Automatic Evaluation Metrics for Multi-Sentence Texts

Eshwar S R Karan Jeswani Rahul John Roy

Abstract

Human evaluations of model generated text are accurate, but expensive and slow for the purpose of model development. Evaluating the output of such systems automatically, saves time, accelerates further research on the text generation tasks and it will also be free of human bias. We provide an in-depth review and comparison of traditional metrics which is based on n-gram word matching to the recently published ones where textual embeddings are compared. We also provide the correlations of these metrics with human evaluation.

1. Problem Description

In this work[1], we provide a detailed overview of evaluation metrics from the n-gram comparison based BLEU and ROUGE to more recent ones like Word mover and Sentence mover similarity scores and BERTScore.

The objective is to find out the relevance of these metrics in tasks like essay evaluation. Various single and multi sentence datasets were considered in this work. The correlations of the metrics are calculated from the available human scores, thus demarcating the different metrics in this application.

Finally, the performance of embeddings from pre-trained transformer models are also examined.

2. Literature Review

BLEU is a modified precision based metric widely used in Machine Translation. It works by calculating n-gram co-occurence statistics between reference and candidate sentence (Papineni et al., 2002), where each word in the candidate text is clipped by the maximum number of the same word occurring in any reference.

BLEU uses the geometric mean of the modified n-gram precisions. A brevity penalty (BP) factor has been introduced, by which the candidate translations are penalized unless they are of the same length as the reference text. The brevity penalty BP is given by:

$$BP = \begin{cases} 1 & \text{if c > r} \\ e^{1-\frac{r}{c}} & \text{if c} \leq \text{r} \end{cases}$$

where, r is the effective reference length or the 'best matching reference length' in the corpus; and c is the length of the candidate translation. The BLEU score is given by:

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} \omega_n log(p_n)\right)$$

where, p_n are the modified n-gram precisions up to length N and ω_n are positive weights summing up to 1. In the Baseline BLEU, N = 4 and $\omega_n = 1/N$.

The ROUGE score (Lin, 2004) is widely used for evaluating summaries. It is a word matching based metric like BLEU. ROUGE-L is the variation of ROUGE which considers 'n-gram' recall where n is the length of the longest common sub-sequence.

Earth mover's distance (EMD) (Rubner et al., 1998) is a measure of closeness of 2 different distributions. Let a signature P have m clusters with $P = \{(p_1, w_{p1}), (p_2, w_{p2}), ..., (p_m, w_{pm})\}$, where p_i is the cluster representative and w_{pi} is the weight of the cluster. Let another signature $Q = \{(q_1, w_{q1}), (q_2, w_{q2}), ..., (q_m, w_{qm})\}$ have n clusters.

Let's define $D = [d_{i,j}]$ to be the ground distance between clusters p_i and q_j.

The idea is to find a flow $F = [f_{i,j}]$ between p_i and q_j, that minimizes the overall cost. subjected to the constraints:

$$f_{i,j} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n$$

$$\sum_{j=1}^{n} f_{i,j} \leq w_{pi}, 1 \leq i \leq m$$

$$\sum_{i=1}^{m} f_{i,j} \leq w_{qj}, 1 \leq j \leq n$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} = \min\left\{\sum_{i=1}^{m} w_{pi}, \quad \sum_{j=1}^{n} w_{qj}\right\}$$

[1] https://github.com/EshwarSR/AutomaticEvaluationMetrics

The optimal flow F is found by solving this linear optimization problem. EMD is defined as the work normalized by the total flow:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}}$$

This metric was initially proposed to measure the similarity of images. After the advent of vector representation of text, it was applied in the NLP domain, for measuring text similarity.

The words are represented as points in a d-dimensional space (dimension of embedding vectors). For calculating the Word mover's distance (WMD) (Kusner et al., 2015) (i.e EMD in this context), the distance between the points can be euclidean distance and the weights of each point given by it's term frequency. In the proposed paper, the authors used Word2Vec for computing the embeddings.

Following the example of (Kilickaya et al., 2017), WMD can be transformed into Word mover's similarity (WMS), given by:

$$WMS(A,B) = \exp\left(-WMD(A,B)\right)$$

Sentence mover's similarity (SMS) (Clark et al., 2019) is very similar to WMS, and proposed two different variations.

SMS, where embeddings are calculated for each sentence instead of individual words and S+WMS, where embeddings from both sentences and words are used in computing the scores. The sentence embeddings can be computed as the mean of the embeddings of the constituent words.

The BERTScore (Zhang et al., 2019) evaluation metric uses contextual embedding for words from pre-trained transformer models. These models are extensions of BERT, they have been trained on more data than BERT and are known to give better results than BERT. They are trained on tasks such as Masked Language Modelling (MLM) and Next Sentence Prediction (NSP).

BERTScore uses this pre-trained transformer model's hidden state as the word piece embeddings and calculates sentence similarity by summing each word's cosine similarity with greedy matching.

BERTScore performs exceedingly well on the WMT18 dataset (single sentence text), and appears to capture paraphrases and word context well.

3. Experiments

As per our objective, we perform a correlation study between the model scores and human scores on 4 different datasets namely:

1. ASAP AES[2]: Automatic essay scoring dataset that has essays written on particular prompts. We use student essays which are 5-15 sentences long with single reference. We have taken the exact setup as that of (Clark et al., 2019).
2. CNN / DailyMail[3]: Multi-sentence machine generated summaries from 3 systems. We have taken a similar setup as that of (Clark et al., 2019). They used the human annotated subset from (Chaganty et al., 2018) and filtered the samples which have at least 2 human grades.
3. ASAP SAS[4]: Short answer scoring with multiple references.
4. WMT 18[5]: Machine Translation dataset is a single sentence dataset from various systems of machine translation.

The datasets are publicly available and were assessed using the metrics discussed above. We have reported results on the ASAP AES, CNN / DailyMail datasets and ASAP SAS, as these are multi-sentence essays and summaries respectively. AES had a common reference for all the candidates while, the DailyMail dataset has separate references for every candidate. In ASAP SAS, essays with a score of 2 was taken as reference, thus giving 441 references for each candidate.

BLEU was implemented as per the study (Papineni et al., 2002). ROUGE has been adapted from (Lin, 2004). It is the also called ROUGE-155. We have used code from their GitHub repo[6].

The WMD[7] implementation was taken and extended for SMS and S+WMS. Embeddings from GloVe, BERT, ELMo and RoBERTa Large were considered in the case of SMS, WMS and S+WMS.

Sentence-BERT[8] (Reimers & Gurevych, 2019) is a recent method to calculate sentence embeddings using siamese and triplet network structures to derive semantically meaningful sentence embeddings. We used 'bert-large-nli-stsb-mean-tokens' with SMS because it has maximum score on the STS (Semantic Textual Similarity) benchmark. This uses cosine similarity to measure closeness of sentences.

The WMT 18 dataset was used to verify the results obtained from the implementation of BERTScore by the authors[9] (Zhang et al., 2019). BERTScore gives high correlations for this dataset as compared to any other metric.

[2] https://www.kaggle.com/c/asap-aes

[3] https://bit.ly/price-of-debiasing

[4] https://www.kaggle.com/c/asap-sas

[5] http://www.statmt.org/wmt18/index.html

[6] https://github.com/pltrdy/rouge/

[7] https://github.com/src-d/wmd-relax

[8] https://bit.ly/36NkG6q

[9] https://github.com/Tiiiger/bert_score/

Table 1. Correlations for different metrics on various datasets

Spearman Rank Correlations			
Score Metric	Datasets		
	CNN	AES	SAS
BLEU	0.037	**0.615**	0.050
ROUGE-L	0.102	0.337	0.027
WMS (GloVe)	0.180	0.429	0.082
SMS (GloVe)	0.257	0.449	0.116
S+WMS (GloVe)	0.214	0.488	0.090
WMS (ELMo)	0.160	0.440	0.015
SMS (ELMo)	0.253	0.438	0.063
S+WMS (ELMo)	0.203	0.483	0.031
WMS (BERT)	0.169	0.368	0.034
SMS (BERT)	0.229	0.412	0.070
S+WMS (BERT)	0.199	0.414	0.048
WMS (RoBERTa Large)	0.163	0.304	0.088
SMS (RoBERTa Large)	0.244	0.463	0.081
S+WMS (RoBERTa Large)	0.200	0.460	0.103
BERTScore (P)	0.178	0.135	0.054
BERTScore (R)	**0.263**	0.536	0.113
BERTScore (F1)	0.255	0.147	0.088
SMS (Sentence-BERT)	0.115	0.297	**0.124**

4. Results

The correlations for each metric with the human evaluations can be seen in Table 1.

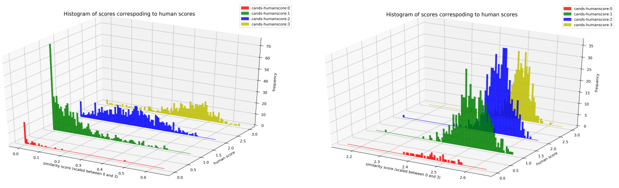
In the CNN/DailyMail dataset, we can see that BERTScore Recall performs the best among all, with SMS (GloVe), SMS (ELMo) and BERTScore F1 score close by. BLEU shows least correlation among all.

BLEU's correlation is the highest in the ASAP AES dataset. The reasoning can be that, the essay dataset is made from a question in the class 10 exam for students, where they had to write the answer to a reading comprehension question. Since the passage was already available, and the students had to answer the questions from the passage, a lot of words from students' answers could match with the passage. So the BLEU metric works spectacularly in this case. But it does not generalize well on other datasets, which is a bad indicator for any metric.

Sentence-BERT performs the best in ASAP SAS dataset. BERTScore Recall and SMS (GloVe) also perform well in this test.

Metrics like BERTScore Recall, SMS (GloVe) and SMS (ELMo) are performing consistently well on all the datasets tried out. This shows that they generalize better.

The plots in Figure 1 show the histogram plots of scores for the best and the worst performing metrics for the ASAP AES dataset. It shows that the metric which is most highly correlated (BLEU), has got maximum separation between the mean scores for each human score (either 0, 1, 2 or 3), and the metric with the minimum correlation (BERTScore P) with human scores has a lot of overlap between scores.



(a) Scores for BLEU (b) Scores for BERTScore P

Figure 1. Histogram plots of scores for different human scores

5. Conclusion

A review of various evaluation metrics were carried out, and the performance of different embeddings for encoding words and sentences were verified. Embeddings from various pre-trained transformer models like BERT, Sentence-BERT and RoBERTa-Large were also used in the experiments. It was seen that these did not provide much improvement when compared to existing type based embeddings like GloVe.

We see reasonably high correlations for metrics when compared to human scores. But, we cannot automate the process of evaluation of essays because we can see that the best correlated scores also show a variation in the scores assigned for the same human score. We reviewed metrics based on contextual embeddings and type embeddings. We observe that there is no significant difference between them.

Supervised methods of learning still hold the upper hand on automatic scoring of essays.

We observe that sentence based embeddings when incorporated give better correlations with human scores than word embeddings only, most of the time. Hence sentence based embeddings are in general more useful.

We also use fundamentally different methods of finding similarity i.e. cosine similarity and word mover distance. We use the same pre-trained model embedding for both and find cosine similarity is more correlated to human scores in all three datasets.

References

Chaganty, A., Mussmann, S., and Liang, P. The price of debiasing automatic metrics in natural language evaluation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 643–653, Melbourne, Australia, July 2018. Association for Computational Linguistics.

doi: 10.18653/v1/P18-1060. URL https://www.aclweb.org/anthology/P18-1060.

Clark, E., Celikyilmaz, A., and Smith, N. A. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2748–2760, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1264. URL https://www.aclweb.org/anthology/P19-1264.

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 199–209, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1019.

Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 957–966. JMLR.org, 2015.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://www.aclweb.org/anthology/D19-1410.

Rubner, Y., Tomasi, C., and Guibas, L. J. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 59–66, 1998.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019. URL http://arxiv.org/abs/1904.09675.