Machine Translation Evaluation: Manual Versus Automatic—A Comparative Study

Chapter *in* Advances in Intelligent Systems and Computing · January 2020

DOI: 10.1007/978-981-15-1097-7_45

4 authors, including:

Kaushal Kumar Maurya
Indian Institute of Technology Hyderabad
19 PUBLICATIONS **103** CITATIONS
SEE PROFILE

Ram Anirudh
University of Hyderabad
5 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Machine Translation Evaluation: Manual Versus Automatic—A Comparative Study

Kaushal Kumar Maurya, Renjith P. Ravindran, Ch Ram Anirudh and Kavi Narayana Murthy

Abstract The quality of machine translation (MT) is best judged by humans well versed in both source and target languages. However, automatic techniques are often used as these are much faster, cheaper and language independent. The goal of this paper is to check for correlation between manual and automatic evaluation, specifically in the context of Indian languages. To the extent automatic evaluation methods correlate with the manual evaluations, we can get the best of both worlds. In this paper, we perform a comparative study of automatic evaluation metrics—BLEU, NIST, METEOR, TER and WER, against the manual evaluation metric (*adequacy*), for English-Hindi translation. We also attempt to estimate the manual evaluation score of a given MT output from its automatic evaluation score. The data for the study was sourced from the Workshop on Statistical Machine Translation WMT14.

Keywords Machine translation (MT) · MT evaluation · Manual metrics · Automatic metrics

1 Introduction

Machine translation (MT) deals with the conversion of natural language texts from one language to another using computers. Developing techniques to adequately judge the quality of machine translation has been a major concern in the MT

K. K. Maurya (✉) · R. P. Ravindran · C. R. Anirudh · K. N. Murthy
School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India
e-mail: kaushalmaurya94@gmail.com

R. P. Ravindran
e-mail: rpr@uohyd.ac.in

C. R. Anirudh
e-mail: ramanirudh28@gmail.com

K. N. Murthy
e-mail: knmuh@yahoo.com

K. S. Raju et al. (eds.), *Data Engineering and Communication Technology*,
Advances in Intelligent Systems and Computing 1079,
https://doi.org/10.1007/978-981-15-1097-7_45

research community. The main concern is whether the meaning of the source sentence is properly preserved in the target sentence. Therefore, the quality of machine translation output is best judged by a human, well versed in both source and target languages. However, like any other cognitive task, manual evaluation of MT is tedious, time-consuming, expensive and can be inconsistent. The general tendency, therefore, is to look for automatic techniques for evaluation. However, automatic evaluation of translation is hard, as computers are incapable of judging the meaning directly. Automatic evaluation metrics try to indirectly capture the meaning by comparing MT output with professional translations of the source sentence, called *reference translations*. Automatic techniques mainly rely on string comparisons [5], and it is well known that they fail to objectively judge the meaning conveyed in all cases. However, automatic techniques, being fast and consistent, can be used to track the progress in MT system development on a fixed data set. Automatic metrics, in spite of their limitations, are also used today by the MT community to judge and compare the performance of MT systems. Correlation studies on automatic and manual metrics show that automatic metrics can be useful in practice [5], although it is hard to adequately judge the quality of translations.

Adequacy and *fluency* are two of the widely used manual evaluation metrics today. *Adequacy* measures how much of the information in the source sentence is preserved in the translation [28]. *Fluency* measures how good the translation is with respect to the target language quality in terms of intelligibility and flow [28]. One or more human annotators score the output produced by an MT system using these metrics. On the other hand, automatic metrics like BLEU [23] and NIST [10] score the MT output based on lexical similarity with reference translations. Lexical similarity is computed through n-gram statistics and therefore is sensitive to word ordering and synonymy. Other automatic metrics like TER [25] and METEOR [1] try to address these issues to a certain degree.

It would be great if we can exploit the advantages of automatic evaluation and at the same time get a better feel for the actual quality of translations. The goal of this paper is to check for correlation between manual and automatic evaluation, specifically in the context of Indian languages. To the extent automatic evaluation methods correlate with the manual evaluations, we can get the best of both worlds. In this paper, we perform a comparative study of automatic evaluation metrics—BLEU, NIST, METEOR, TER and WER [26], against the manual evaluation metric (*adequacy*), in the context of English-Hindi translation. We also attempt to estimate the manual evaluation score of a given MT output from its automatic evaluation score. The data for the study was sourced from the Workshop on Statistical Machine Translation WMT14 [3].

2 Literature Survey

Intelligibility and *fidelity* were the first two manual evaluation metrics used by Automatic Language Processing Advisory Committee (ALPAC) [6]. In early 1990s, Advanced Research Projects Agency (ARPA) proposed three manual evaluation metrics, viz. *adequacy*, *fluency* and *comprehension* [28] in different MT evaluation campaigns. Few more extended criteria such as *suitability*, *interoperability*, *reliability*, *usability*, *efficiency*, *maintainability* and *portability* were discussed by King et al. [18]. A task-oriented metric was developed by White et al. [29] which can be used to judge whether an MT system is suitable for a given task such as publishing, gisting or extraction. Farrús et al. [12] and Costa-jussà [8] proposed an objective method for manual evaluation which takes various linguistic aspects like orthography, morphology, syntax and semantics into account. They provide guidelines to classify the MT output errors. In recent evaluation campaigns of WMT [4], *segment ranking* is used where judges rank the sentences from different systems according to their quality. This gives a relative scoring between the MT systems, and to choose the best MT system. This cannot be used to judge the quality of MT output as such. One of the problems with manual evaluation is that different human evaluators may disagree on the scores for the same MT output. The results may be subjective and irreproducible. Also, human evaluators are expensive in terms of money and time.

The core idea behind automatic evaluation is: "*the closer the machine generated translation is to a professional human translation, the better it is*" [23]. Automatic evaluation techniques use lexical similarity measures like the edit distance and overlap in n-gram sequences for measuring the closeness between machine output and a reference translation. Translation Error Rate (TER) and Word Error Rate (WER) [26] are edit-distance-based metrics, and BLEU, NIST, METEOR, etc., are n-gram sequence-based metrics. Edit distance is concerned with the calculation of minimum number of edit operations required to transform a given translation into a reference translation. WER finds the proportion of *insertions*, *substitutions* and *deletions* in the output with respect to reference translation. Hence, the higher the WER, the lower is the performance of MT system. It counts reordering between words as *deletions* and *insertions*, increasing the WER score. TER overcomes this problem, by also considering *shift* in addition to the above three operations.

BLEU is one of the most widely used automatic metrics today. The BLEU score is calculated by taking the product of the geometric mean of modified n-gram precision scores with brevity penalty. Brevity penalty penalizes the score if the output sentence is shorter than the reference sentence. The drawback of BLEU is that it gives equal weightage to all words and it fails if exact n-grams are not present in the reference translations [5].

NIST is a modification of BLEU. It uses arithmetic mean of n-gram matching precision scores instead of geometric mean. This is to avoid the nullification of the score if one or more n-grams do not match with the reference. The precision scores are weighted by information weights that heavily weigh infrequently occurring

n-grams. Further, this approach modifies the brevity penalty so that small variations in sentence length do not affect the score much.

METEOR [1] is based on the matching of unigrams, matching of morphological variants based on their stems and matching synonyms. METEOR requires linguistic resources such as morphological analyzers for stemming and WordNet [21] for matching synonyms. For Indian languages, Gupta et al. [17] proposed an automatic evaluation metric *METEOR-Hindi* for Hindi as target language which is a modified version of original METEOR (uses Hindi-specific language resources).

Giménez et al. [16] proposed a metric that involves various linguistic features from shallow syntactic similarity, dependency parsing, shallow semantic similarity and semantic roles. They later extended it to include discourse representation structures also [15].

Gautam et al. [14] proposed another automatic metric called "LAYERED," based on three layers of NLP: lexical, syntactic and semantic. In lexical layer, BLEU is considered as the baseline metric. Syntactic layer focuses on reordering of sentences. Semantic layer uses features from a dependency parse of the sentence.

More recently proposed, COBALTF [13] uses target *language models* (LM)-based features to measure fluency of candidate translations. Features have been classified as adequacy-based and fluency-based features. Adequacy features are based on counts of words and n-grams aligned in the target and reference translations. Fluency-based features rely on the LM probability of candidate translation and reference translation, linguistic features like POS information, percentage of content/function words, etc.

3 Setup of the Experiments

3.1 Source of Data

All data required for our experiments were taken from the English-Hindi translation task of the Workshop on Statistical Machine Translation, 2014 edition (WMT14) [3]. In this translation task, participants were required to translate a shared test set. WMT14 hosted ten translation tasks—between English and each of Czech, French, German, Hindi and Russian in both directions. The more recent editions of WMT have not included the English-Hindi translation task; we could therefore source data only from WMT14. The test set for the English-Hindi translation task in WMT14 had 2507 English sentences (source) along with corresponding translations in Hindi (reference). As manual evaluation is expensive, we have restricted our studies to a subset of this whole data. We have made a random selection of 450 source-reference pairs from the 2507 source-reference pairs available. For each of the 450 source-reference pairs, we selected corresponding system-outputs from three MT systems, out of a total of 12 systems that competed in the English-Hindi translation task in WMT14. Thus, we have a total of $450 \times 3 = 1350$ triples where

each triple is <source, reference, system-output>. The three MT systems we have chosen are: online-B, IIT-BOMBAY(IIT-B) [11] and MANAWI-RMOOVE [27] which were ranked in English-Hindi task in WMT14 as 1st, 5th, and 9th, respectively. Our choice of systems was to ensure a representative range of MT output quality. The IIT-B system is a factored SMT system. MANAWI-RMOOVE system is an improvement over the MOSES toolkit [20], and it uses Multi-Word expression and Named-Entity recognition. Online-B system is an online machine translation service that was anonymized by WMT14, for which translations were collected by the WMT organizing committee.

3.2 *Choice of Metrics*

Manual evaluation metrics are supposed to capture two different aspects of translation quality: *Adequacy* and *Fluency*. There can be fluent translations that may not be adequate and there may be adequate translations that are not fluent. Preservation of meaning being the most essential requirement in translation, adequacy is more important. We use adequacy alone for all our manual evaluations. Each sentence is assigned a score ranging from 1 to 5 based on the criteria mentioned in Table 1 [19].

We use the following metrics for automatic evaluation: BLEU [23], NIST [10], METEOR [1], TER [25] and WER [26]. These were chosen as their open implementations are available on the Internet and also they do not require elaborate linguistic resources. METEOR-Hindi is tailor-made for Hindi as a target language, but an open implementation is not available and we were not able to source it from the authors. METEOR allows inclusion of linguistic resources as modules, but only a few of those are openly available for Hindi. Therefore, in our experiments, we have used only the unigram module and the synonymy module through the Hindi-WordNet [22].

Automatic evaluation scores for all the five metrics mentioned above were computed on the entire test corpus (1350 sentences). Both manual and automatic evaluations are performed at segment level. A segment is a unit of translation which is usually one or a few sentences [10].

Segment-level scores are computed using the script `mtevalv13a.pl`[1] which is a part of Moses tool kit and was used in WMT [4]. We used the tool `meteor-1.5`[2] [9] for computing METEOR scores. TER and WER scores were obtained using the tool `tercom.7.25`[3] which implements the original idea of TER proposed by Snover et al. [25]. TER and WER are error metrics with values

[1] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl.

[2] http://www.cs.cmu.edu/∼alavie/METEOR/.

[3] http://www.cs.umd.edu/∼snover/tercom/.

Table 1 Manual evaluation: adequacy

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

ranging from 0 to 100. Higher scores indicate worse quality. We subtract the TER and WER scores from 100 for making them consistent with other metrics.

3.3 Manual Evaluation Setup

Manual evaluation was done by nine bilingual annotators. None of our annotators are professional translators, and they are graduate-level students with Hindi as their first language and English their second language during their studies. The evaluation experiment is set up as follows: (1) Each annotator will annotate 300 system-outputs in two rounds with 150 sentences in each. (2) Each will get equal proportions from all three MT systems. (3) No two annotators will get same system-outputs from the same MT systems. (4) Every system-output will be annotated by exactly two annotators (for getting inter-annotator agreement).

Before the actual evaluations were conducted, a pilot run with two annotators was carried out to get a fair understanding of common mistakes and difficulties the annotators would face during evaluation. The findings of the pilot run are as follows: (1) Annotators were sometimes inconsistent in the way they judge the sentence—sometimes they look at the source sentence first, sometimes they look at the reference translations, sometimes they directly judge the output. (2) The annotators were using the English word to fill the meaning gap in case of untranslated words. (3) A few annotators did not take any breaks during the evaluation process. This can cause fatigue.

Before the actual evaluation, we discussed the above points with all nine annotators and gave the following instructions that would help them do a fair and consistent evaluation. (1) Read the source sentence and the reference translation before scoring system-output. (2) Untranslated words should simply be treated as untranslated words. (3) At least one break was made mandatory to avoid fatigue and boredom during the evaluation. For statistics regarding the manual evaluation refer to Table 2.

Table 2 Statistics of manual evaluation

Min./Max./Avg. Time taken per annotator	50/250/82 min
Min./Max./Avg. Time per sentence (overall)	0.33/1.66/0.54 min

Table 3 Interpretation of *k*-values for inter-annotator agreement

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

Table 4 *K*-values obtained for measuring inter-annotator 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

Inter-Annotator Agreement: Inter-annotator agreement scores are a measure of reliability of manual evaluation. We measure the pairwise inter-annotator agreement for each system as well as for the whole data using kappa coefficient (k) [7].

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

where $P(A)$ is the proportion of times the annotators agree and $P(E)$ is proportion of times they would agree by chance. The range of k value lies between 0 and 1 where 1 indicates perfect agreement and 0 no agreement. Interpretation of the kappa coefficient (k) is shown in Table 3. Results (k-values) are shown in Table 4.

From Table 4, it can be concluded that there is a fair inter-annotator agreement in all cases. We take these manual evaluations as reliable.

As mentioned earlier, each segment is annotated by two annotators. Average of adequacy scores from both annotators is considered as final manual evaluation score [19].

3.4 Correlation: Manual Versus Automatic

To find the segment-level correlation between automatic and manual evaluation, we use the Pearson's rho (ρ) correlation coefficient and Kendall's tau (τ) rank correlation coefficient.

Pearson's Correlation Coefficient: Pearson's correlation coefficient [24] ρ is given by:

Table 5 Interpretation of Pearson's ρ correlation coefficient

Correlation	Negative	Positive
Small	−0.29 to −0.10	0.10–0.29
Medium	−0.49 to −0.30	0.30–0.49
Large	−1.00 to −0.50	0.50–1.00

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

where H_i is the manual evaluation score of segment i. M_i is the automatic evaluation score of segment i. $\overline{H}$ and $\overline{M}$ are the average of manual and automatic scores, respectively. Range of ρ lies between −1 and +1 where $\rho = 1$ counts as perfect correlation, $\rho = 0$ is total independence between two evaluation scores, and $\rho = -1$ indicates very strong negative correlation. Interpretation of ρ value is given in Table 5.

Kendall's τ Rank Correlation Coefficient: The advantage of Kendall's τ rank correlation coefficient over Pearson's ρ is that it does not assume a normal distribution of data. But it needs the data to be ranked. We order the sentences based on their adequacy scores and use them to calculate Kendall's tau. We used a variant of Kendall's tau [2] called Kendall's tau b:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{3}$$

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 and t_j = number of tied values in the j^{th} group of ties for the second quantity.

For a given set of manual score and automatic score pairs, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, any pair of scores, (x_i, y_i) and (x_j, y_j) such that $i \neq j$, are said to be concordant if $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant if $x_i < x_j$ and $y_i > y_j$; or if $x_i > x_j$ and $y_i < y_j$. The pair is a tie, if $x_i = x_j$ and $y_i = y_j$.

4 Results and Analysis

4.1 Correlation: Manual Versus Automatic

Manual versus automatic evaluation scores for the five automatic metrics are given as separate scatter plots in Fig. 1. Each plot contains 1350 data points corresponding to 1350 system-outputs. The x-axis in each plot gives the automatic metric

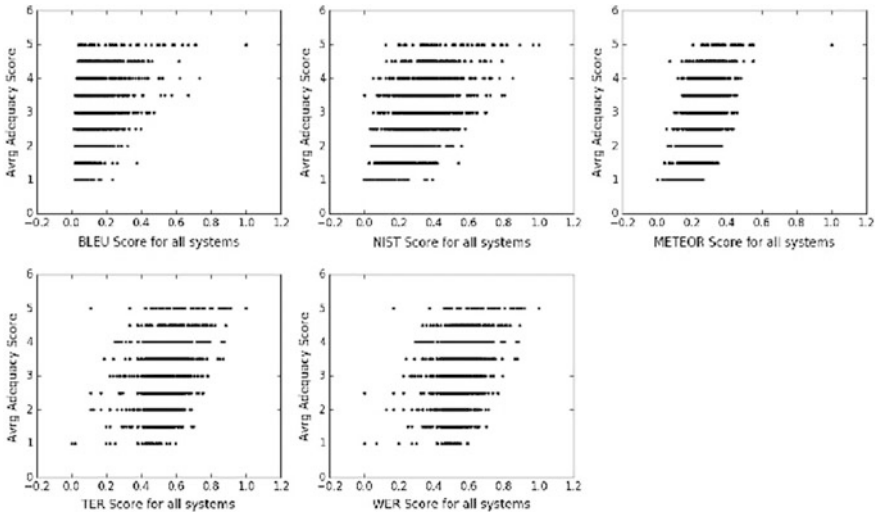

Fig. 1 Automatic metric score versus average adequacy score

score for a given MT system-output, and the *y*-axis gives the corresponding adequacy score. The plot gives a fair idea about the correlation between automatic and manual evaluation scores. It is evident that the correlation is weak as there is a substantial spread in the data points. The spread is minimum for METEOR; therefore, we can infer that METEOR scores agree the most with human judgments. However, to quantify the correlation we test the same using Pearson's and Kendall's correlation coefficients.

Tables 6 and 7 show segment-level Pearson's correlation coefficients and Kendall's tau correlation coefficients, respectively. Both the correlation coefficients suggest that METEOR correlates the best with manual evaluation scores.

4.2 Distribution: Manual Versus Automatic

One of the goals of this study is to see if manual evaluation score for a given MT output can be reliably estimated given its automatic evaluation score. Automatic

Table 6 Pearson's ρ correlation coefficient for different metrics

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

Table 7 Kendall's τ correlation scores for different metrics

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

metrics are used to compare the relative performance of MT systems. But it is not clear whether automatic metrics can be used to absolutely judge translation quality. The correlation study above showed that METEOR has the best correlation with human judgments at segment level. Therefore, we now consider METEOR scores to estimate segment-level manual evaluation scores.

First, we observe how adequacy scores are distributed in the METEOR score range of 0.0–1.0. For this, we bin the METEOR scores into 10 bins each having an interval of 0.1. For each bin, we depict the distribution of adequacy scores as histograms in Fig. 2. The histogram shows how many segments scored adequacy scores of 1–5 in each interval of METEOR scores. Note that the range of *y*-axis for the bins is not normalized and is therefore different for each bin.

A majority of segment scores fall in the METEOR score interval of 0.2–0.4 (C, D in Fig. 2). The distribution of adequacy scores is very regular in the METEOR interval 0.1–0.5 (B–E in Fig. 2), with the central trend of adequacy score shifting positively with the METEOR scores. Bins in the interval 0.6–0.9 (G in Fig. 2) are empty as no segments received a METEOR score in this interval. Also, the first bin and the last bin have very little data points to make any valid conclusions.

This distribution study shows that there is some statistical evidence for the correlation between automatic scores and adequacy scores in the range 0.1–0.6 of METEOR scores. In order to get a rough estimate of manual scores for a given range of METEOR scores, we calculate 95% confidence interval [4] for the mean of adequacy scores in each bin. A 95% confidence interval for a given bin does not indicate that there is a 95% probability that the mean for a given bin lies within the interval. Instead, it tells us that the calculated interval will include the true mean for a given bin with a probability of 95%. Table 8 lists the most probable (95% confidence interval) range of adequacy scores for a given interval of METEOR score.

[4]We used the tconfint_mean function available in the statsmodels package in Python.

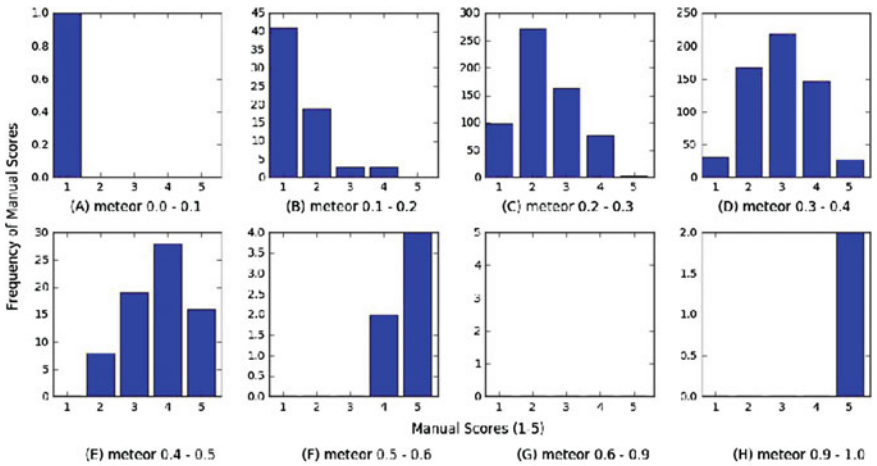

Fig. 2 Distribution of manual scores for each interval of METEOR scores

Table 8 95% Confidence interval for the mean of manual scores for each bin 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

5 Conclusions and Future Work

In this paper, we have presented an empirical study to compare automatic machine translation evaluation metrics with the manual evaluation metric *Adequacy*. We see that automatic metrics have a weak correlation with *adequacy*. However, among the various metrics considered, METEOR correlates best with *adequacy*. We see that for METEOR scores in the 0.1–0.6 interval, we can get a somewhat better idea of the adequacy scores. Thus, the quality of MT can be estimated from METEOR scores in certain situations. Our data did not contain METEOR scores in the interval 0.6–1.0, although we used the MT system that performed best in the WMT14 [3]. A more thorough study including other carefully selected language pairs and MT paradigms including rule-based and neural MT systems will be useful.

References

1. Banerjee, S., Lavie, A.: 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, vol. 29. University of Michigan, Ann Arbor, pp. 65–72 (2005)
2. Bojar, O. et al.: Findings of the 2013 workshop on statistical machine translation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Sofia, Bulgaria, pp. 1–44 (2013)
3. Bojar, O. et al.: Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Association for Computational Linguistics. Baltimore, Maryland, USA, pp. 12–58 (2014)
4. Bojar, O. et al. Findings of the 2016 conference on machine translation (WMT16). In: Proceedings of the First Conference on Machine Translation (WMT), vol. 2. Berlin, Germany, pp. 131–198 (2016)
5. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluation the role of Bleu in machine translation research. In: EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy, pp. 249–256 (2006)
6. Carroll, J.B.: An experiment in evaluating the quality of translations. Mech. Transl. Comput. Linguist. **9**(3-4), 55–66 (1966)
7. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
8. Costa-jussà, M.R., Farrús, M.: Towards human linguistic machine translation evaluation. Digit. Scholarsh. Humanit. **30**(2), 157–166 (2015)
9. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, Maryland, USA, pp. 376–380 (2014)
10. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc. San Diego, California, pp. 138–145 (2002)
11. Dungarwal, P. et al.: The IIT Bombay Hindi-English translation system at WMT 2014. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, Maryland, USA, pp. 90–96 (2014)
12. Farrús, M., Costa-jussà, M.R., Popović Morse, M.: Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. J. Assoc. Inf. Sci. Technol. **63**(1), 174–184 (2012)
13. Fomicheva, M. et al.: CobaltF: a fluent metric for MT evaluation. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, vol. 2, pp. 483–490 (2016)
14. Gautam, S., Bhattacharyya, P.: LAYERED: metric for machine translation evaluation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, Maryland, USA, pp. 387–393 (2014)
15. Giménez, J., Màrquez, L.: A smorgasbord of features for automatic MT evaluation. In: Proceedings of the Third Workshop on Statistical Machine Translation. StatMT '08. Association for Computational Linguistics, Columbus, Ohio, pp. 195–198 (2008). ISBN: 978-1-932432-09-1
16. Giménez, J., Màrquez, L.: Linguistic features for automatic evaluation of heterogenous MT systems. In: Proceedings of the Second Workshop on Statistical Machine Translation. StatMT '07. Association for Computational Linguistics, Prague, Czech Republic, pp. 256–264 (2007)
17. Gupta, A., Venkatapathy, S., Sangal, R.: Meteor-Hindi: automatic MT evaluation metric for hindi as a target language. In: Proceedings of ICON-2010: 8th International Conference on Natural Language Processing. Macmillan Publishers, Kharagpur, India (2010)

18. King, M., Popescu-Belis, A., Hovy, E.: FEMTI: creating and using a framework for MT evaluation. In: Proceedings of MT Summit IX. New Orleans, USA, pp. 224–231 (2003)
19. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2009). Chap. 8
20. Koehn, P. et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics. Prague, Czech Republic, pp. 177–180 (2007)
21. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
22. Narayan, D. et al.: An experience in building the indo WordNet-a WordNet for Hindi. In: First International Conference on Global WordNet. Mysore, India (2002)
23. Papineni, K. et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, pp. 311–318 (2002)
24. Pearson, K.: 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. London, Edinburgh, Dublin Philos. Mag. J. Sci. **50**(302), 157–175 (1900)
25. Snover, M. et al. 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, "Visions for the Future of Machine Translation". Cambridge, Massachusetts, USA, pp. 223–231 (2006)
26. Su, K.-Y., Wu, M.-W., Chang, J.-S.: A new quantitative quality measure for machine translation systems. In: Proceedings of the 14th Conference on Computational linguistics, vol. 2. Association for Computational Linguistics, Nantes, pp. 433–439 (1992)
27. Tan, L., Pal, S.: Manawi: using multi-word expressions and named entities to improve machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, Maryland, USA, pp. 201–206 (2014)
28. White, J., O'Connell, T., O'Mara, F.: The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In: Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas. Columbia, Maryland, USA, pp. 193–205 (1994)
29. White, J.S., Taylor, K.B.: A task-oriented evaluation metric for machine translation. In: Proceedings of Language Resources and Evaluation Conference, LREC-98, vol. 1. Granada, Spain, pp. 21–27 (1998)