Some Like It Small: Czech Semantic Embedding Models for Industry Applications

Jiří Bednář, Jakub Náplava, Petra Barančíková, Ondřej Lisický

Seznam.cz, Prague, Czech Republic
{jiri.bednar,jakub.naplava,petra.barancikova,ondrej.lisicky}@firma.seznam.cz

Abstract

This article focuses on the development and evaluation of *Small*-sized Czech sentence embedding models. Small models are important components for real-time industry applications in resource-constrained environments. Given the limited availability of labeled Czech data, alternative approaches, including pre-training, knowledge distillation, and unsupervised contrastive fine-tuning, are investigated. Comprehensive intrinsic and extrinsic analyses are conducted, showcasing the competitive performance of our models compared to significantly larger counterparts, with approximately 8 times smaller size and 5 times faster speed than conventional *Base*-sized models. To promote cooperation and reproducibility, both the models and the evaluation pipeline are made publicly accessible. Ultimately, this article presents practical applications of the developed sentence embedding models in Seznam.cz, the Czech search engine. These models have effectively replaced previous counterparts, enhancing the overall search experience for instance, in organic search, featured snippets, and image search. This transition has yielded improved performance.

Introduction

In recent years, the field of natural language processing (NLP) has experienced remarkable progress. One key contributing factor to this progress is the development of more sophisticated distributed text representation, i.e., word and sentence embeddings (Devlin et al. 2019; Zhuang et al. 2021; Reimers and Gurevych 2019).

In this work, our primary focus lies in the development and evaluation of sentence embeddings. Sentence embeddings are instrumental in various NLP tasks such as facilitating efficient information retrieval, sentiment analysis, machine translation, question answering, or providing interpretability. The importance of quality sentence embeddings simply cannot be overstated.

While sentence embeddings are becoming more sophisticated and capable of capturing richer semantic and syntactic information, there is a trade-off between embedding size and computational efficiency. In industry applications, the choice of an embedding model should be carefully evaluated. Larger embeddings may offer better representation of sentence semantics, but they can significantly increase computational complexity and inference times, which may not be practical in real-time applications or resource-constrained environments.

Furthermore, it is essential to acknowledge that much of the research and development in sentence embeddings has predominantly focused on English language, both in supervised and unsupervised settings. This creates challenges and limitations when applying these models to other languages, where data availability, linguistic characteristics, and semantic structures may differ significantly from English.

This paper aims to address the existing gap by focusing on the development of *Small*[1] sentence embeddings specially tailored for real-time industry applications in the Czech language. Nevertheless, the underlying techniques and evaluation can be adapted and applied to other languages as well. The research and strategies presented in the paper serve as a valuable foundation for enhancing sentence embeddings, with the goal of optimizing computational efficiency without sacrificing their representational capacity and performance.

The main contributions of our paper are as follows:

1. ***Small* Czech models training**: We trained and evaluated multiple *Small* Czech BERT (Devlin et al. 2019) based models for sentence embeddings. Despite being approximately 8 times smaller and 5 times faster compared to conventional *Base* models, the *Small* models exhibit competitive performance in diverse downstream tasks.
2. **Czech sentence embeddings evaluation**: We conducted a thorough evaluation of existing Czech sentence embeddings, both intrinsically and extrinsically. This evaluation provides valuable insights into the effectiveness and applicability of different embeddings in various NLP tasks.
3. **Evaluation pipeline**: To ensure transparency and facilitate the reproducibility and verification of our results, we have made our evaluation pipeline openly accessible.

[1]We follow the naming convention of Clark et al. (2020) in which *Small* models refer to BERT-based models with roughly 14 million parameters, 12 encoder layers, and have a hidden size of 256. *Base* models comprise circa 110 million parameters, 12 encoder layers and have a hidden size of 768.

4. **Public release of models**: Our developed models are publicly available under the CC-BY-4.0 and CC-BY-NC-4.0 licenses, depending on the training data used. This allows other researchers and practitioners to utilize and build upon our work in their own projects, fostering collaboration and advancement in the NLP community.

All the developed models and the evaluation pipeline are available at the following link:

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

Related Work

Recently, two prominent methodologies, cross-encoder (CE) and bi-encoder (BE) (Reimers and Gurevych 2019), have emerged as leading approaches in comparison of text pairs. Both methods encode sentence pairs and have demonstrated significant advancements in capturing the semantic meaning and relationships between sentences.

The CE approach entails encoding a pair of sentences together into a joint representation. By considering the interaction between the sentences, CE excels at capturing fine-grained semantic nuances and understanding complex relationships between text pairs (Qu et al. 2021). On the other hand, the BE approach takes an independent encoding approach (Zhao et al. 2022) – it generates a separate embedding for each sentence, allowing for greater flexibility and computational efficiency. BEs are well-suited for large-scale applications like semantic search (Reimers and Gurevych 2019), where the goal is to find the most similar sentence or document to a given query.

While CEs excel at providing rich semantic information, BEs can partially achieve similar performance by integrating multiple text representations, as demonstrated in ColBERT (Khattab and Zaharia 2020) or MADRM (Kong et al. 2022). However, the use of multiple representations can significantly increase the index size, making it impractical for certain applications. ColBERTer (Hofstätter et al. 2022) aims to mitigate this issue, thereby promoting wider adoption of single-representation BE.

Improving the performance of sentence embedding models often relies not only on refining task-specific architectures but also on utilizing rich labeled English datasets. The Sentence Transformers[2] framework is widely used for comparing sentence embeddings and provides a range of models, including *all-mpnet-base-v2*, fine-tuned with a contrastive objective (Song et al. 2020), and *all-MiniLM-L6-v2*, which provides high-quality embeddings while being computationally efficient. These models are trained on diverse datasets, covering over 1 billion sentence pairs.

Regrettably, there are no similarly extensive labeled datasets available for the Czech language. As a result, alternative strategies are necessary to optimize model performance. One effective approach is knowledge distillation, where the excellent performance of larger, more complex models can be utilized to train smaller student models (Hinton, Vinyals, and Dean 2015). In this context, a CE can be employed as a teacher to train a BE (Qu et al. 2021). Interestingly, BE and CE can also be learned simultaneously in an unsupervised manner using the Trans-Encoder (Liu et al. 2021), where one component generates pseudo-labels to update the other, eliminating reliance on labeled data. To address the scarcity of language-specific datasets, one possible solution is multilingual distillation (Reimers and Gurevych 2020), where a strong teacher model, trained in English for example, is used to train a student model in the target language. The main advantage of this approach is that the student model learns representations for both languages. However, this training process requires a bilingual dataset whose quality directly impacts student model performance.

To further enhance model performance, an alternative strategy involves refining pre-training methods. One such approach involves leveraging autoencoders to compress text data into the "CLS" token using an encoder and then using a decoder to reconstruct the original text. This technique has been demonstrated in models such as Condenser (Gao and Callan 2021), TSDAE (Wang, Reimers, and Gurevych 2021), and coCondenser (Gao and Callan 2022). However, a strong decoder may negatively impact sequence representation quality (Lu et al. 2021). Recent approaches such as SimLM (Wang et al. 2023) and RetroMAE (Xiao et al. 2022) address this issue by adopting shallow decoders with limited past context access and enhanced decoding mechanisms.

Another approach to enhancing text representation is through contrastive learning. A successful implementation of this approach is SimCSE (Gao, Yao, and Chen 2021). In SimCSE, authors applied a simple dropout mask as noise to the input text, creating positive pairs along with in-batch negatives. Despite its simplicity, this implementation has yielded surprisingly good results and has inspired numerous similar approaches, further improving model performance. Notable examples include DiffCSE (Chuang et al. 2022), InfoCSE (Wu et al. 2022), and the recently introduced RankCSE (Liu et al. 2023), which ensures ranking consistency between text pairs.

The effective utilization of pre-training strategies, such as employing autoencoders or contrastive learning, as well as incorporating knowledge distillation, can yield substantial improvements in capturing semantic meaning within text and consequently enhance the performance of models for languages that lack labeled datasets.

Models

Three distinct methodologies were employed in our work to train effective *Small* bi-encoder models and generate high-quality sentence embeddings: Auto-encoder training, unsupervised contrastive fine-tuning, and multilingual distillation. The primary advantage of these approaches is their independence from large supervised datasets, which are often unavailable for low-resource languages like Czech.

RetroMAE

The recent retrieval-oriented pre-training method, RetroMAE (Xiao et al. 2022), was used in combination with *Small* configurations of BERT as an encoder. The setup included

[2]https://www.sbert.net

12 layers, a hidden size of 256, and a WordPiece tokenizer with vocabulary size of 57,226, incorporating both Czech and English tokens. Asymmetrical masking ratios were applied: 30% for the encoder and 50% for the decoder. The Adam optimizer (Kingma and Ba 2014) with decoupled weight decay regularization (Loshchilov and Hutter 2019) (AdamW) was used in the training process with a learning rate of 5e-4, along with a cosine scheduler (Loshchilov and Hutter 2017) with a linear warmup of 10% training steps, and a total batch size of 512. The model was trained on the Czech corpus (see Section Training Data) for 2 epochs, or 250,000 steps.

Multilingual Distillation

In this experiment, we trained two models using multilingual distillation (Reimers and Gurevych 2020) with the English *all-mpnet-base-v2* model as the teacher, chosen for its high MTEB benchmark (Muennighoff et al. 2023) performance relative to its size. Despite the original paper using a pre-trained multilingual model to initialize the student model, we found a newly initialized BERT model with a merged Czech-English vocabulary to be more effective.

We addressed the discrepancy in embedding size between the teacher (768) and the student (256) by adding a linear projection to the student model and normalizing all embeddings during training. After training, the projection was removed, resulting in a bilingual student model with the desired embedding size.

We utilized two parallel datasets for our training: a high-quality dataset for non-commercial use, the *czeng20-csmono* (Kocmi, Popel, and Bojar 2020) and a commercial one, the *Paracrawler v9* (Bañón et al. 2020) (see Section Training Data). Our method was efficient, providing successful training on both, which resulted in the Dist-MPNet-CzEng and Dist-MPNet-ParaCrawl models, respectively.

Unsupervised Fine-Tuning

For enhanced performance, all pre-trained language models, including RetroMAE-Small and those employing multilingual distillation, underwent additional unsupervised fine-tuning. This process aimed to optimize their representations.

In our research, we found that models trained using conventional masked language modelling (MLM) objectives such as BERT, RoBERTa, or even more sophisticated ELECTRA pre-training, often performed sub-optimally in a bi-encoder setup. Therefore, we also fine-tuned *Small-E-Czech* (Kocián et al. 2022), a Czech *Small* ELECTRA model (Clark et al. 2020) , to evaluate the impact of unsupervised fine-tuning on models not tailored for sentence embedding.

Contrastive Learning The contrastive learning approaches used Wikipedia data for their experiments. The experiments involved three methodologies: SimCSE (Gao, Yao, and Chen 2021), RankCSE (Liu et al. 2023), and InfoCSE (Wu et al. 2022). Due to time constraints, we did not conduct an exhaustive hyperparameter grid search. Instead, we consistently adopted the recommended values provided by the original authors of each specific method. In each approach, consistent with the original SimCSE methodology, identical sentences were used as positive pair examples. The sole difference came from a dropout mask applied during creation, leading to minimal augmentation. In-batch negatives were also included.

The AdamW optimizer was used for all experiments, with learning rate 3e-5 for SimCSE and RankSCE, and learning rate 7e-6 for InfoCSE. Each model was trained for 3 epochs with a batch size of 128.

In RankCSE, the SimCSE model variants were utilized as teachers for ranking distillation. In line with Liu et al. (2023), both ListNet and ListMLE were experimented with, but no significant differences were observed, leading to the adoption of ListNet for further experiments.

TSDAE The TSDAE (Wang, Reimers, and Gurevych 2021) is a notable approach used for improving model representations or domain adaptation. However, its training is more computationally and memory intensive than previous methods. TSDAE employs an auto-encoder training, which is considerably slower than RetroMAE as it uses a deep (multi-layer) decoder rather than a shallow one.

In this experiment, Small-E-Czech was fine-tuned with TSDAE using a batch size of 16 and a learning rate of 5e-5 on a single GPU, following the original author's code. The encoder and decoder weights were tied during the training process. The model was trained on a sample of sentences from the Czech corpus for one epoch.

We also attempted to pre-train a BERT model from scratch using the TSDAE method. However, this resulted in exceedingly poor performance on downstream tasks. The trained embeddings might be overly complex for general NLP tasks, requiring a sophisticated decoder, which isn't suitable for typical fine-tuning scenarios that employ a simpler classifier or regressor.

Training Data

The pre-training process utilized a non-public corpus internally called *Czech corpus*, which is 253 GB in size and was obtained by Seznam.cz. This corpus comprises post-processed texts extracted from Czech web pages, encompassing diverse quality levels and lacking a specific domain specification. During the corpus cleaning phase, documents that were deemed too short, non-Czech, duplicated, or classified as spam were excluded.

For the multilingual distillation process, the English-to-Czech transfer was conducted using two distinct datasets. The first one, czeng20-csmono (Kocmi, Popel, and Bojar 2020), is a high-quality corpus containing 50 million pairs of Czech sentences and their corresponding synthetic English translations, intended strictly for non-commercial use. The second dataset, Paracrawler v9 (Bañón et al. 2020), is an open-source resource that also comprises 50 million Czech-English sentence pairs. Despite its less refined nature and origin through bitext mining, the dataset allows commercial applications.

In the unsupervised fine-tuning experiments, the training set was the Czech Wikipedia dump.[3] Pre-processing steps

[3] https://dumps.wikimedia.org

involved splitting paragraphs into sentences, filtering out special characters and removing sentences deemed too short or too long (Gao, Yao, and Chen 2021). Contrastive learning methods, such as SimCSE, RankCSE, InfoCSE, typically utilize a sample size of 10^6 sentences from the English Wikipedia with appropriate filtering. In this study, a similar approach was extended to include also the full training split from the Czech Wikipedia dataset, resulting in a total of 5.2×10^6 sentences after pre-processing. All models were fine-tuned using SimCSE on both datasets. Our experiments showed that using the larger dataset, on average, improved the score on the STS task by 1 percentage point for small models and 2 percentage points for base models.

Evaluation

To explore the performance of each model and understand its behavior across various NLP tasks, multiple evaluation tasks were conducted. Both intrinsic and extrinsic evaluations were performed to thoroughly assess the models.

Intrinsic evaluation aims to test how effectively the embeddings capture the semantic meaning and syntactic structure of sentences. Extrinsic evaluation focuses on the models' performance in a range of NLP challenges within real-world applications and downstream tasks.

To facilitate these evaluations, publicly available Czech datasets were utilized.

Intrinsic Evaluation

Costra Costra (Barančíková and Bojar 2020) is a dataset designated for evaluating the quality of sentence embeddings spaces. It examines the proficiency of sentence embeddings in capturing intricate linguistic phenomena such as paraphrases, tense, or style. For instance, given a triplet of sentences – an original, a paraphrased and an antonymous sentence – the vector similarity (cosine) between the original and paraphrased sentence is expected to be greater than the similarity between the original and antonymous sentence. The percentage of such cases is given as accuracy score.[4]

STS In the Semantical Textual Similarity (STS) task, we employ three datasets. The first, CNA, sourced from Sido (2021), contains 1,100 sentence pairs from Czech journalistic texts, using only the test data. The label for each sentence pair was determined by taking the average from 9 different annotations. The second, SVOB-IMG, comprises 850 pairs from image descriptions, and the third, SVOB-HL, has 525 pairs from headlines. Both were translated from the English SentEval dataset (Conneau and Kiela 2018) and subsequently annotated (Svoboda and Brychcín 2018).

Extrinsic Evaluation

Multi-Label Classification The Czech text document corpus CTDC (Kral and Lenc 2018) is a dataset designed for direct comparison of document classification on Czech data. It comprises 12,000 news articles labeled with 37 categories used for classification. A *5-fold cross-validation* procedure was conducted. A micro-averaged F1 score was used as an evaluation metric, reported with a standard deviation.

Sentiment Analysis The Czech Facebook Dataset (CFD) (Habernal, Ptáček, and Steinberger 2013) was employed for sentiment analysis. It consists of 2,587 positive, 5,174 neutral, and 1,991 negative posts. To assess model performance, a *10-fold cross-validation* procedure was conducted, following the methodology outlined in Straka et al. (2021). The evaluation metric used was the macro-averaged F1 score, and the reported results include the standard deviation.

Relevance Ranking Ranking documents based on user queries is a fundamental task in information retrieval, with applications in search engines and recommendation systems. To evaluate model performance in this context, the DaReCzech dataset (Kocián et al. 2022) was utilized. This dataset consists of 1.6 million Czech query-document pairs split into training, development and testing sets, each labeled with a relevance score.

In assessing the effectiveness of ranking algorithms, the precision at 10 (*p@10*) metric was employed. This metric is commonly used in information retrieval to measure how well the model performs in returning the relevant documents within the top 10 results for a given query.

Experiments

We present an evaluation of semantic models' quality in various settings, including zero-shot evaluation, linear-probing experiments, and whole model fine-tuning.

Zero-Shot Evaluation

Zero-shot evaluation assesses a model's generalization capabilities through embedding space quality measured with Costra and STS. Additionally, we conduct a zero-shot evaluation on DaReCzech, based on the concept that relevant documents typically show high semantic similarity to a query.

We use cosine similarity and CLS pooling (Devlin et al. 2019). However, for models that have been pre-trained using methods such as MLM, the choice of pooling method can greatly influence their performance in zero-shot scenarios. In these cases, MEAN or MAX pooling methods tend to give the best results. We determine the optimal setting by using

[4]The method of evaluation used in our study slightly differs from the one originally proposed by Barančíková and Bojar (2020). The reason for this difference is straightforward. In their study, Barančíková and Bojar (2020) assess models across six categories: *basic*, *modality*, *time*, *style*, *generalization*, and *opposite*. However, upon conducting our evaluations, we find that the first two categories (*basic* and *modality*) are too hard for all current models. In these categories, all examined embeddings performed below the random baseline. Consequently, it is unfeasable to distinguish whether good performance in these categories can be attributed to model quality or randomness. In order to keep the evaluation reliable, the *basic* and *modality* categories are omitted from our evaluation and instead only the average performance across the remaining four categories (*time*, *style*, *generalization*, and *opposite*) is reported as our Costra score. By adjusting the evaluation approach in this manner, we aimed to provide a more accurate and meaningful assessment of the model's performance in the selected categories, while acknowledging the limitations and challenges posed by certain linguistic phenomena in the evaluation process.

Model	Spearman's correlation				Accuracy	P@10
	SVOB-IMG	**SVOB-HL**	**CNA**	**STS-Average**	**Costra**	**DaReCzech**
Random baseline	2.40	3.85	29.09	11.78	49.54	38.10 ± 0.31
Random-small	64.61	55.30	67.69	62.53	68.73	40.38 ± 0.35
Avg. fastText	54.81	47.52	72.47	58.26	65.75	37.88 ± 0.31
Small-E-Czech	39.67	42.43	62.80	48.30	64.29	37.31 ± 0.33
RetroMAE-Small	78.88	66.21	83.82	76.30	69.66	42.16 ± 0.36
Dist-MPNet-ParaCrawl	90.11	77.66	84.99	84.25	70.42	42.33 ± 0.32
Dist-MPNet-CzEng	**90.94**	83.89	87.97	87.60	71.22	42.01 ± 0.37
SimCSE-Small-E-Czech	61.70	59.75	77.26	66.24	66.44	39.20 ± 0.38
TSDAE-Small-E-Czech	77.45	66.17	83.16	75.59	69.42	40.54 ± 0.37
SimCSE-RetroMAE	78.88	71.92	85.19	78.66	69.63	42.04 ± 0.37
RankCSE-RetroMAE	79.91	72.03	85.10	79.01	69.79	41.97 ± 0.36
InfoCSE-RetroMAE	79.30	65.58	84.31	76.40	69.89	41.77 ± 0.38
SimCSE-Dist-MPNet-ParaCrawl	90.29	78.80	85.91	85.00	71.12	**42.38** ± 0.35
SimCSE-Dist-MPNet-CzEng	90.73	**84.22**	**88.56**	**87.83**	**71.77**	42.18 ± 0.38
OpenAI Ada Embedding	83.51	78.04	86.21	82.59	69.01	42.21 ± 0.31

Table 1: Zero-shot evaluation of our models and several baselines on STS, Costra and DaReCzech. The table is organized into four horizontal sections: baselines, pre-trained models, models fine-tuned for sentence embeddings, and external sentence embedding services.

the Spearman's correlation with the CNA train dataset. In the case of Small-E-Czech the MEAN pooling is used.

Alongside our trained models, we evaluate a sentence embedding service for comparison, specifically the OpenAI's Ada embeddings (*text-embedding-ada-002*). These multilingual embeddings, each with a length of 1536, are expected to provide quality semantic features and a robust embedding space, making them suitable for zero-shot settings. In measuring similarity, we adhere to official recommendation of using cosine similarity. Although these embeddings could be a viable solution for industrial applications, their cost-effectiveness may be questionable for larger datasets, rendering them potentially unsuitable.

For our baselines, we employ the *random baseline* with random embeddings, *random-small* which is a model with the *Small* ELECTRA architecture initialized randomly without pre-training, and averaged fastText (Bojanowski et al. 2017) embeddings. We use the MEAN pooling for the random-small.

Zero-Shot Results As shown in Table 1, the initial performance of Small-E-Czech's embeddings in zero-shot evaluations was poor, but improved significantly with unsupervised methods such as SimCSE and TSDAE. Despite InfoCSE and RankCSE being theoretically advanced successors of SimCSE, they did not exhibit any improvements over their predecessor in our experiments. Therefore, in the remainder of this work, we will only present the performance of SimCSE.

RetroMAE pre-training yielded even better results than the SimCSE-Small-E-Czech. Multilingual distillation performed exceptionally, even surpassing OpenAI embeddings on all datasets despite having 6x smaller embeddings.

RetroMAE and distilled models do not exhibit significant enhancements in STS after undergoing SimCSE training. The lack of improvement is attributed to the already favorable spatial properties of the embeddings generated by these models. This pattern is obvious in distilled models, which imitate a model trained with a contrastive objective but using supervised data. However, the situation is not immediately clear for autoencoder models like TSDAE and RetroMAE, as they display similar behavior. This suggests that autoencoders and contrastive learning may have some degree of interchangeability.

Model	F1 score	
	CFD	**CTDC**
Random baseline	23.11 ± 0.49	21.44 ± 0.24
Random-small	25.66 ± 1.21	21.45 ± 0.23
Avg. fastText	64.12 ± 1.41	66.56 ± 0.58
Small-E-Czech	32.38 ± 2.35	25.92 ± 0.56
RetroMAE-small	68.56 ± 1.37	78.18 ± 0.08
Dist-MPNet-ParaCrawl	71.30 ± 1.40	80.18 ± 0.10
Dist-MPNet-CzEng	72.84 ± 1.62	**82.38** ± 0.25
SimCSE-Small-E-Czech	54.78 ± 2.59	50.06 ± 0.45
SimCSE-RetroMAE-small	68.70 ± 1.49	77.32 ± 0.39
TSDAE-Small-E-Czech	66.99 ± 1.24	75.84 ± 0.24
SimCSE-Dist-MPNet-ParaCrawl	71.87 ± 1.06	79.41 ± 0.28
SimCSE-Dist-MPNet-CzEng	72.66 ± 1.35	81.63 ± 0.37
OpenAI Ada Embedding	**75.20** ± 1.12	75.89 ± 0.46

Table 2: Linear probing evaluation of our models on CFD and CTDC.

Linear Probing

Linear probing clarifies a model's internal representations, which are not explicitly discernible based on a position in a vector space, and provides a novel perspective on sentence embeddings, making them suitable for tasks where cosine similarity is inadequate. We train a classifier head on top

of pre-trained sentence embeddings and evaluate its performance on CFD and CTDC, reserving other datasets for the sentence-pair bi-encoder setup using simple cosine similarity. We chose CLS pooling as the standard for all models due to the classification head's adaptability, which reduces disparities between different pooling methods in our tests.

Linear Probing Results As outlined in Table 2, ELECTRA pre-training enhances the performance of linear probing compared to a randomly initialized model, yet it remains less effective than other methods. This may be attributed to the anisotropy problem observed within the semantic spaces of Small-E-Czech embeddings. In other words, the cosine similarity of nearly any meaningful texts is very close to 1, limiting their expressiveness. While the embeddings can be effectively redistributed through full model fine-tuning, the direct utilization of embeddings may prove inadequate, leading to subpar performance.

Contrastive methods, designed to enhance the spatial properties of sentence embeddings, notably also enrich the generated features. This effect helps Small-E-Czech to reduce the performance gap, leading to significant score improvements of about 22 and 24 percentage points on the CFD and CTDC datasets, respectively. However, it still doesn't surpass fastText, which performs notably better in linear-probing than in zero-shot. TSDAE furthers these gains, exceeding fastText. However, such improvements are not observed in our other models after contrastive fine-tuning, where there is insignificant change in performance. RetroMAE-Small consistently generates robust features for linear probing, and bilingual models consistently outperform others, particularly on the CTDC dataset in a linear-probing setup.

OpenAI's embeddings top the sentiment analysis rankings, possibly benefiting from the abundance of data and the task's prevalence. However, for more complex tasks like CTDC, our models outperform them.

Fine-Tuning

In this experiment, we permit models to directly optimize their sentence embeddings for specific tasks, enabling an in-depth assessment of their potential on the DaReCzech, CFD, and CTDC datasets. We continue optimizing the classification head, as we previously did with CFD and CTDC. However, for the DaReCzech training, we exclude a classification head, maintaining it within a bi-encoder setup, with the optimization focused solely on the embedding space to closely align a query with its relevant documents.

We do not fine-tune OpenAI's embeddings as we lack access to the model and control over the fine-tuning process.

Fine-Tuning Results The results in Table 3 reveal that ELECTRA pre-training effectively adapts its sentence embeddings to the given task, achieving commendable performance, confirming the results of Kocián et al. (2022). Despite contrastive learning enhancing fine-tuning performance for tasks involving sentence pair comparisons, like DaReCzech, it can lead to a decline in performance for other tasks, which confirms previous findings of Gao, Yao, and Chen (2021). However, this isn't the case for Small-E-Czech augmented with SimCSE and subsequently fine-tuned on the CTDC dataset, where it improves by another 10 points. This suggests that contrastive learning can still be advantageous for fine-tuning specific types of downstream tasks.

RetroMAE's pre-training demonstrates strong performance in these settings, showing its capacity to learn powerful encodings for any piece of text. Notably, the fine-tuned bilingual models surpass others in this setup, making them a generally excellent choice when both a robust English model and a large bilingual dataset are available.

Other Experiments

Comparison to Existing Czech Base Models

We compared the performance of our distilled models (Dist-MPNet-ParaCrawl, Dist-MPNet-CzEng) with the existing Czech *Base* models: *Czert-b-base-cased* (Sido et al. 2021), *FERNET-C5* (Lehečka and Švec 2021), *RobeCzech* (Straka et al. 2021) and multilingual *LaBSE* (Feng et al. 2022). As shown in Table 4, despite being approximately 8 times smaller than the existing Czech *Base*-sized model, our *Small*-sized models have proven to be highly competitive. The intrinsic evaluation of their embeddings' semantic features (*STS-AVERAGE*, *COSTRA*) revealed that they outperform the majority of BASE-sized systems. Surprisingly, our models perform on par with most BASE-sized models in relevance ranking (DaReCzech) and are only surpassed by the LABSE model, whose embeddings have a semantic origin.

Fine-Tuning Data Size

One of our hypotheses posited that ample data for downstream tasks during model fine-tuning could mitigate differences among initial models. To empirically validate this hypothesis, we leveraged the DaReCzech training dataset, comprising in excess of 1.4 million query-document pairs, and proceeded to train diverse models on its subsets. Specifically, each model underwent training across four instances, each involving random subsets of sizes 1 000; 5 000; 10 000; 100 000; 500 000; and 1 400 000. The aggregate P@10 values, alongside their corresponding standard deviations, are visualised in Figure 1. Evidently, the distinctions among initial models remains discernible despite variations in data size, even following comprehensive fine-tuning with the entire available dataset.

Model Usage in Seznam.cz

Seznam.cz started its proprietary search engine in 2005, initially solely based on lexical matching. In 2021, a semantic branch was integrated, utilizing bi-encoder neural networks, notably Small-E-Czech model. This augmentation extended the search engine's capabilities beyond mere word matching, enabling deeper semantic comprehension of user queries and web content. Consequently, search accuracy and relevance has been significantly improved, providing users with more precise search results. Small-E-Czech has also proven valuable in additional tasks at Seznam.cz, such as query correction (Filip 2021), named-entity recognition (Hávová 2023) or clickbait article detection.

Model	F1 score CFD	F1 score CTDC	P@10 DaReCzech
Random-small	69.67 ± 1.35	40.32 ± 1.40	42.66 ± 0.35
Small-E-Czech	76.94 ± 1.18	58.12 ± 1.52	43.64 ± 0.37
RetroMAE-Small	76.85 ± 1.16	84.58 ± 0.37	45.29 ± 0.34
Dist-MPNet-ParaCrawl	77.42 ± 1.60	86.02 ± 0.12	45.55 ± 0.33
Dist-MPNet-CzEng	**78.73** ± 1.39	85.85 ± 0.21	**45.75** ± 0.34
SimCSE-Small-E-Czech	76.27 ± 1.19	68.33 ± 1.34	44.64 ± 0.38
SimCSE-RetroMAE	76.16 ± 1.53	84.95 ± 0.05	45.26 ± 0.34
TSDAE-Small-E-Czech	75.31 ± 1.00	73.15 ± 0.29	44.84 ± 0.35
SimCSE-Dist-MPNet-ParaCrawl	77.31 ± 1.40	**86.10** ± 0.32	45.66 ± 0.33
SimCSE-Dist-MPNet-CzEng	**78.73** ± 1.43	85.25 ± 1.15	**45.75** ± 0.33

Table 3: Fine-tuning evaluation of our models on CFD, CTDC and DaReCzech datasets.

Model	Costra	STS-Average	CFD	DaReCzech	CTDC
Czert-b-base-cased	**72.08**	74.79	78.73 ± 1.25	45.63 ± 0.34	88.69 ± 0.21
FERNET-C5	67.57	65.46	**82.00** ± 1.43	45.87 ± 0.34	**89.56** ± 0.19
RobeCzech	63.94	69.41	80.54 ± 0.93	45.54 ± 0.31	86.01 ± 0.24
LaBSE	70.63	82.91	79.79 ± 1.07	**46.15** ± 0.34	87.97 ± 0.31
Dist-MPNet-CzEng	71.22	**87.60**	78.73 ± 1.39	45.75 +-0.34	85.85 ± 0.21
Dist-MPNet-ParaCrawl	70.42	84.25	77.42 ± 1.60	45.55 ± 0.33	86.02 ± 0.12

Table 4: Comparison of *Base* Czech models (top four) to two of our best *Small* models (bottom – Dist-MPNet-CzEng, Dist-MPNet-ParaCrawl) on all datasets.

Recognizing the limitations of the Small-E-Czech model, as previously discussed, we have gradually replaced it in our primarily retrieval-based applications with its semantic alternatives, namely the RetroMAE-Small and Dist-MPNet-ParaCrawl models. These models now hold a crucial role, particularly in organic search, featured snippets, and image search, enhancing the overall search experience.

Organic Search

Seznam.cz's search engine utilizes a sequence of cascade stages for organic search, each characterized by varying complexity and time constraints (Kocián et al. 2022). The initial stage operates across the entire document index, demanding rapid processing. Subsequent stages handle documents pre-selected by prior stages, they are allowed longer processing times to retrieve only documents relevant to user queries. Across all these stages, the Small-E-Czech model has consistently maintained a crucial role.

In the final component of the organic search (*stage-2* Catboost (Prokhorenkova et al. 2018) tree), Small-E-Czech-based bi-encoder model trained on the DaReCzech dataset has been employed as an additional feature leading to 1.5 percentage points improvement on the offline test set and a 3.8 percentage points improvement, as assessed by human annotators after deployment in production (measured using P@10) (Kocián et al. 2022). Presently, this model has been substituted with the new RetroMAE-Small model, yielding additional 2 percentage points improvement. This change has led to an improved user experience with our search functionality. Upon analysis, the model primarily enhances results for intricate and information-rich queries.

We are currently engaged in an active effort to replace outdated models in the preceding stages.

Featured Snippets

Since 2022, Seznam.cz introduced featured snippets to its search results, providing direct answers from documents on the search engine results page (SERP) alongside organic results. Two-step architecture is employed: a bi-encoder first pre-selects relevant paragraphs, and a cross-encoder subsequently re-ranks them. In this process, the Small-E-Czech model, enhanced with annotations, plays a key role.

Unlike organic results, featured snippets are considered an optional addition. Hence, it is essential to display only accurate and reliable featured snippets. To ensure this, we establish a threshold, based on our annotated data, ensuring at least 95% accuracy for the snippets. Accordingly, our primary metric is recall, conditioned on the model maintaining a minimum precision of 95%.

Despite the already high recall of our retrieval model, there are instances when the bi-encoder falls short in identifying the appropriate paragraph. Transitioning from the Small-E-Czech model to RetroMAE-Small reduced this error rate by a significant 20%. Consequently, our refined pipeline has resulted in an average 3.5% boost in recall, with a consistent precision of 95%.

Image Search

To accommodate the image queries in Seznam.cz, the engine employs a joint embedding space for both images and texts.

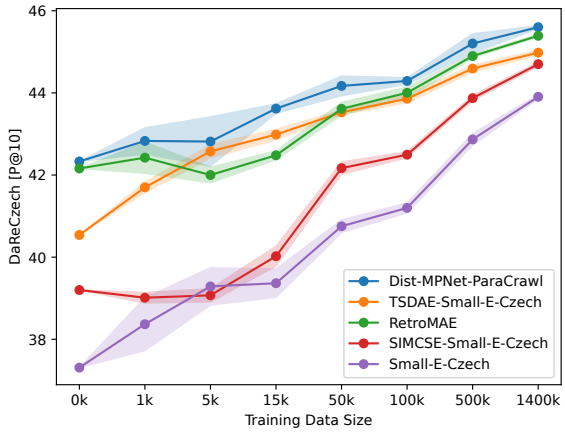

Figure 1: Comparison of different models when fine-tuned with various training data sizes evaluated on the DaReCzech test set. Each data point represents the average P@10 of four models, with standard deviation indicated.

Texts associated with specific images form cohesive clusters in the embedding space, while texts and images lacking semantic relevance remain farther apart. This embedding space is developed through joint training of both text and image encoders, where the model learns to understand the connections between texts and images much like *CLIP* (Radford et al. 2021).

Our ultimate metric for evaluating the quality of the embedding space is the performance of the production image search model, which uses the embedding space similarity of query and image as one of its features. By replacing the Small-E-Czech text encoder with the Dist-MPNet-ParaCrawl model, the relative NDCG (normalized discounted cumulative gain) error rate reduced by 3.2%, similar to the improvement observed when the Small-E-Czech text encoder replaced the original FastText encoder. Notably, this improvement is significant considering that the production model comprises numerous complex features and has limited room for enhancement.

We also assess the quality of the embedding space separately and compare different text encoders. In this comparison, Dist-MPNet-ParaCrawl allows retrieving 7% more relevant images than the previous best Small-E-Czech model.

Deployment Information

In our deployment configuration, both organic and featured snippet's components are computed on the CPU with AVX512 instructions. Our models are stored in the ONNX format and quantized using FP16, with negligible impact on performance. Processing efficiency is maintained through our in-house queuing system, integrated with Kubernetes. Documents and paragraphs are processed offline whenever there is a modification or a new document is identified, while queries are handled in real-time. The average query processing time is approximately 4 milliseconds. Regarding images, similar to queries, computations are conducted online only for queries, and images are embedded using image-specific models developed in-house.

Conclusion

In this work, we trained and evaluated multiple *Small*-sized Czech models for sentence embeddings. Despite their relative small size, these models matched or exceeded the performance of *Base* models across diverse tasks. The intrinsic evaluations proved the models' effectiveness in capturing semantic and syntactic information, while the extrinsic tasks demonstrated their practical use in real-world applications in NLP. We also established that a powerful bilingual model can be derived through multilingual distillation from a proficient English model, outperforming several pre-trained Czech models. The effectiveness of RetroMAE's pre-training and language distillation methods was confirmed in various Seznam.cz tasks. We anticipate our released models could promote wider applications and inspire further exploration into efficient, smaller-scale models for complex language tasks.

Acknowledgments

We thank Barbora Rišová, Martin Dvořák, Martin Habrovec and Dominika Kozlová for experiments done on image-text encoders. We also thank Milan Straka and Jana Straková for their valuable comments on this article.

References

Bañón, M.; Chen, P.; Haddow, B.; Heafield, K.; Hoang, H.; Esplà-Gomis, M.; Forcada, M. L.; Kamran, A.; Kirefu, F.; Koehn, P.; Ortiz Rojas, S.; Pla Sempere, L.; Ramírez-Sánchez, G.; Sarrías, E.; Strelec, M.; Thompson, B.; Waites, W.; Wiggins, D.; and Zaragoza, J. 2020. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *ACL*.

Barančíková, P.; and Bojar, O. 2020. Costra 1.1: An Inquiry into Geometric Properties of Sentence Spaces. In *TSD*.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *TACL*.

Chuang, Y.-S.; Dangovski, R.; Luo, H.; Zhang, Y.; Chang, S.; Soljačić, M.; Li, S.-W.; Yih, W.-t.; Kim, Y.; and Glass, J. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *NAACL*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

Conneau, A.; and Kiela, D. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *LREC*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*.

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2022. Language-agnostic BERT Sentence Embedding. In *ACL*.

Filip, O. 2021. Rychlá oprava dotazů ve vyhledávači pomocí neuronových sítí. https://www.root.cz/clanky/rychla-

oprava-dotazu-ve-vyhledavaci-pomoci-neuronovych-siti/. Accessed: 2023-12-04.

Gao, L.; and Callan, J. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. In *EMNLP*.

Gao, L.; and Callan, J. 2022. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In *ACL*.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*.

Habernal, I.; Ptáček, T.; and Steinberger, J. 2013. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In *WASSA*.

Hávová, M. 2023. Jak Vyhledávání na Seznamu rozpozná jména, příjmení a osobnosti? https://blog.seznam.cz/2023/07/vylepsili-jsme-rozpoznavani-jmen-prijmeni-a-osobnosti-ve-vyhledavani-na-seznamu/. Accessed: 2023-12-04.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Hofstätter, S.; Khattab, O.; Althammer, S.; Sertkan, M.; and Hanbury, A. 2022. Introducing Neural Bag of Whole-Words with ColBERTer: Contextualized Late Interactions using Enhanced Reduction. In *CIKM*.

Khattab, O.; and Zaharia, M. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR*.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *ICLR*.

Kocián, M.; Náplava, J.; Štancl, D.; and Kadlec, V. 2022. Siamese BERT-Based Model for Web Search Relevance Ranking Evaluated on a New Czech Dataset. In *AAAI*.

Kocmi, T.; Popel, M.; and Bojar, O. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *arXiv preprint arXiv:2007.03006*.

Kong, W.; Khadanga, S.; Li, C.; Gupta, S. K.; Zhang, M.; Xu, W.; and Bendersky, M. 2022. Multi-Aspect Dense Retrieval. In *SIGKDD*.

Kral, P.; and Lenc, L. 2018. Czech Text Document Corpus v 2.0. In *LREC*.

Lehečka, J.; and Švec, J. 2021. Comparison of Czech Transformers on Text Classification Tasks. In *SLPS*. Springer.

Liu, F.; Jiao, Y.; Massiah, J.; Yilmaz, E.; and Havrylov, S. 2021. Trans-encoder: unsupervised sentence-pair modelling through self-and mutual-distillations. *arXiv preprint arXiv:2109.13059*.

Liu, J.; Liu, J.; Wang, Q.; Wang, J.; Wu, W.; Xian, Y.; Zhao, D.; Chen, K.; and Yan, R. 2023. RankCSE: Unsupervised Sentence Representations Learning via Learning to Rank. In *ACL*.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, S.; He, D.; Xiong, C.; Ke, G.; Malik, W.; Dou, Z.; Bennett, P.; Liu, T.-Y.; and Overwijk, A. 2021. Less is More: Pretrain a Strong Siamese Encoder for Dense Text Retrieval Using a Weak Decoder. In *EMNLP*.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. arXiv:2210.07316.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *NeurIPS*, 31.

Qu, Y.; Ding, Y.; Liu, J.; Liu, K.; Ren, R.; Zhao, W. X.; Dong, D.; Wu, H.; and Wang, H. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *NAACL*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *EMNLP*.

Sido, J.; Pražák, O.; Přibáň, P.; Pašek, J.; Seják, M.; and Konopík, M. 2021. Czert–Czech BERT-like Model for Language Representation. *arXiv preprint arXiv:2103.13031*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *NeurIPS*.

Straka, M.; Náplava, J.; Straková, J.; and Samuel, D. 2021. RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. In *TSD*.

Svoboda, L.; and Brychcín, T. 2018. Czech dataset for semantic textual similarity. In *TSD*. Springer.

Wang, K.; Reimers, N.; and Gurevych, I. 2021. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoderfor Unsupervised Sentence Embedding Learning. In *EMNLP*.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2023. SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval. In *ACL*.

Wu, X.; Gao, C.; Lin, Z.; Han, J.; Wang, Z.; and Hu, S. 2022. InfoCSE: Information-aggregated Contrastive Learning of Sentence Embeddings. In *EMNLP*.

Xiao, S.; Liu, Z.; Shao, Y.; and Cao, Z. 2022. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In *EMNLP*.

Zhao, W. X.; Liu, J.; Ren, R.; and Wen, J.-R. 2022. Dense Text Retrieval based on Pretrained Language Models: A Survey. *arXiv preprint arXiv:2211.14876*.

Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *CCL*.