *https://nlp.stanford.edu/projects/mt.shtml*. 7:49AM, 01-08-2017.

Thank You !!!