eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings

Muhammad ElNokrashy
Microsoft
Cairo, Egypt
muelnokr@microsoft.com

Tom Kocmi
Microsoft
Munich, Germany
tomkocmi@microsoft.com

Abstract

We propose eBLEU, a metric inspired by BLEU metric that uses embedding similarities instead of string matches. We introduce *meaning diffusion vectors* to enable matching n-grams of semantically similar words in a BLEU-like algorithm, using efficient, non-contextual word embeddings like fastText. On WMT23 data, eBLEU beats BLEU and ChrF by around 3.8% system-level score, approaching BERTScore at -0.9% absolute difference. In WMT22 scenarios, eBLEU outperforms f101spBLEU and ChrF in MQM by $2.2\%-3.6\%$. Curiously, on MTurk evaluations, eBLEU surpasses past methods by $3.9\%-8.2\%$ (f200spBLEU, COMET-22). eBLEU presents an interesting middle-ground between traditional metrics and pretrained metrics.

1 Introduction

The machine translation field has improved significantly, with various metrics developed to measure translation quality. Translation quality in human eyes is usually a delicate balance to convey meaning, style, tone, and other dimensions of text from one language into another with different idioms and concept ontologies. After all, translation is not only about translating words from one language to another literally, but ensuring that the core meaning behind is also accurately conveyed.

Traditional metrics, like the **BLEU** score (Papineni et al., 2002) or ChrF (Popović, 2015), have proven effective over last 20 year. However, there has been growing evidence that they have not kept pace with the performance of recent NMT and LLM MT systems (Kocmi et al., 2021; Freitag et al., 2022). BLEU essentially computes a score based on string n-grams matches. One clear limitation of this approach is that it fails to recognize semantically similar words. For instance, in the eyes of BLEU, the words (*cat*, *kitty*) are as different as (*cat*, *book*) or (*fire*, *water*).

Recent neural metrics, on the other hand, have explored the potential of leveraging pretrained language models for encoding entire sentences. These models either compare encoded sentences in a shared embedding space or employ a trained classifier to predict human scores, as demonstrated by Rei et al. (2020); Zhang et al. (2020); Freitag et al. (2022). These methods are more capable of capturing semantic nuances.

In this paper, we introduce **eBLEU**, a metric designed to address the mentioned limitation of the BLEU score while keeping the calculation as close to BLEU as possible by using the word embedding similarities instead of string matching. By doing so, eBLEU enhances the metric by recognizing semantically similar n-grams. Our method relies on the *meaning diffusion map* to approximate n-gram matching in a BLEU-like algorithm. The core implementation leverages efficient, non-contextual word embeddings, such as fastText embeddings.

2 Related work

In machine translation, measuring quality is a balance of many potentially competing factors. The most prominent are language quality (fluency) and accuracy of meaning conveyed (adequacy). Other factors may be critical in special scenarios. Consider the conveyance of tone or cultural register in translated dialog (see for example registers in East-Asian languages). Or the conveyance of flow in a translated play (see some examples of translations of the Greek epic Iliad in Mendelsohn, 2011).

Traditional automatic quality assessment methods, like BLEU and METEOR (Banerjee and Lavie, 2005), rely on string matching against a reference. The more matches, the more a candidate captures of the intended meaning in the reference, as proxy for adequacy. While features like n-gram matching in BLEU and explicit ordering penalties in METEOR act as proxy for fluency.

Such metrics suffer from limitations inherent to

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 746–750
December 6–7, 2023.

their literal string matching core, which some try to mitigate (e.g. via lemmatization or synonym dictionaries). These limitations are clearer in light of the more complex and semantically rich language produced by recent Neural MT systems and Large Language Model MT systems.

BERTScore (Zhang et al., 2020) utilizes a similar idea to ours, matching contextual encodings of words in candidate/reference pairs. While it uses unigrams only, eBLEU uses n-grams as well, and calculates token matches differently. Other systems, like COMET (Rei et al., 2020), are finetuned on human judgement scores for machine translation evaluation specifically.

3 Preliminaries

3.1 BLEU

The BLEU formula applied to a single candidate/reference sentence pair X, Y is:

$$\underset{N}{\texttt{BLEU}}(X,Y) = \texttt{bp}(X,Y) \prod_{n \in 1..N} \texttt{pr}_n(X,Y)^{w_n} \tag{1}$$

where $\texttt{bp}(X,Y)$ is the brevity penalty. This score ranges between $[0,1]$ for lowest and highest match.

The n-gram precision $\texttt{pr}_n(X,Y)$ is:

$$\frac{\sum_{s \in [X]^n} \min(C(s,X), C(s,Y))}{\sum_{s \in [X]^n} C(s,X)} \tag{2}$$

Set of n-gram substrings in candidate

Count of s in candidate

3.2 Embeddings

At the core, our method utilizes simple word embeddings that can be generated from sub-word information or memorized for full words as appropriate. We do not require tokenization of words into sub-words. Here we use the fastText word embeddings (Bojanowski et al., 2017). Other simple word embeddings should be appropriate as-is but were not tested. FastText is trained for every language separately and we require a trained fastText model for the target language in any translation pair.

3.3 String Matching

Strings under strict equality are literal representations of unique identities: the string *abc* is equal only to *abc* itself. This works for BLEU. Now we want to match based on the closeness of meaning instead, where (*cat*, *cats*) would be closer together than (*cat*, *book*).

4 eBLEU description

We propose the following formulation for an embedding-based matching in the style of BLEU precision from eq. (2).

Let X refer to the candidate sentence, and Y refer to the reference sentence. Now, given an asymmetric similarity function $\texttt{mdSim}(a \mid b)$ from a with reference to b, we can define the following analogous values for "precision" and "recall":

$$\text{precision: } \texttt{pr}(X,Y) = \texttt{mdSim}(Y \mid X) \tag{3}$$

$$\text{recall: } \texttt{re}(X,Y) = \texttt{mdSim}(X \mid Y) \tag{4}$$

$\overline{m}_n$ is the n-gram score of the pair, defined as the geometric mean of the n-gram precision and recall from eq. (3, 4):

$$\overline{m}_n = \left(\texttt{pr}_n(X,Y) \cdot \texttt{re}_n(X,Y)\right)^{\frac{1}{2}} \tag{5}$$

The final score is a weighted geometric average of the n-gram-based scores $\overline{m}_n$ between candidate and reference, for $N = 4$ and $w_n = N - n$.

$$\underset{N}{\texttt{eBLEU}}(X,Y) = \texttt{lp}(X,Y) \prod_{n \in 1..N} \overline{m}_n^{\,w_n/N} \tag{6}$$

where $\texttt{lp}$ is a modified length penalty which penalizes longer candidates as well.

(e)BLEU This shows the analogous structure of eBLEU compared to BLEU, given an appropriate definition for $\texttt{mdSim}$ as used in eq. (3, 4).

4.1 Aggregating Similarity Values

Similar to $\texttt{pr}_n(X,Y)$, we want $\texttt{mdSim}(Y \mid X)$ to be a single value for a candidate/reference sentence pair, as if aggregating the meaning diffusion values m_x for $x \in X$:

$$\texttt{mdSim}(Y \mid X) = \frac{\sum_x \min(\mathbf{m}_X, \mathbf{m}_{X|Y})}{\sum_x \mathbf{m}_X} \tag{7}$$

Index over words in cand.

MD vector of words in cand.

Compare Equation (7) for eBLEU with Equation (2) for BLEU.

4.2 Meaning Diffusion

Meaning Diffusion (MD) is a value for each word in a sentence indicating *the ratio of similar words*

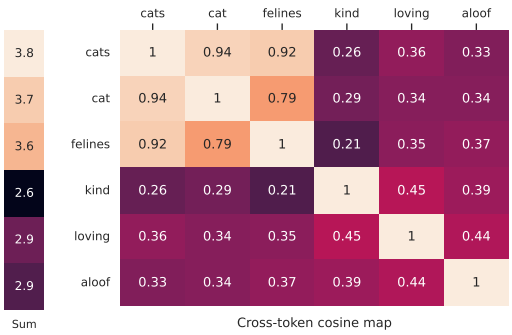

Figure 1: *(Right)* Meaning Diffusion Map between some words for illustration. Notice the similarity of *cats, cat, felines* and the relative similarity of *kind, loving* but not *aloof*. *(Left)* Meaning Diffusion Vector is the sum of a word's similarity to all words in the sentence.

in the same sentence. This allows claims such as "*there exists* $2/7$ `eats` " in:

The cat `eats` *, no,* `devours` *the food.*

That is: $m_{\text{eats}} \approx 2/7$. See also Figure 1 *(Left)*.

MD Map $\hat{\mathbf{S}}_{y,y'}$ is a weighted sum over the candidate side (x) with `softmax`-normalized weights. It approximates the similarity matrix of the reference Y against itself (see Figure 1 *(Right)*), as seen *through* the candidate X. The **L1** variant replaces `softmax` with a simple division by the sum of values: $\mathbf{S}_{y,x} / \sum_x \mathbf{S}_{y,x}$.

MD Vectors $\mathbf{m}_*$ represent each word's total closeness to all other words in the same sentence ($\mathbf{m}_Y$) or through a candidate sentence ($\mathbf{m}_{Y|X}$).

$$\hat{\mathbf{S}}_{y,y'} = \underset{x}{\texttt{softmax}}(\mathbf{S}_{y,x}) \cdot \mathbf{S}_{x,y} \tag{8}$$

$$\mathbf{m}_{Y|X} = \sum_{y'} \hat{\mathbf{S}}_{y,y'} \tag{9}$$

$$\mathbf{m}_Y = \sum_{y} \mathbf{S}_{y,y} \tag{10}$$

Vector Similarity For the candidate/reference X, Y, let $\mathbf{X}, \mathbf{Y}$ be the embedding matrices shaped as *token* $\times$ *embedding*. $\mathbf{S}_{x,y}$ is Cosine vector similarity clipped within $[0, 1]$, defined as:

$$\mathbf{S}_{x,y} = \texttt{clip}_{[0,1]} \cos_{\textit{embedding}}(\mathbf{X}, \mathbf{Y}^\top) \tag{11}$$

4.3 N-gram Scores

For each $n \in 1..N$, we calculate the n-gram score of a sentence pair using $\mathbf{S}^n_{x,y}$: the geometric mean of the cosine scores of adjacent words in each sentence, such that the n-gram-aware $\mathbf{S}^n_{x,y}$ is of shape $|X| - n + 1 \times |Y| - n + 1$.

4.4 Length Penalty

The length penalty penalizes length mismatch between candidate and reference, as used in eq. (6):

$$\texttt{lp}(X, Y) = \begin{cases} 1.0 & \texttt{ratio} \leq 0.5 \\ e^{0.5 - \texttt{ratio}} & \text{else} \end{cases} \tag{12}$$

$$\texttt{ratio} = \frac{\texttt{abs}(|X| - |Y|)}{|Y|} \tag{13}$$

5 Evaluation and Results

In this section, we describe the evaluation of the metric and the results

5.1 Meta-evaluation

We use the WMT Metrics 2022 test set (Freitag et al., 2021) which contains human judgments based on three different protocols: MQM, DA+SQM and MTurk DA. The translation systems are mainly from participants of the WMT22 General MT shared task (Kocmi et al., 2022). The source segments and human reference translations for each language pair contain around 2,000 sentences from four different texts domains: news, social, conversational, and e-commerce.

Human labels are produced via three methods:

- MQM - annotated by professionals who mark individual errors in each translation, as described in (Freitag et al., 2021)
- DA+SQM - professional annotators are asked to rate each translation on a scale 0-100 (Kocmi et al., 2022)
- MTurk DA - low paid crowd of MTurk annotators is asked to rate each translation on a scale 0-100, for how much it resembles human reference (Kocmi et al., 2022)

To determine the correlation of automatic metrics with humans, we measure system-level, pairwise accuracy (Kocmi et al., 2021), which is defined as the number of system pairs ranked correctly by the metric with respect to the human ranking divided by the total number of system pair comparisons. Formally:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) == \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

We reproduced scores reported in the WMT22 Metrics shared task findings paper with the official

System	MQM	DA-SQM	MTurk
COMET-22	**83.94%**	**84.19%**	62.61%
COMET-20	83.58%	82.17%	63.53%
BERTScore	77.37%	75.92%	66.57%
f101spBLEU	**74.45%**	74.26%	65.96%
f200spBLEU	74.09%	74.26%	**66.87%**
chrF	73.36%	**75.92%**	66.57%
BLEU	70.80%	70.22%	65.35%
eBLEU-FastText			
↪ L1	**76.64%**	74.45%	68.69%
↪ Softmax*	74.82%	72.43%	**70.82%**

Table 1: System-level WMT22 results on 3 human labeling scenarios. The Softmax variant of eBLEU was submitted to WMT23.

System	Rank	Score
COMET	2	**93.5%**
BERTScore	7	90.2%
f200spBLEU	11	**86.8%**
BLEU	12	85.9%
ChrF	12	85.2%
eBLEU-FastText		
↪ Softmax	9	**89.3%**

Table 2: WMT23 System-level ranking clusters and correlations on en-de, he-en, zh-en language pairs.[2]

WMT22 script.[1] Scores match Table 8 (DA+SQM and DA) and Table 11 for MQM of the WMT22 Metrics findings paper (Freitag et al., 2022).

5.2 Results

On **WMT23** scenarios (Table 2), eBLEU scores $\mathbf{89.3}\%$, improving noticeably on ChrF, BLEU, and f200spBLEU, beating the latter by $\mathbf{2.5}\%$ points. Its ranking cluster (**9**) puts it much closer to more sophisticated embedding-based metrics (like BERTScore) than string metrics like BLEU. Notably, this was achieved by the Softmax variant, which scored below the L1 variant on the more accurate human MQM and DA-SQM scenarios.

On **WMT22** scenarios (Table 1), eBLEU outperforms both f101spBLEU and ChrF in MQM by $2.2\% - 3.6\%$ in system-level accuracy.

eBLEU shows SOTA correlation with MTurk evaluations at $\mathbf{70.82}\%$, beating existing methods by $3.9\% - 8.2\%$ (f200spBLEU, COMET-22). Although Freitag et al. (2022) shows them to be of sub-optimal quality, this is interesting as MTurk evaluations often involve manual n-gram matching—a nice result given the intuition behind our method.

6 Conclusion

6.1 eBLEU: Between Strings and Neural Eval

In this paper, we introduced eBLEU, a novel metric that adapts the BLEU algorithm by adding embedding-based semantic understanding. By incorporating word embedding similarities and leveraging *meaning diffusion vectors*, eBLEU bridges the gap between literal and semantic matching.

We show that eBLEU can outperform widely adopted metrics like (sp)BLEU and ChrF, and approaches some pretrained contextual embedding-based metrics, like BERTScore, using simpler, cheaper-to-compute embeddings like fastText.

On WMT23, eBLEU scores 89.3%, placing almost halfway between BLEU, and COMET, an especially finetuned model for MT evaluation.

Although eBLEU lags behind the latest pretrained metrics, it presents an interesting approach for a simple semantically informed metric.

6.2 Limitations

However, it is important to recognize the limitations. Fundamentally, eBLEU does not attempt to improve the BLEU formula as a proxy for adequacy and fluency. Thus predictably, it lags far behind the latest pretrained metrics such as COMET or BLEURT. As language models, the core of these systems holds the advantages of large pre-training data, contextual understanding of input candidates and references, and potentially task-specific fine-tuning for the translation domain. Their more general nature allows for much improved measurement of adequacy and fluency among the range of possible translations that humans may produce and judge acceptable.

In summary, eBLEU offers a semantically-aware machine translation evaluation metric extending standardized BLEU algorithm. There may exist other such methods that bridge the gap further while improving inference time, efficiency, or interpretability where needed.

[1] https://github.com/google-research/mt-metrics-eval
[2] As provided by the WMT team.

References

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. Association for Computational Linguistics. [Cited on page 1.]

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. [Cited on page 2.]

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. [Cited on page 3.]

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. [Cited on pages 1 and 4.]

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. [Cited on page 3.]

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics. [Cited on pages 1 and 3.]

Daniel Mendelsohn. 2011. Englishing the iliad: Grading four rival translations. [Cited on page 1.]

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. [Cited on page 1.]

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. [Cited on page 1.]

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. [Cited on pages 1 and 2.]

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. [Cited on pages 1 and 2.]