

Inside Sentence Embeddings: Leveraging Token Representations for Machine Translation Evaluation and Word Alignment

Mana Makinae, Takumi Goto, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology (NAIST)

{makinae.mana.mh2, goto.takumi.gv7, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

Contextualized token-level embeddings from encoders, e.g., mBERT, are used for several cross-lingual tasks, e.g., machine translation (MT) evaluation and alignment tasks, but fine-tuning is required for the specific task settings. In contrast, multilingual sentence embeddings are trained under sentence-level objectives on parallel data, but their token-level representations remain unclear whether they carry cross-lingual semantics, although not explicitly trained with token-level objectives. In this paper, we investigate token representations extracted from multilingual sentence embedding models for MT evaluation and word alignment. Experimental results and analysis show that sentence embedding models can provide useful token-level representations for direct similarity-based tasks. Although the observed advantage is not consistent across all settings, the results indicate that multilingual sentence-level training may support token-level semantic similarity without explicit token-level training.

1 Introduction

Natural language processing has shifted from surface-level matching to contextual embedding-based similarity. This shift has also influenced machine translation (MT) evaluation and word alignment. In MT evaluation, embedding-based metrics such as BERTScore (Zhang et al., 2020) compute semantic similarity between candidate and reference translations using contextualized token representations. In word alignment, methods such as SimAlign (Jalili Sabet et al., 2020) and AwesomeAlign (Dou and Neubig, 2021) use multilingual contextual embeddings to identify corresponding words across languages. Because contextualized embedding models (Devlin et al., 2019) are trained with token-level objectives such as masked language modeling, their token representations are expected to encode rich information.

Meanwhile, multilingual sentence embedding models have made progress. Models such as LaBSE (Feng et al., 2022), multilingual E5 (Wang et al., 2024) are trained to produce sentence-level representations that perform well in multilingual retrieval and semantic similarity tasks. Although these models are not explicitly trained for token-level objectives, their sentence-level training may still induce token representations that support cross-lingual token matching. Such representations could be useful for similarity-based tasks such as MT evaluation and word alignment, without task-specific training.

In this paper, we investigate whether token representations extracted from multilingual sentence embedding models are useful for two token-level similarity tasks: MT evaluation and word alignment. For MT evaluation, we use these representations in a BERTScore-style framework in place of standard contextualized token representations. For word alignment, we apply a simple mutual nearest-neighbor method based on cosine similarity between source and target token representations. Both settings directly use pretrained sentence embedding models and require no task-specific fine-tuning or alignment training.

Experimental results and analysis show that token-level representations extracted from sentence embedding models can be used for direct similarity-based tasks. While their effectiveness varies across settings, the results suggest that multilingual sentence-level training may also learn token representations suitable for token-level semantic similarity tasks. A layer-wise analysis reveals different tendencies across tasks: higher layers are effective for MT evaluation, while intermediate layers are suited to word alignment. This pattern is consistent between contextualized embedding models and sentence embedding models, suggesting that sentence embedding models also preserve both abstract semantic information and token-level

correspondence information across layers despite the absence of explicit token-level training.

2 Related Work

Word and Sentence Embeddings Contextualized encoder models such as BERT (Devlin et al., 2019) produce context-dependent token representations, improving many semantic similarity tasks, such as general language understanding and question answering. Sentence embedding models have also advanced in recent years (Feng et al., 2022; Wang et al., 2024; Chen et al., 2024). These models are continually trained to produce semantically meaningful sentence-level embeddings using contrastive learning from pretrained encoder models and achieve strong performance on multilingual semantic similarity and retrieval tasks. Although they are optimized for sentence-level semantics, prior work has shown that multilingual sentence encoders can be adapted into cross-lingual lexical encoders using only a small bilingual dictionary (Vulić et al., 2023). Other work has also demonstrated that sentence embedding models can capture semantic changes within sentences by representing edits in sentence embedding space (Goto et al., 2026).

MT Evaluation With the rise of multilingual pretrained language models, contextualized embedding evaluation methods have also been proposed. COMET (Rei et al., 2020) utilizes cross-lingual pretrained language models to predict human judgment scores by incorporating source, reference, and system outputs. COMET-QE (Rei et al., 2021) is an extension of COMET to reference-free evaluation by eliminating references and directly evaluates the similarity between the source and generated sentences by predicting human evaluation score. BERTScore (Zhang et al., 2020) compute token-level semantic similarity directly using pretrained language models such as BERT and multilingual BERT (mBERT) (Devlin et al., 2019).

Word Alignment Contextualized embedding methods have been proposed along with the pretrained language models. SimAlign (Jalili Sabet et al., 2020) finds word alignment using multilingual contextualized embeddings and achieves strong performance without explicit alignment supervision or parallel training data. AwesomeAlign (Dou and Neubig, 2021) further improves embedding-based alignment by fine-tuning

multilingual pretrained language models using alignment-oriented objectives and proposes alignment extraction methods based on this method, extract alignments from these fine-tuned models.

3 Proposed Method

In this work, we investigate the use of token representations extracted from multilingual sentence embedding models for MT evaluation and word alignment. Instead of using the final aggregated sentence embedding, we utilize contextualized token representations extracted from intermediate hidden layers. Our approach directly utilizes pretrained sentence embedding models without training or parallel data.

Token Representation Extraction Given a parallel sentence pair $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ and $\mathbf{y} = \{y_1, y_2, \dots, y_m\}$, we extract token representations from the i -th layer of the sentence embedding model:

$$\mathbf{X}^{(i)} = (\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_n^{(i)}), \mathbf{Y}^{(i)} = (\mathbf{y}_1^{(i)}, \dots, \mathbf{y}_m^{(i)}). \quad (1)$$

where $\mathbf{x}_j^{(i)}$ and $\mathbf{y}_k^{(i)}$ denote contextualized token representations of the j -th and k -th subword tokens at layer i , respectively. Each representation is a d -dimensional vector: $\mathbf{x}_j^{(i)}, \mathbf{y}_k^{(i)} \in \mathbb{R}^d$.

MT Evaluation For MT evaluation, we incorporate token representations extracted from sentence embedding models into similarity-based evaluation in the same manner as BERTScore (Zhang et al., 2020). Specifically, contextualized token representations from sentence embedding models are used in place of token representations derived from pretrained encoder models such as BERT (Devlin et al., 2019). Given a candidate translation $\mathbf{x}$ and a reference translation $\mathbf{y}$, we extract token representation sequences $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(i)}$ from the i -th layer, as described above. Token-level semantic similarity is computed using cosine similarity:

$$s^{(i)}(j, k) = \cos(\mathbf{x}_j^{(i)}, \mathbf{y}_k^{(i)}). \quad (2)$$

Following BERTScore (Zhang et al., 2020), each token is matched with the most similar token in the opposite sentence. Precision is computed as:

$$P_{\text{BERT}} = \frac{1}{|\mathbf{x}|} \sum_{x_j \in \mathbf{x}} \max_{y_k \in \mathbf{y}} s^{(i)}(j, k), \quad (3)$$

where $s^{(i)}(j, k)$ denotes the cosine similarity between the j -th candidate token representation and

the k -th reference token representation. Recall is computed symmetrically, and F1 is obtained from precision and recall.

Word Alignment Based on token representations extracted from sentence embedding models, we first compute a cosine similarity matrix between source and target subwords. We then construct alignments using mutual nearest-neighbor matching, also known as mutual argmax matching in similarity-based word alignment (Jalili Sabet et al., 2020). Concretely, we apply L2 normalization to the extracted token representations and compute cosine similarity between source and target subwords defined in Equation (2). A source subword $x_j^{(i)}$ and target subword $y_k^{(i)}$ are aligned when they are mutual nearest neighbors in the embedding space:

$$k = \arg \max_{1 \leq t \leq m} s^{(i)}(j, t), j = \arg \max_{1 \leq u \leq n} s^{(i)}(u, k) \quad (4)$$

Here, t ranges over all target subwords and u ranges over all source subwords. A source subword $x_j^{(i)}$ and target subword $y_k^{(i)}$ are aligned only when they are the most similar tokens to each other in both directions. Finally, subword-level alignments are merged into word-level alignments.

4 Experimental Settings

Embedding Models We use the following multilingual encoder models as baselines: mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), and the following multilingual sentence embedding models: (i) LaBSE (Feng et al., 2022), which is based on mBERT and trained for sentence embeddings; (ii) E5 (Wang et al., 2024), and (iii) BGE-M3 (Chen et al., 2024)¹. E5 and BGE-M3 are based on XLM-R and further trained with contrastive learning. For MT evaluation, we first perform layer selection on the development set and then evaluate the selected layer on the test set. For word alignment, we analyze all hidden layers and report the best-performing layer as an oracle measure of token-level representation quality.

MT Evaluation MT evaluation experiments use data from the WMT Metrics Shared Tasks. For layer selection, we use WMT 2022 (Freitag et al., 2022) as the development set and WMT 2021 as the test set. WMT 2022 provides MQM annotations for

en-de, en-ru, and zh-en, as well as DA scores for {cs, ja, ru, de, uk}-en. Automatic metrics are evaluated at the system level using Pearson correlation with human judgments. For each model, we select the layer with the highest mean Pearson correlation across the three WMT 2022 MQM language pairs: en-de, en-ru, and zh-en. The selected layer and its corresponding mean Pearson correlation are also reported. We then fix the selected layer and evaluate the model on the WMT 2021 News MQM dataset, which contains the same three language pairs, without any further layer selection or tuning. Therefore, we report the resulting system-level Pearson correlations on WMT 2021.

Word Alignment We use the fine-tuned Awesome-Align (Dou and Neubig, 2021) as a supervised baseline. For all other models, we extract source and target subword representations from each model and apply mutual nearest-neighbor matching, where a source-target subword pair is aligned only when they are each other’s nearest neighbors under cosine similarity. We report Alignment Error Rate (AER) following Dou and Neubig (2021). We evaluate performance on five language pairs: de-en, fr-en, ro-en, ja-en, and zh-en, following Dou and Neubig (2021). Because our main setting does not use task-specific training, we directly evaluate the models on benchmark datasets. We use the de-en dataset from Vilar et al. (2006), the fr-en and ro-en datasets from Mihalcea and Pedersen (2003), the ja-en dataset based on the Kyoto Free Translation Task word alignment corpus (Neubig, 2011), and the zh-en dataset from the TsinghuaAligner benchmark (Liu and Sun, 2015). To account for sampling variation in the evaluation data (Koehn, 2004), we use bootstrap resampling over all annotated sentence pairs for each language pair: 508 for de-en, 447 for fr-en, 248 for ro-en, 582 for ja-en, and 450 for zh-en. For each language pair, we resample sentence pairs with replacement, compute AER on each resampled dataset, and repeat this procedure 1,000 times. We then compute 95% bootstrap confidence intervals for AER. These intervals are used to assess the robustness of observed performance differences between models.

5 Results

5.1 MT Evaluation

Layer Selection Results Table 1 shows the Pearson correlation with MQM on the development set

¹Although BGE-M3 have been jointly trained on sentence-level and token-level, we also included as representative of sentence embedding model.

Model ($\uparrow$)	Layer	en-de	en-ru	zh-en
COMET-22	-	0.771	0.900	0.942
COMET-20	-	0.876	0.936	0.970
BS(mBERT)	12	0.460	0.822	0.841
BS(XLM-R)	10	0.489	0.864	0.830
BS(LaBSE)	12	0.608	0.746	0.927
BS(E5)	12	0.611	0.783	0.936
BS(BGE-M3)	24	0.648	0.815	0.955

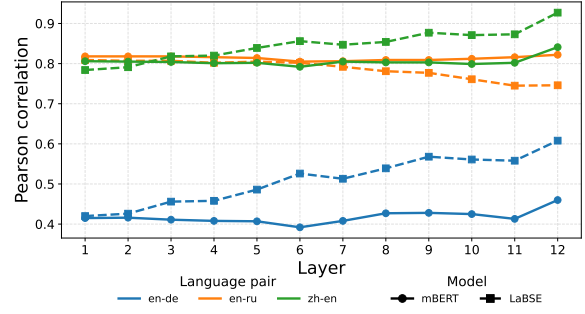
Table 1: Pearson correlation coefficients with human evaluation on the WMT 2022 as the development set. BS denotes under BERTScore calculation framework.

Model ($\uparrow$)	Layer	en-de	en-ru	zh-en
COMET-22	-	0.771	0.900	0.942
COMET-20	-	0.876	0.936	0.970
BS(mBERT)	12	0.907	0.657	0.500
BS(XLM-R)	10	0.809	0.388	0.441
BS(LaBSE)	12	0.805	0.544	0.515
BS(E5)	12	0.916	0.721	0.523
BS(BGE-M3)	24	0.796	0.605	0.498

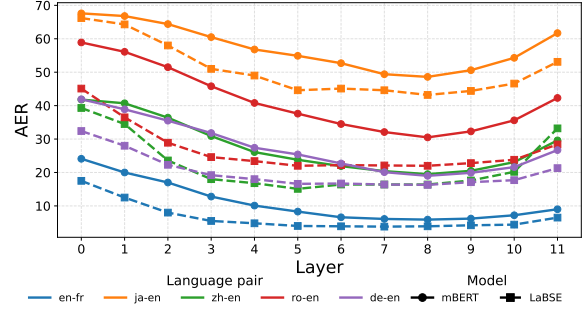
Table 2: Pearson correlation coefficients with human evaluation on the WMT 2021 as the test set. BS denotes under BERTScore calculation framework.

across different multilingual embedding models under the same BERTScore calculation framework. Results with DA are provided in Appendix A. Because the scoring architecture is fixed across all settings, the performance differences reflect the quality of the token representations. COMET is included as a reference for the supervised setting. Sentence embedding models show strong performance on en-de and zh-en, whereas the contextualized encoder baselines remain stronger on en-ru. In particular, BGE-M3 achieves the highest correlation among the BERTScore-based models on zh-en and performs close to COMET, while E5 and LaBSE also perform strongly on en-de.

Test Set Results Next, we evaluate the layers selected on the development set using the held-out test set. Table 2 reports the Pearson correlation with MQM under the same BERTScore calculation framework. On the test set, E5 achieves the highest correlation among the BERTScore-based models across all three language pairs. In contrast, LaBSE does not retain the same relative advantage observed during development set layer selection. Although the performance advantage of sentence embedding models is not consistent across all settings, the test results suggest that they can nevertheless provide useful token-level representations



(a) MT evaluation task ($\uparrow$).



(b) Word alignment task ($\downarrow$).

Figure 1: Layer-wise performance for mBERT and LaBSE. Higher is better in MT evaluation task and lower is better in word alignment task.

for direct similarity-based MT evaluation. Results with DA under oracle layer selection are reported separately in Appendix A.

5.2 Word Alignment

Table 3 shows AER across different multilingual embedding models. Our main comparison is among models without alignment-specific fine-tuning. The fine-tuned Awesome-Align result is included as a task-specific fine-tuned reference rather than as a directly comparable baseline. Under unsupervised setting, LaBSE shows the strongest performance across language pairs, outperforming Awesome-Align without fine-tuned setting. This is consistent with the sentence-level retrieval analysis in Appendix 6, where LaBSE also shows the strongest cross-lingual retrieval performance among the sentence embedding models. These results suggest that, despite its simplicity, mutual nearest-neighbor matching achieves strong alignment performance when combined with strong multilingual sentence embedding representations.

6 Analysis

Layer-wise Performance We analyze how token representations from different layers affect MT

Model (↓)	Layer	de-en	fr-en	ro-en	ja-en	zh-en
Awesome Align (w/ ft)	-	15.2	4.1	22.6	37.4	13.4
Awesome Align (w/o ft)	-	17.4 [16.4, 18.4]	5.59 [4.98, 6.23]	27.9 [25.9, 30.0]	45.6 [44.1, 47.0]	17.9 [17.1, 18.7]
mBERT	8	20.1 [19.1, 21.1]	6.1 [5.5, 6.8]	32.1 [30.0, 34.2]	49.4 [47.9, 50.9]	20.4 [19.6, 21.3]
XLM-R	8	19.5 [18.5, 20.6]	6.7 [5.9, 7.4]	25.9 [24.0, 28.1]	59.1 [57.6, 60.4]	25.3 [24.3, 26.3]
LaBSE	8	16.4 [15.5, 17.3]	3.8 [3.3, 4.3]	22.1 [20.2, 24.2]	44.6 [43.4, 45.9]	16.4 [15.6, 17.2]
E5	7	19.5 [18.4, 20.5]	5.2 [4.6, 5.9]	26.3 [24.3, 28.6]	55.8 [54.4, 57.2]	23.6 [22.7, 24.4]
BGE-M3	13	20.1 [19.1, 21.1]	5.8 [5.2, 6.4]	26.0 [24.0, 28.1]	60.4 [59.1, 61.7]	25.1 [24.3, 25.8]

Table 3: AER across multilingual embedding models. Sentence embedding models are used with mutual nearest neighbor matching. Values in brackets denote 95% bootstrap confidence intervals.

Model (↑)	en-de	en-ru	zh-en
LaBSE	0.914	0.802	0.939
E5	0.890	0.774	0.922
BGE-M3	0.904	0.788	0.936

Table 4: Sentence-level retrieval results across multilingual sentence embedding models.

evaluation and word alignment. Figure 1 shows the layer-wise results for mBERT and LaBSE. Overall, token representations extracted from LaBSE outperform those from mBERT across most layers in both tasks. For MT evaluation, higher layers achieve better performance, with the final layer obtaining the best or near-best correlation across language pairs. For word alignment, intermediate layers perform better than the final layer, with the best alignment performance observed around the middle layers. These layer-wise tendencies are consistent between mBERT and LaBSE. This suggests that sentence embedding models can encode token representations effective for token-level semantic similarity tasks, similar to contextualized encoder models, despite being optimized for sentence-level objectives rather than explicit token-level training. However, even with sentence embedding models, the challenge of requiring layer selection remains. The most useful layer depends on whether the task requires sentence-level semantic similarity or local token correspondence, indicating that practical use of these representations would require layer selection based on development data rather than selecting layers directly on the evaluation set.

Cross-Lingual Sentence Retrieval To examine whether the observed word alignment performance is specific to word alignment or reflects a broader limitation of multilingual sentence embedding models, we additionally conduct a sentence-level retrieval analysis using the Tatoeba MT

dataset (Tiedemann, 2020). Given a source sentence, we retrieve its corresponding target-language translation from a pool of candidate sentences using cosine similarity between sentence embeddings. Table 4 reports retrieval accuracy for each multilingual sentence embedding model, measuring whether the correct translation is ranked as the most similar sentence. Among the evaluated sentence embedding models, LaBSE achieves the strongest cross-lingual retrieval performance. This tendency is consistent with the word alignment results, suggesting that stronger sentence-level cross-lingual semantic alignment may also contribute to better token-level alignment quality.

7 Conclusion

In this paper, we investigated whether multilingual sentence embedding models contain useful token-level representations for MT evaluation and word alignment. Experimental results and analysis show that, despite variation across settings, multilingual sentence-level training may produce useful token representations for direct token-level semantic similarity evaluation without explicit token-level training. These findings motivate future work on how multilingual sentence-level training supports token-level semantic representation and which training factors contribute to this capability.

8 Limitation

This study has several limitations. First, we focus on two token-level similarity tasks, MT evaluation and word alignment. Although these tasks provide complementary evidence, our findings may not directly generalize to other token-level tasks, such as syntactic analysis, morphological analysis, or token classification.

Second, our experiments cover a limited set of language pairs, and the results show clear varia-

tion across languages. In particular, the behavior observed for en–ru in MT evaluation suggests that the usefulness of token-level similarity may depend on language-specific factors, such as tokenization, morphology, and contextual token representations. The observed differences may instead be related to token-level matching or aggregation rather than sentence-level embedding quality. Future work should examine a broader range of language pairs and more carefully separate sentence-level and token-level effects.

Third, our methods are limited to direct similarity-based approaches without task-specific fine-tuning. This design helps isolate the effect of the representations themselves, but it does not show whether sentence embedding models remain competitive when used with supervised metrics or alignment-specific training.

Finally, our layer selection protocol is limited. We report the best-performing layer for each setting, which should be interpreted as an oracle analysis of representation quality rather than a practical model-selection procedure. This choice helps us examine whether sentence embedding models contain useful token-level information without underestimating their representational capacity. However, it may overestimate performance in practical use. In a realistic setting, the layer should be selected using development data or a fixed layer-selection strategy before evaluation. This is particularly important because our layer-wise analysis shows that the most useful layer differs by task: higher layers are more effective for MT evaluation, while intermediate layers are more effective for word alignment.

9 Ethical Consideration

Dataset Usage and Licenses We use publicly available benchmark datasets for MT evaluation, word alignment, and sentence-level retrieval. We follow the intended research use of these datasets and refer users to the original dataset releases for license and usage conditions. Our work does not use new human annotations or collect personal data.

Harmful Content Because the experiments are conducted on existing public datasets, the data may contain harmful, biased, or offensive content inherited from the original sources. Our work does not aim to generate or amplify such content; instead, it evaluates representation quality using automatic metrics and benchmark annotations. Nevertheless, model representations and evaluation results may

reflect biases present in the underlying datasets and pretrained models.

AI-assistant Usage We used an AI assistant to support parts of the writing process, such as improving wording and linguistic clarity. All technical content, experimental design, analysis scripts, interpretation of results, and final decisions were made by the authors.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work was partly supported by JST SPRING Grant Number JPMJSP2140 and JSPS KAKENHI Grant Numbers JP25K24369, JP26K21312.

References

- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zi-Yi Dou and Graham Neubig. 2021. [Word alignment by fine-tuning embeddings on parallel corpora](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

- 878–891, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Takumi Goto, Yusuke Sakai, and Taro Watanabe. 2026. [Grammatical error correction evaluation by optimally transporting edit representation](#). *Transactions of the Association for Computational Linguistics*, 14:1692–1710.
- Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. [SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Yang Liu and Maosong Sun. 2015. [Contrastive unsupervised word alignment with non-local features](#). In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Rada Mihalcea and Ted Pedersen. 2003. [An evaluation exercise for word alignment](#). In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.
- Graham Neubig. 2011. [The kyoto free translation task](#).
- Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. [Are references really needed? unbabel-IST 2021 submission for the metrics shared task](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.
- 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.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- David Vilar, Maja Popovic, and Hermann Ney. 2006. [AER: do we need to “improve” our alignments?](#) In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*, Kyoto, Japan.
- Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti, and Anna Korhonen. 2023. [Probing cross-lingual lexical knowledge from multilingual sentence encoders](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual e5 text embeddings: A technical report](#). *Preprint*, arXiv:2402.05672.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

A MT Evaluation on DA

Table 5 shows the Pearson correlation with DA across different multilingual embedding models under the same BERTScore calculation framework. The trend is substantially different from that observed for MQM. In MQM, the final layer achieved the highest correlation with human evaluation across all models. In contrast, for DA, the layer with the highest correlation differs depending on the model. The differences between models are also smaller, suggesting that DA may be less sensitive to the choice of representation layer than MQM.

B Additional Word Alignment Analysis

Additionally, we added one more alignment extraction methods, then compared with MNN. The addition is Awesome-style extraction, which follows the extraction process used in Awesome-Align (Dou and Neubig, 2021) while replacing the underlying token representations with those from each model.

Table 6 shows AER across different multilingual embedding models. “Awesome-Align-Style Extraction” denotes the alignment extraction method of Awesome-Align. Replacing mBERT representations with token representations extracted from sentence embedding models affects alignment quality. Under Awesome-Align-Style extraction without fine-tuning, mBERT still achieves strong overall performance.

Models ($\uparrow$)	Layer	cs-en	de-en	ja-en	ru-en	uk-en	zh-en
COMET-22	-	0.821	0.446	0.976	0.857	0.714	0.135
COMET-20	-	0.827	0.424	0.989	0.847	0.723	0.330
BS(mBERT)	7	0.381	0.381	0.991	0.849	0.721	0.494
BS(XLM-R)	12	0.825	0.389	0.991	0.834	0.712	0.520
BS(LaBSE)	4	0.821	0.385	0.993	0.835	0.724	0.487
BS(E5)	4	0.834	0.400	0.992	0.832	0.706	0.491
BS(BGE-M3)	9	0.834	0.399	0.992	0.834	0.712	0.494

Table 5: Pearson correlation coefficients with DA on the WMT 2022 Metrics Shared Task. BS denotes BERTScore calculation framework.

Model ($\downarrow$)	Align.	Layer	fr-en	ja-en	ro-en	zh-en
Awesome Align	-	-	4.1	37.4	22.6	13.4
mBERT	ASE	-	5.6	45.6	27.9	18.1
	MNN	8	5.9	48.6	30.5	19.5
XLM-R	ASE	-	9.2	62.0	28.6	30.3
	MNN	8	6.5	57.5	25.9	24.2
LaBSE	ASE	-	12.6	51.6	24.1	28.7
	MNN	8	3.9	43.2	22.0	16.4
E5	ASE	-	8.6	57.3	27.6	27.6
	MNN	7	5.4	55.5	25.8	22.9
BGE-M3	ASE	-	9.6	73.5	34.3	38.8
	MNN	13	5.7	58.1	25.8	25.0

Table 6: Word alignment AER across multilingual embedding models and alignment methods. ASE denotes Awesome-Align-Style Extraction, and MNN denotes mutual nearest neighbor matching.

C Language Pair Abbreviations

We use the following language pair abbreviations throughout the paper: English–German (en–de), English–Russian (en–ru), French–English (fr–en), Romanian–English (ro–en), Japanese–English (ja–en), and Chinese–English (zh–en).