Intrinsic vs. Extrinsic Evaluation of Czech Sentence Embeddings: Semantic Relevance Doesn't Help with MT Evaluation

Petra Barančíková and **Ondřej Bojar**
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{barancikova,bojar}@ufal.mff.cuni.cz

Abstract

In this paper, we compare Czech-specific and multilingual sentence embedding models through intrinsic and extrinsic evaluation paradigms. For intrinsic evaluation, we employ Costra, a complex sentence transformation dataset, and several Semantic Textual Similarity (STS) benchmarks to assess the ability of the embeddings to capture linguistic phenomena such as semantic similarity, temporal aspects, and stylistic variations. In the extrinsic evaluation, we fine-tune each embedding model using COMET-based metrics for machine translation evaluation.

Our experiments reveal an interesting disconnect: models that excel in intrinsic semantic similarity tests do not consistently yield superior performance on downstream translation evaluation tasks. Conversely, models with seemingly over-smoothed embedding spaces can, through fine-tuning, achieve excellent results. These findings highlight the complex relationship between semantic property probes and downstream task, emphasizing the need for more research into "operationalizable semantics" in sentence embeddings, or more in-depth downstream tasks datasets (here translation evaluation).

1 Introduction

Machine translation (MT) evaluation has advanced significantly in recent years, finally moving beyond traditional surface-level metrics like BLEU (Papineni et al., 2002) towards more sophisticated approaches based on neural networks and contextualized embeddings.

State-of-the-art MT evaluation metrics such as COMET (Rei et al., 2022b) and BLEURT (Sellam et al., 2020) use sentence embeddings to better capture semantic similarity between translations and references, achieving much higher correlation with human judgments than traditional metrics.

However, the rapid development of new embedding models presents MT researchers with a challenging choice. Although multilingual models such as LaBSE (Feng et al., 2022) and XLM-RoBERTa (Conneau et al., 2019) have shown strong cross-lingual capabilities, there are also language-specific models that claim superior performance for a selected language. For morphologically rich languages like Czech, it remains unclear whether these specialized sentence embeddings offer advantages over multilingual alternatives when used in MT evaluation.

In this paper, we examine the evaluation of English-to-Czech machine translation and compare several state-of-the-art Czech-specific models against multilingual models using both intrinsic evaluation and extrinsic evaluation. To this end, we see the task of machine translation evaluation (MTE) and quality estimation (QE), i.e. MTE without professionally translated reference sentences, as methods for extrinsic evaluation of sentence embeddings. For intrinsic evaluation, we assess how well the examined sentence embeddings reflect semantic properties exemplified in two datasets: Costra (Barančíková and Bojar, 2020) and Semantic Textual Similarity (STS, Bednář et al., 2024). In sum, our goal is to understand whether the performance of a model in intrinsic semantic tasks correlates with its usability for MT evaluation, potentially simplifying the selection of embeddings.

2 Related Work

Several studies have raised concerns about the use of STS as an evaluation metric. For instance, Reimers et al. (2016), Eger et al. (2019), and Zhelezniak et al. (2019) argue that, while STS can capture certain semantic similarities, it does not reliably predict how effective sentence representa-

Proceedings of Machine Translation Summit XX Volume 1, pages 265–275, June 23-27, 2025

tions will be for downstream tasks. These works highlight how STS tasks often encourage surface-level heuristics or oversimplified semantic similarity patterns that may not generalize to more complex applications like entailment or paraphrasing detection.

To address these limitations, new intrinsic evaluation methods such as EvalRank (Wang et al., 2022) and SentBench (Xiaoming et al., 2023) have been proposed, both of which exhibit a stronger correlation with extrinsic evaluation measures. These benchmarks evaluate sentence representations through information retrieval, sentence ordering, and probing tasks, offering a more holistic view of embedding quality that aligns better with actual downstream task performance.

It is important to note that these previous experiments did not specifically focus on machine translation evaluation, which seems to be very close to STS—it also involves comparing pairs of sentences to assess their semantic closeness. Cífka and Bojar (2018) report a negative correlation between the translation quality of Transformer models measured by BLEU and the semantic properties (assessed using STS) of the sentence embeddings derived from the Transformer model. In contrast, Libovický and Madhyastha (2019) demonstrate a strong positive correlation between STS performance and translation quality for both Transformer and RNN-based models.

More recently, Freitag et al. (2022) have advocated for the use of semantic-aware metrics such as BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020) in MT evaluation, showing that these outperform BLEU in correlating with human judgments. These models incorporate contextual embeddings and often exhibit closer alignment with human-perceived meaning, bringing MT evaluation closer to the goals of intrinsic semantic understanding.

3 Models

For our experiments, we used several state-of-the-art sentence embedding models, employing both Czech-specific and multilingual variants. The Czech encoders include three *base*-size Transformer architectures, each using masked language modeling as their primary pretraining objective—CZERT-b-cased (`Czert`, Sido et al., 2021), FERNET-C5 (`FERNET`, Lehečka and Švec, 2021) and `RobeCzech` (Straka et al., 2021).

To provide a broader comparative analysis, we also experimented with multilingual sentence embedding models trained on datasets that contain Czech texts. These include `LaBSE` (Feng et al., 2022), a model generating language-agnostic representations for more than one hundred languages with remarkable cross-lingual alignment, since its training objective was machine translation. Furthermore, we evaluated two *large* models: XLM-RoBERTa-large (`XLM-R`, Conneau et al., 2019) and multilingual-e5-large (`mE5`, Wang et al., 2024), a model pretrained using a contrastive learning approach on a diverse range of tasks, including natural language inference and question answering across multiple languages.

As a baseline model, we employed a BERT architecture (Devlin et al., 2019) with randomly initialized weights. The only component inherited from the pretrained *'bert-base-multilingual-cased'* model is the tokenizer. This means that while the model processes input according to the tokenization patterns learned from multilingual data, it does not benefit from any pretrained language representations. We refer to this configuration as `random BERT`. This setup isolates and assess the contribution of the tokenizer alone, establishing a lower performance bound and offering a meaningful point of comparison to evaluate the benefits of pretraining.

4 Intrinsic Evaluation

We first evaluate the embeddings using a series of semantic benchmarks to determine their ability to accurately capture various semantic properties of a sentence.

4.1 Costra

As the first dataset for intrinsic evaluation, we used the Costra[1] dataset (Barančíková and Bojar, 2020). It was created manually, specifically to test the quality of Czech embeddings, focusing on complex transformations of sentences beyond standard paraphrasing or simple word-level changes. The sentence embeddings are tested across the following six categories:

- **Basic**: evaluates whether paraphrases are positioned closer together in embedding space compared to transformations that significantly alter the meaning of the original sentence.

[1] `https://github.com/barancik/costra`

- **Modality**: measures whether paraphrases are more similar to their original sentence than transformations that change the sentence's modality (e.g., possibility or prohibition).

- **Time, Style, Generalization, Opposite**: these categories test embeddings' ability to reflect linear ordering of sentence variations (e.g., from the least general to the most general) as proposed by annotators.

Each category is scored on a scale from 0 (worst) to 1 (best), reflecting the proportion of Costra sentence triplets for which the relations in the sentence vector space align with human annotations. For example, consider a triplet consisting of a seed sentence S, its paraphrase P, and its opposite sentence O. Ideally, the cosine similarity S_C should satisfy $S_C(S,P) > S_C(S,O)$ and $S_C(S,P) > S_C(P,O)$, indicating that the model correctly identifies the paraphrase closer to the seed sentence than the opposite sentence.

The results are presented in Table 1, with the overall Costra score calculated as the arithmetic mean across all six categories. In particular, the evaluation shows that all sophisticated models failed to outperform randomly generated embeddings[2] in the first two categories, **Basic** and **Modality**. In fact, these categories were designed to be particularly challenging, including comparisons of paraphrases with substantial lexical variation and sentences that, despite the close lexical similarity to a paraphrase, differ significantly in meaning. These results suggest that all models were fooled by surface-level similarity, making randomly generated embeddings the overall winner in these two categories. Consequently, it is impossible to distinguish whether slight improvements in these categories can be attributed to model quality or to randomness.

To address this limitation, we introduce the **Costra–** score, calculated as the average of the four remaining categories: **Time**, **Style**, **Generalization**, and **Opposite**. However, the **Costra–** scores revealed only marginal differences across models. The smallest model, `SimCSE`, slightly outperformed its counterparts but the improvement was not substantial. In fact, the models performed only marginally better than the `random BERT` model, suggesting limited success in capturing phenomena tested in the Costra dataset, such as linearity of time or generalization. Several models, including large `XLM-R`, even performed worse than `random BERT`.

4.2 Semantic Textual Similarity

Table 2 presents the results of our evaluation of sentence embeddings on the Semantic Textual Similarity (STS) task. Performance is measured using the automated evaluation tool[3] provided by Bednář et al. (2024). This tool computes similarity for pairs of sentences in three STS datasets. For precomputed sentence embeddings, it explores different embedding similarity metrics including cosine similarity, dot product, and Manhattan distance. Additionally, it applies various sentence embeddings pooling strategies and selects the highest average score as the final result.

Interestingly, the results are consistent with findings from Costra, with `SimCSE` being the overall best performing model, followed by `mE5` and `LaBSE` in the next two positions. Surprisingly, `XLM-R`, despite being a powerful multilingual model, may not be well-optimized for Czech-specific STS tasks, ranking last in the evaluation, performing even worse than `random BERT`.

5 Extrinsic Evaluation—MTE and QE

Extrinsic evaluation utilizes sentence embeddings as feature vectors for machine learning algorithms in downstream NLP tasks—MTE and QE in our tasks. It serves well to choose the best method for a particular task but not as an absolute metric of embedding quality, as the performance of the embeddings does not correlate across different tasks (Bakarov, 2018).

5.1 Data

In the following experiments, we utilize datasets from the Workshop on Machine Translation (WMT), selecting data from English-to-Czech translations. These datasets include English source sentences, Czech hypotheses (i.e., machine translated outputs), Czech reference sentences, and the human translation quality scores collected using the Direct Assessment (DA) method (Graham et al., 2013) and subsequently z-normalized.

Data from WMT17 to WMT19 (Bojar et al., 2017, 2018; Barrault et al., 2019) were used to

[2]Not to be confused with `random BERT`, we evaluated Costra also using completely *random vectors*.

[3]https://github.com/seznam/czech-semantic-embedding-models

Embeddings	Size	Basic	Mod.	Time	Style	Gen.	Opp.	Costra	Costra–
SimCSE	256	0.20	0.35	**0.74**	0.63	0.73	**0.78**	**0.57**	**0.72**
mE5	1,024	0.24	0.34	0.71	0.62	**0.75**	0.77	**0.57**	0.71
LaBSE	768	0.20	0.26	0.71	0.63	**0.75**	0.75	0.55	0.71
RetroMAE	256	0.06	0.06	0.69	0.63	0.70	0.76	0.48	0.70
RobeCzech	768	0.15	0.13	0.69	**0.65**	0.69	0.75	0.51	0.70
random BERT	768	0.08	0.06	0.65	0.60	0.72	0.73	0.47	0.68
Czert	768	0.31	0.35	0.66	0.64	0.69	0.69	0.56	0.67
XLM-R	1,024	0.16	0.11	0.65	0.61	0.67	0.68	0.48	0.65
FERNET	768	0.33	0.38	0.65	0.61	0.63	0.68	0.54	0.64
random vectors	256	**0.50**	**0.51**	0.49	0.50	0.49	0.50	0.50	0.50

Table 1: This Table presents the results of intrinsic evaluation using the Costra dataset. The Costra score ranges from 0 (worst) to 1 (best) in each category. The overall Costra score is calculated as the arithmetic mean across all categories. Costra– represents the mean score excluding the first two categories (Basic and Mod.), as these categories appear excessively challenging for all pretrained encoders evaluated.

Embeddings	avg. similarity
SimCSE	**87.83**
LaBSE	82.91
mE5	78.39
RetroMAE	76.30
Czert	74.79
RobeCzech	70.28
FERNET	65.46
random BERT	60.48
XLM-R	57.88

Table 2: Results of intrinsic evaluation on three STS datasets.

train the COMET estimators (Rei et al., 2020). The validation of the models was performed on the WMT20 dataset (Barrault et al., 2020), and the performance of the models was tested using the WTM21 (Akhbardeh et al., 2021) and WMT22 (Kocmi et al., 2022) datasets.

5.2 MTE Baseline Approach

Before fine-tuning the sentence embedding models for machine translation evaluation, we conducted a preliminary analysis to assess their default ability to evaluate translation quality. Specifically, we examined Pearson's correlation between human judgments and the cosine similarities computed between (i) a hypothesis and a reference translation and (ii) a hypothesis and a source sentence. We expected high cosine similarity for multilingual models, reflecting their ability to capture cross-lingual semantic relationships, whereas Czech-specific models—lacking such cross-lingual information—were anticipated to have random similarity scores.

Furthermore, we examined the intrinsic quality of the embedding spaces by measuring the cosine similarity between the source and reference embeddings. We also performed a random shuffling experiment designed to evaluate the discriminative ability of the embeddings.

The results presented in Table 3 reveal that even without fine-tuning, a slight correlation between human judgments and cosine similarity of hypotheses and references is observable in certain models—particularly mE5, RetroMAE, and SimCSE. However, contrary to expectations, this does not hold for source sentences; no relationship was detected between human evaluation scores and the cosine similarity computed between a translated sentence and its source sentence, even among the multilingual models.

The analysis of the embedding space via similarity between source and reference sentences provides further insights. In line with our hypothesis, XLM-R exhibits a near perfect similarity between the source and reference sentences, indicative of a tightly clustered or language-agnostic representation; however, the same holds for random BERT.

To further investigate this behavior, we repeated the experiment using random shuffle of source and

Sentence Embeddings	WMT21 test set			WMT22 test set			
	$\rho_{H,R}$	$\rho_{H,S}$	$S_C(S,R)$	$\rho_{H,R}$	$\rho_{H,S}$	$S_C(S,R)$	$S_C(S_R,R_R)$
`mE5`	**0.29**	0.04	0.89	0.26	0.01	0.90	0.75
`RetroMAE`	0.26	-0.10	0.76	**0.27**	**0.09**	0.76	0.69
`SimCSE`	0.24	**0.13**	0.85	0.25	0.05	0.82	0.09
`Czert`	0.20	-0.03	0.63	0.18	-0.06	0.62	0.52
`XLM-R`	0.17	-0.08	1.00	0.05	-0.10	1.00	0.99
`RobeCzech`	0.15	-0.16	0.92	0.11	-0.06	0.91	0.89
`LaBSE`	0.11	0.03	0.89	0.19	0.06	0.88	0.31
`FERNET`	0.07	-0.11	0.45	0.11	-0.03	0.40	0.35
`random BERT`	0.06	-0.16	0.99	0.03	-0.20	0.98	0.98

Table 3: Results for baseline MTE approach—using sentence embeddings for direct evaluation without fine-tuning. $\rho_{H,R}$ represents Pearson correlation between human quality assessments and the cosine similarity between the translation hypothesis and the reference translation, while $\rho_{H,S}$ shows the correlation between human judgments and the cosine similarity between the hypothesis and the source sentence. $S_C(S,R)$ represents cosine similarity between references and sources. The last column represents cosine similarity between randomly shuffled source and reference sentences averaged over 100 runs.

reference sentences; see the last column of Table 3. The similarity remained perfect for both `XLM-R` and `random BERT` even on shuffled pairs, indicating an overly invariant embedding space, where even pairs of semantically unrelated sentences tend to cluster together. This *over-smoothing* reduces the model's capacity to distinguish subtle differences that are essential for evaluating translation quality. In such cases, even bad translations can receive high similarity scores, lowering the correlation with human judgment. This also explains the poor performance of `XLM-R` in our intrinsic evaluation task, especially in STS (Table 2). More broadly speaking, it casts doubts on any results based on the direct similarity of `XLM-R` embedding vectors in the Czech language, given that `XLM-R` assigns similar vectors to random Czech sentences.

5.3 Models fine-tuning for MTE and QE

For all sentence encoders, we fine-tuned two COMET-based estimators (CE, Rei et al., 2020), one for machine translation evaluation using reference sentences and the other for quality estimation without reference sentences. The COMET models use a dual-encoder architecture: the source sentence, reference translation, and hypothesis are each processed independently using transformer encoder models followed by two hidden layers of sizes 3072 (resp. 2048 for QE) and 1024.

We used the default training settings with the AdamW optimizer ($1.5 \cdot 10^{-5}$ for the regression layers and $1.0 \cdot 10^{-6}$ for the encoder) and a layer-wise decay of 0.95. To preserve encoder generalization, the embeddings were frozen for the first 0.3 epochs. Both models used mixed-layer pooling with a sparsemax-based transformation before pooling and were optimized with mean squared error loss (using a dropout of 0.1). Training was conducted over five epochs, and we selected the checkpoint with the highest Kendall's tau validation on a held-out validation dataset.

These settings were applied consistently across all models without extensive hyperparameter tuning. In total, we trained a total of 18 COMET estimators. To avoid confusion with the original embeddings, we refer to a trained COMET estimator for given embeddings X as to $\mathrm{CE}_{\mathrm{MTE}}$(X) for machine translation with reference sentences and $\mathrm{CE}_{\mathrm{QE}}$(X) for the referenceless quality estimation metric (e.g., for the `Czert` embeddings, we use $\mathrm{CE}_{\mathrm{MTE}}$(`Czert`) and $\mathrm{CE}_{\mathrm{QE}}$(`Czert`), respectively).

5.4 Results of MTE and QE evaluation

We compare the performance of trained evaluation metrics at the system level with traditional string-matching MTE metrics. Specifically, we include BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and ChrF (Popović, 2015), using their default configurations as implemented in SacreBLEU (Post, 2018). Additionally, we employ METEOR-NEXT (Denkowski and Lavie, 2010), a metric that include paraphrase support, on both system and segment levels.

Furthermore, we compute scores using the official pretrained COMET models for machine translation evaluation, namely `wmt22-comet-da` (Rei

	system-level		segment-level	
MTE metrics	2021	2022	2021	2022
$\mathrm{CE}_{\mathrm{MTE}}$(FERNET)	**0.98**	**0.97**	0.60	0.47
wmt22-comet-da	0.97	0.93	**0.66**	**0.51**
$\mathrm{CE}_{\mathrm{MTE}}$(Czert)	0.96	0.93	0.57	0.43
$\mathrm{CE}_{\mathrm{MTE}}$(XLM-R)	0.96	0.93	0.62	0.47
$\mathrm{CE}_{\mathrm{MTE}}$(RobeCzech)	0.97	0.92	0.58	0.44
$\mathrm{CE}_{\mathrm{MTE}}$(mE5)	0.96	0.92	0.59	0.46
METEOR-NEXT	**0.98**	0.84	0.24	0.21
chrF2	0.97	0.84	-	-
$\mathrm{CE}_{\mathrm{MTE}}$(RetroMAE)	0.91	0.82	0.43	0.34
$\mathrm{CE}_{\mathrm{MTE}}$(LaBSE)	0.89	0.79	0.56	0.45
$\mathrm{CE}_{\mathrm{MTE}}$(SimCSE)	0.96	0.74	0.44	0.37
1-TER	0.95	0.60	-	-
BLEU	0.94	0.54	-	-
$\mathrm{CE}_{\mathrm{MTE}}$(random BERT)	0.35	-0.35	0.23	0.22

	system-level		segment-level	
QE metrics	2021	2022	2021	2022
$\mathrm{CE}_{\mathrm{QE}}$(FERNET)	**0.98**	**0.96**	0.60	0.46
wmt22-cometkiwi-da	0.95	0.91	**0.67**	**0.49**
$\mathrm{CE}_{\mathrm{QE}}$(XLM-R)	0.96	0.88	0.63	**0.49**
$\mathrm{CE}_{\mathrm{QE}}$(RobeCzech)	0.97	0.87	0.58	0.39
$\mathrm{CE}_{\mathrm{QE}}$(Czert)	0.95	0.86	0.57	0.39
$\mathrm{CE}_{\mathrm{QE}}$(mE5)	0.93	0.76	0.59	0.45
$\mathrm{CE}_{\mathrm{QE}}$(LaBSE)	0.83	0.39	0.54	0.40
$\mathrm{CE}_{\mathrm{QE}}$(RetroMAE)	0.64	0.15	0.39	0.23
$\mathrm{CE}_{\mathrm{QE}}$(random BERT)	0.47	-0.19	0.26	0.20
$\mathrm{CE}_{\mathrm{QE}}$(SimCSE)	0.12	-0.92	0.38	0.24

Table 4: Correlations between human scores and evaluation metrics, including both fine-tuned COMET-based metrics and traditional metrics, computed at the system and segment levels.

et al., 2022a), and for quality estimation, specifically wmt22-cometkiwi-da (Rei et al., 2022b). These COMET models extend beyond a simple trained COMET estimator, as they incorporate an ensemble approach combining a COMET estimator trained on DA data and sequence predictors trained on MQM annotations.

The results in Table 4 indicate a clear advantage for COMET-based evaluation metrics over traditional metrics in MTE. In the system-level analysis, the COMET variants $\mathrm{CE}_{\mathrm{MTE}}$(FERNET) and $\mathrm{CE}_{\mathrm{QE}}$(FERNET) achieved consistently remarkably high correlation outperforming even the official COMET ensemble metrics – wmt22-comet-da and wmt22-cometkiwi-da, which were the top-performing metrics at the segment level.

In contrast, classical metrics, although competitive in 2021, showed significant performance degradation in 2022. $\mathrm{CE}_{\mathrm{MTE}}$(random BERT) failed completely, highlighting the importance of using pretrained sentence embeddings, even though $\mathrm{CE}_{\mathrm{QE}}$(random BERT) outperformed $\mathrm{CE}_{\mathrm{QE}}$(SimCSE), even though SimCSE was the best performing encoder in intrinsic evaluation.

Another interesting observation is the small difference in correlations of the top performing embeddings between MTE and QE. The correlation of $\mathrm{CE}_{\mathrm{MTE}}$(FERNET) and $\mathrm{CE}_{\mathrm{QE}}$(FERNET) is practically equal at both the system and segment levels, as if these metrics no longer have use for reference translations. This is consistent with recent research showing that reference-free evaluation has become competitive with reference-based evaluation (Rei et al., 2021) or even outperforms it (Moosa et al., 2024).

6 Results and Discussion

When comparing the results of MTE and QE with those of the intrinsic evaluation tasks, we can observe an interesting inversion. Although both evaluation approaches aim to capture semantic similarity, the performance of the embeddings changed significantly after fine-tunings. Specifically, XLM-R and FERNET embeddings, which performed poorly in intrinsic evaluation, became the best performing MTE and QE metrics. In contrast, SimCSE, which dominated intrinsic evaluations, ranks among the worst performing metrics in MTE and QE. These results are in line with related research (Section 2), which shows that STS performance may not accurately predict effectiveness in downstream tasks.

There are several plausible hypotheses that might explain these discrepancies. Let us at least mention them here— unfortunately, their thorough testing is beyond the scope of this article.

First, XLM-R and FERNET might perform poorly in intrinsic tasks because their representation spaces are not tuned for fine-grained semantic differences. However, when fine-tuned on a translation quality task, the model might learn to emphasize those aspects of the embedding space that are important for distinguishing translation quality.

The fine-tuning process for COMET-based evaluation might be effectively reconfiguring the XLM-R embedding space, transforming its initially oversmoothed representations into task-specific features that are highly discriminative for translation quality. Although XLM-R raw embeddings appear to be all clustered together (see Table 3), the fine-tuning may introduce transformations that reweight and separate the dimensions relevant for capturing translation errors. In contrast, SimCSE embeddings, which are already optimized for intrinsic semantic discrimination, might leave less room for adjustments necessary to learn the new training objective.

We should also not forget about the different embedding sizes, which played an important role in the observed behavior. The *small* embeddings—SimCSE and RetroMAE—were among the worst performing COMET estimators. Large embeddings, such as those produced by XLM-R, offer a higher-dimensional space that can capture more nuanced semantic and syntactic features. When fine-tuning with the COMET estimator—which adds two hidden layers with sizes 3072 (resp. 2048) and 1024—the richer representation provided by larger embeddings could allow the model to extract and emphasize the translation-specific signals more effectively.

Interestingly, we can see not too much difference between the monolingual vs. multilingual embedding performances—they seem to be equally represented among the best performing embeddings in both intrinsic and extrinsic tasks. The size of the embeddings seem not to matter in the intrinsic tasks—the top 3 best performing embeddings (SimCSE, LaBSE and mE5) are *small*, *base* and *large*, respectively.

The correlation analysis between different evaluation methodologies, visualized by heatmaps in Figure 1, reveals interesting patterns of how different evaluation methodologies relate to each other. These patterns provide valuable insight into the reliability and consistency of various embedding evaluation approaches.

The heatmaps highlight a strong alignment among all intrinsic tasks (Costra–, STS, and MTE baselines). Moreover, there is a strong correlation between the segment-level and the system-level metrics, indicating that aggregated segment scores provide reliable system-level insights. In particular, we observe strong correlations between segment-level metrics (*segment MTE* and *segment QE* showing correlations of 0.97 and 0.91 for 2021 and 2022 respectively), suggesting that these evaluation approaches capture similar aspects of translation quality despite their methodological differences.

However, one of the most striking findings is the weak and sometimes even negative correlation between intrinsic evaluation metrics (Costra–, STS) and the system-level quality estimation scores *system QE*. This discrepancy is particularly evident in the 2022 data, where Costra– shows a negative correlation (-0.52) with *system QE*, challenging the assumption that better semantic representation capabilities necessarily translate to improved MT evaluation performance.

These findings indicate that intrinsic measures, while useful for general semantic similarity, may not sufficiently reflect translation-specific nuances required for MTE or QE. Consequently, intrinsic criteria alone appear inadequate for selecting optimal sentence embeddings for these specific tasks. Further research is needed to identify intrinsic evaluation methods that better capture the subtleties relevant to machine translation. Additionally, it would be valuable to explore in more detail the types of errors penalized in manual MT quality assessments

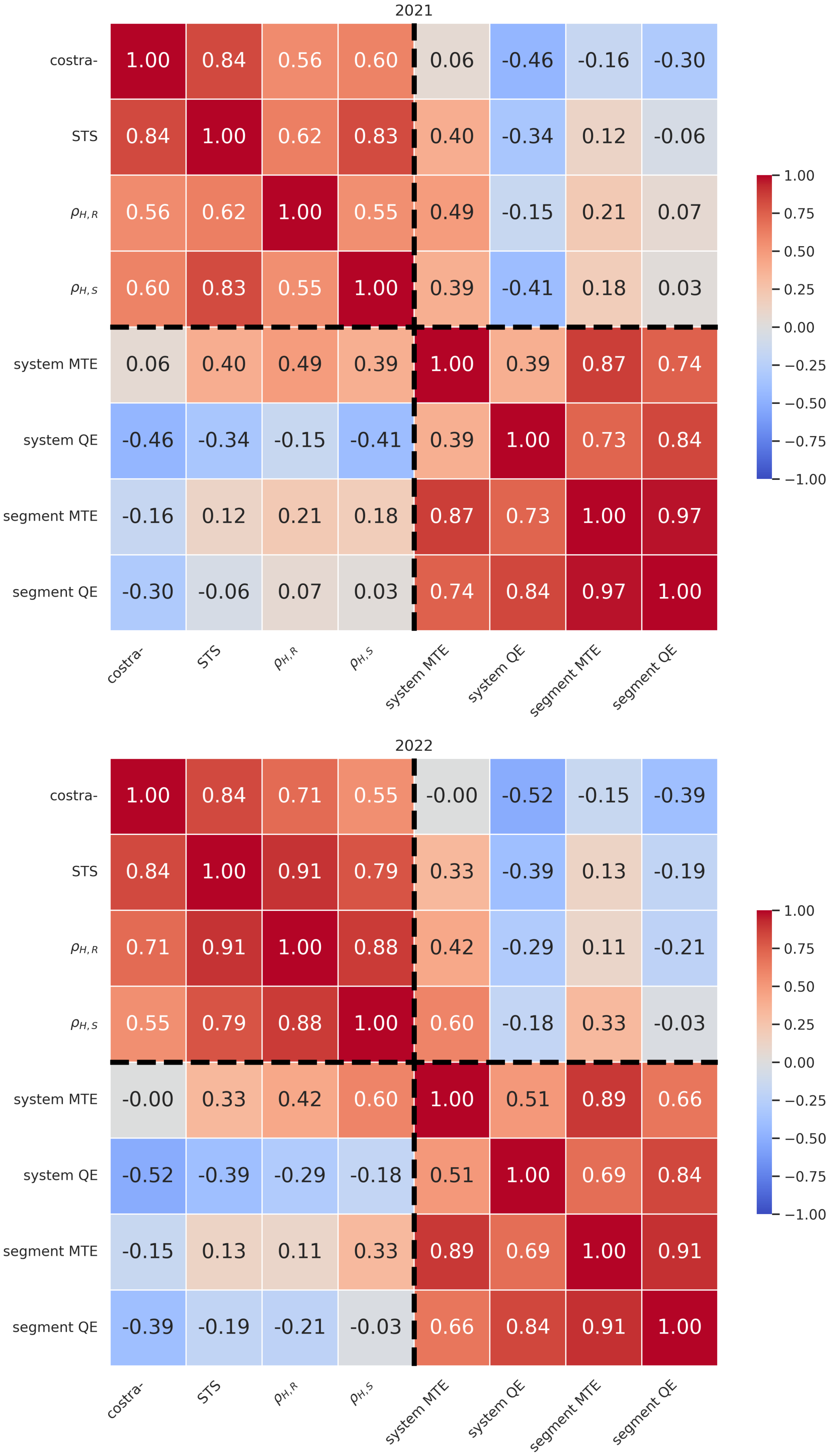

Figure 1: Correlation heatmaps for different method of embedding evaluations. The heatmaps are intentionally kept in a rectangular shape to emphasize the mismatch in correlation patterns between intrinsic evaluation (Costra–, STS, $\rho_{H,R}$, $\rho_{H,S}$) and extrinsic evaluation (system MTE/QE and segment MTE/QE).

to determine whether these errors predominantly concern sentence meaning or other aspects that should be preserved in translation.

7 Conclusion and Future Work

We experimented with several evaluation methods for both Czech and multilingual sentence embeddings, considering intrinsic semantic tasks and downstream application in machine translation evaluation and quality estimation. Our key findings include the following:

- **Intrinsic vs. Extrinsic Discrepancy:** The lack of correspondence between the intrinsic and extrinsic metrics used in our experiments suggests that intrinsic evaluation methods employing these metrics cannot reliably predict a model's performance in MT evaluation tasks. This finding suggests the need for better targeted intrinsic evaluation approaches that reflect downstream application requirements (Figure 1).

- **Temporal Stability:** The stability of the correlations over time between the segment-level metrics provides encouraging evidence for the reliability of these evaluation approaches.

- **Language-Specific vs. Multilingual Models:** There are no strong differences in performance between language-specific and multilingual models. Both categories are comparably represented among the top-performing models in intrinsic and extrinsic tasks.

- **Model Size Might Matter:** In contrast to intrinsic tasks, fine-tuning embeddings for MTE/QE reveals that model size does matter, with the *small* embeddings consistently showing poor performance.

In future work, we intend to replicate these experiments across multiple languages to investigate whether the observed behavior is specific to the Czech language or if it generalizes to other languages. In addition, we plan to conduct a more thorough analysis to better understand the underlying reasons for the differences in performance between the evaluation methods.

Acknowledgments

The authors acknowledge the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO and the project OP JAK Mezisektorová spolupráce Nr. CZ.02.01.01/00/23_020/0008518 named "Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím." This work has been conducted using data provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Amir Bakarov. 2018. A Survey of Word Embeddings Evaluation Methods. *CoRR*, abs/1801.09536.

Petra Barančíková and Ondřej Bojar. 2020. COSTRA 1.0: A Dataset of Complex Sentence Transformations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3535–3541, Marseille, France. European Language Resources Association.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019.

Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Jiří Bednář, Jakub Náplava, Petra Barančíková, and Ondřej Lisický. 2024. Some Like It Small: Czech Semantic Embedding Models for Industry Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22734–22742.

Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, and Julia Kreutzer, editors. 2017. *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and Meaning Representation in Opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1362–1371, Melbourne, Australia. 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. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden. 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.

Steffen Eger, Andreas Rücklé, and Iryna Gurevych. 2019. Pitfalls in the Evaluation of Sentence Embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 55–60, Florence, Italy. 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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Jan Lehečka and Jan Švec. 2021. Comparison of Czech Transformers on Text Classification Tasks. In *Statistical Language and Speech Processing*, pages 27–37, Cham. Springer International Publishing.

Jindrich Libovický and Pranava Madhyastha. 2019. Probing Representations Learned by Multimodal Recurrent and Transformer Models. *CoRR*, abs/1908.11125.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. MT-Ranker: Reference-free machine translation evaluation by inter-system ranking. In *The Twelfth International Conference on Learning Representations*.

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.

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.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – Czech BERT-like Model for Language Representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1326–1338, Held Online. INCOMA Ltd.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Bin Wang, C.-C. Jay Kuo, and Haizhou Li. 2022. Just Rank: Rethinking Evaluation with Word and Sentence Similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077, Dublin, Ireland. 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. *arXiv preprint arXiv:2402.05672*.

Liu Xiaoming, Lin Hongyu, Han Xianpei, and Sun Le. 2023. SentBench: Comprehensive Evaluation of Self-Supervised Sentence Representation with Benchmark Construction. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 813–823, Harbin, China. Chinese Information Processing Society of China.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation Coefficients and Semantic Textual Similarity. 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 951–962, Minneapolis, Minnesota. Association for Computational Linguistics.