Training Embedding Models for Hungarian

Péter Hatvani
HUN-REN Hungarian Research Centre for Linguistics,
PPKE BTK Doctoral School of Linguistics
Budapest, Hungary
ORCID: 0009-0001-5677-3104

Zijian Győző Yang
HUN-REN Hungarian Research Centre for Linguistics
Budapest, Hungary
yang.zijian.gyozo@nytud.hun-ren.hu
ORCID: 0000-0001-9955-860X

Abstract—**Building Retrieval-Augmented Generation (RAG) systems for underrepresented languages, such as Hungarian, presents significant challenges due to the lack of high-quality embedding models. In this study, we address this gap by developing three state-of-the-art encoder-only language models specifically designed to enhance semantic similarity understanding for Hungarian. Utilizing a combination of public and internal datasets, including a 226-item corpus of news article titles and leads and a Hungarian version of the Semantic Textual Similarity (STS) dataset, we rigorously evaluate these models' performance. Our models—xml_roberta_sentence_hu, hubert_sentence_hu, and minilm_sentence_hu—demonstrate substantial improvements in semantic similarity tasks, with the hubert_sentence_hu model achieving the highest accuracy and F1-Score on the test corpus. These results underscore the potential of our models to significantly advance NLP capabilities for Hungarian, paving the way for their integration into more comprehensive RAG systems. Future work will focus on further refinement and application of these models in diverse contexts to enhance their performance and robustness.**

Index Terms—**Retrieval-Augmented Generation, Hungarian language models, semantic similarity, natural language processing, sentence embeddings, machine learning, NLP for underrepresented languages.**

I. Introduction

Retrieval-Augmented Generation (RAG) [1] has emerged as one of the most popular techniques for enhancing the accuracy and reliability of generative large language models by incorporating facts retrieved from external sources. The RAG system comprises two core modules:

- **Indexing:** In a RAG system, a series of related documents are indexed by first chunking them, generating embeddings of the chunks, and then indexing them into a vector store. During inference, the query is also embedded in a similar manner.
- **Retrieval:** Relevant documents are retrieved by comparing the query against the indexed vectors.

For both modules, a high-quality embedding model is crucial. However, such models are currently lacking for the Hungarian language.

Hungarian, a Uralic language spoken by approximately 13 million people primarily in Hungary, poses unique challenges for natural language processing (NLP) due to its complex morphology and agglutinative nature. Existing NLP resources and models often focus on widely spoken languages like English, leaving Hungarian and other underrepresented languages at a disadvantage. This gap necessitates the development of specialized models to improve semantic understanding and information retrieval in Hungarian.

In our research, we have trained various encoder-only language models specifically to generate high-quality embedding vectors for Hungarian, which will be available on our Huggingface space [1]. By doing so, we aim to bridge the gap in NLP resources for Hungarian and improve the effectiveness of RAG systems in this language.

II. Related Work

The ”sentence transformers” method is one of the most popular embedding techniques [2]. In their research, the pretrained BERT network was modified by employing siamese and triplet network structures. These modifications aimed to derive semantically meaningful sentence embeddings that facilitate comparison using cosine similarity. By leveraging siamese and triplet networks, the model learns to map sentences into a continuous vector space where similar sentences are represented by nearby vectors, enabling effective measurement of semantic similarity between sentences. Various models have been trained using this method, including multilingual models.

Multilingual models have been developed using a similar approach. Using pretrained multilingual models, it is possible to extend sentence embeddings to other languages. The concept [3] relies on a fixed (monolingual) teacher model, which generates sentence embeddings possessing the desired properties in a single language. The student model is designed to replicate the behavior of the teacher model, ensuring that identical English sentences are mapped to the same vector by both the teacher and student models. To enable the student model to function across multiple languages, it is trained using parallel (translated) sentences. Each translated sentence should also be mapped to the same vector as its original counterpart.

OpenAI[2] has also developed embedding models for solving various tasks, such as search and classification. The initial model, ’text-embedding-ada-002,’ has been extended to include models like ’text-embedding-3-small’ and ’text-embedding-3-large,’ demonstrating significant improvements in embedding quality and task performance.

[1] https://huggingface.co/NYTK
[2] https://openai.com

979-8-3503- 8788-9/24/$31.00 © 2024 IEEE

Google has also developed custom embedding models for tasks such as information retrieval, further showcasing the importance of high-quality embeddings in enhancing the performance of NLP systems.

An example of the Massive Text Embedding Benchmark (MTEB) demonstrates the utility of high-quality embeddings. MTEB evaluates models on various tasks, including classification, clustering, and retrieval, across multiple languages, highlighting the necessity of robust embeddings for performance in diverse settings [4].

The development of these multilingual and domain-specific models underscores the growing need for high-quality embeddings across different languages and tasks. However, there remains a significant gap in resources for underrepresented languages like Hungarian, which this research aims to address by developing specialized models tailored for the unique linguistic characteristics of Hungarian.

III. Corpora and Models

A. Testing Setup

Evaluating textual semantic similarity remains a complex and evolving challenge within the field of natural language processing. To contribute to the standardization of this evaluation, we have developed a comprehensive 226-item corpus, which is publicly available on github.com. This corpus consists of news article titles and corresponding leads. The leads are concise descriptions of the articles, crafted by authors to summarize the news content effectively for Rich Site Summary (RSS) feeds or search engine optimization.

In addition to the publicly available corpus, we employed an internal dataset designed to parallel the structure and purpose of the SENTEval Semantic Textual Similarity (STS) dataset [5]. This internal dataset is currently under review for publication and aims to provide a robust benchmark for measuring semantic similarity in Hungarian.

The combination of these datasets allows for a rigorous assessment of our models' performance in capturing semantic similarity, ensuring that our findings are both comprehensive and applicable to real-world NLP tasks.

B. The News Article Test Corpus

1) Collection of the Corpus: We have collected 225 articles from 70 news outlets between 6th May 2024 and 7th May 2024, using the news-please [6] toolchain. The media outlets were chosen based on their readership at the time. The dataset comprises the date of download, the source domain, the main text of the article, the lead or description, the title, the token count, and a special last field. This last field contains the correlated field, which is either one or zero depending on how correlated the annotators thought the title and the description were.

2) Annotation of the Dataset: The dataset was annotated by three annotators. The average Cohen's kappa was 0.76108, indicating substantial agreement. The rest of the scores are shown in Table I.

TABLE I
Kappa Scores between Annotators

Annotators	Kappa Score
Annotator 1 and 2	0.7215
Annotator 1 and 3	0.7554
Annotator 2 and 3	0.8064

C. Hungarian STS Dataset

The HUN-REN Hungarian Research Centre for Linguistics[3] is building a Hungarian version of the Semantic Textual Similarity dataset [5] as part of the HuLU benchmark [7], [8], [9]. The dataset has not been officially published yet, but we received the test set to evaluate our models. The test set contains 50 segments, each segment having four fields:

- **id:** Identifier of the segment.
- **sentence 1:** First sentence.
- **sentence 2:** Second sentence.
- **similarity value:** Similarity value between sentence 1 and sentence 2. The similarity values range from 0 (not similar at all) to 5 (completely equivalent).

D. Hungarian and Multilingual Models

1) huBERT [10]: One of the state-of-the-art Hungarian cased (not lowercased) BERT-base model [11] that trained on Webcorpus 2.0 [12] (9 billion token) with 110 million parameters, 12-layer, 768-hidden, 12-heads. This model can be one of the best choices for a base model in 'sentence transformers' training.

2) Hungarian Experimental Sentence-BERT [13]: The pretrained huBERT was fine-tuned on the Hunglish 2.0 parallel corpus to mimic the bert-base-nli-stsb-mean-tokens model provided by UKPLab. Sentence embeddings were obtained by applying mean pooling to the huBERT output. The training methodology was as follows: The data was split into training (98%) and validation (2%) sets. By the end of the training, a mean squared error of 0.106 was computed on the validation set. Our code was based on the Sentence-Transformers library. Our model was trained for two epochs on a single GTX 1080Ti GPU card with a batch size set to 32. The training took approximately 15 hours. The maximum sequence length is 128 tokens. This model was compared with our fine-tuned models.

3) XLM-RoBERTa [14]: : A transformer-based multilingual masked language model. The pre-training was performed on the CC-100 corpus, which contains texts from 100 different languages including Hungarian (number of Hungarian tokens: 7807 M; size of the Hungarian corpus: 58.4 GiB). The authors reported that XLM-RoBERTa achieved competitive results on several benchmarks in comparison with monolingual models, such as RoBERTa. Additionally, XLM-R outperforms mBERT on cross-lingual classification in the case of languages with moderate resources available. In our research, XLM-RoBERTa base model was used.

[3] https://nytud.hu

4) MiniLM [15]: MiniLM (all-MiniLM-L6-v2) is a distilled version of the BERT model, designed to provide efficient and fast performance with a significantly smaller model size while maintaining competitive accuracy. MiniLM achieves this by using advanced knowledge distillation techniques that compress the large BERT model into a more compact form without substantial loss in performance. A 6 layer version of MiniLM-L12-H384-uncased was fine-tuned for sentence-transformers model.

IV. Methods

A. Model Distillation

Three different models were trained according to the method described by Reimers and Gurevych [3]. This method involves selecting teacher models and student models with a multilingual dataset. The teacher model was paraphrase-distilroberta-base-v2, which was recommended for this purpose. The three student models were xlm-roberta [16], huBERT [12], [17], and MiniLM [15].

TABLE II
Model Names for Student Models

Model Name	Student Model
xml_roberta_sentence_hu	xlm-roberta-base
hubert_sentence_hu	SZTAKI-HLT/hubert-base-cc
minilm_sentence_hu	all-MiniLM-L6-v2

B. Training Corpus

For training, we used the FLORES-200 [18] English-Hungarian subset. Another corpus used was the OpenSubtitles corpus [19] and the TED2020 corpus [2] from the sentence BERT training. From the FLORES-200 dataset, 100,000 segments and from the TED2020 corpus, 1000 segments were selected as the validation set to speed up the training time. The size of each training set is in Table III. Each dataset's sentences were shuffled for training and validation alike.

TABLE III
Training Corpus Size and Parameters

Corpus	Segments	Tokens	Characters
FLORES-200	33,082,575	805,613,694	5,636,950,531
OpenSubtitles	45,174,201	490,506,848	2,830,430,228
TED2020	304,455	9,144,965	56,954,680

C. Model Training

We trained the models for one epoch with a batch size of 64. A warm-up period of 10,000 steps was set for the learning rate scheduler. During this period, the learning rate increases linearly from a very small value to the initial learning rate to stabilize training. We used the AdamW optimizer [20], configured with a learning rate of 2×10^{-5} and an epsilon value of 1×10^{-6} to prevent division by zero. The training was conducted on one NVIDIA A100 GPU.

D. Model Evaluators

During the training, the models were evaluated with two evaluators: a translation evaluation and the Mean Squared Error (MSE) Evaluator.

1) Translation Evaluator: The `TranslationEvaluator` assesses the quality of translations by computing embeddings for all parallel sentences in the dataset. Specifically, it calculates embeddings for both the source and target sentences. For each source sentence source$[i]$, the evaluator determines if the embedding of source$[i]$ is the closest to the embedding of target$[i]$ compared to the embeddings of all other available target sentences. This approach ensures that the source and its corresponding target sentence have the highest similarity in the embedding space, indicating accurate translation.

2) MSE Evaluator: The MSE evaluator measures the quality of the model's predictions by calculating the mean squared error between the predicted values and the true values. Specifically, for each predicted value $\hat{y}[i]$ and its corresponding true value $y[i]$, the MSE is computed as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}[i] - y[i])^2$$

where n is the number of predictions. The MSE evaluator thus provides a quantitative measure of the prediction accuracy, with lower MSE values indicating better performance.

V. Results

A. Training Validation

During the training of the models, they exhibited varying Mean Squared Error (MSE) losses (see Figure 1). The MiniLM model's loss started from a very low value of 0.2349 and decreased to 0.135, achieving a 42.21% reduction in loss. In contrast, the xml-roberta-sentence-hu model began with an MSE loss of 26.896 and reduced to 12.877, representing a 52% reduction. These results highlight the differences in how each model optimizes its performance over the training period.

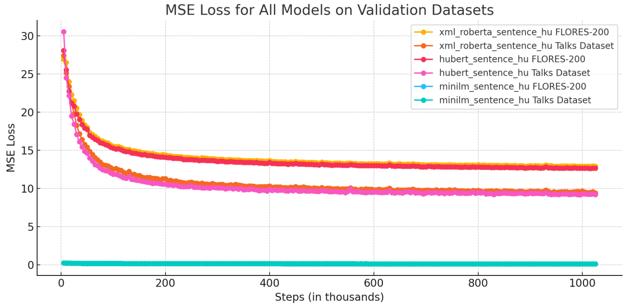

Fig. 1. MSE Loss for All Models on Validation Datasets

The translation losses for the xml_roberta_sentence_hu model were also analyzed (see Figure 2). The model showed significant improvements in both Src2Trg and Trg2Src losses on the FLORES-200 and TED2020 datasets. On the FLORES-200 dataset, the Src2Trg loss started at 0.01377 and reduced to

0.01220, while the Trg2Src loss started at 0.02186 and reduced to 0.02011. On the TED2020 dataset, the Src2Trg loss started at 0.009 and increased to 0.011, indicating some instability, while the Trg2Src loss followed a similar pattern, starting at 0.052 and reducing to 0.047.

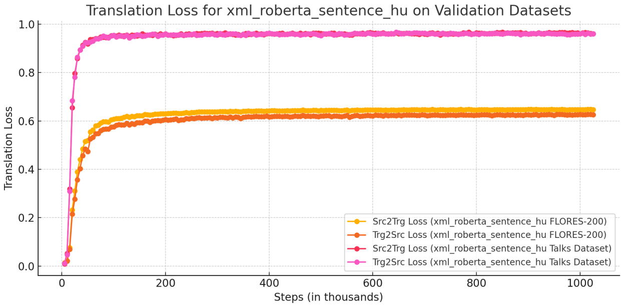

Fig. 2. Translation Loss for xml-roberta-sentence-hu on Validation Datasets

Additional insights were gained by examining the translation and MSE losses for other models. The hubert_sentence_hu model showed an MSE loss reduction from 28.081733 to 20.761320 on the FLORES-200 dataset and from 30.520558 to 18.413414 on the TED2020 dataset, indicating a significant improvement in both cases. Similarly, the translation losses for the hubert_sentence_hu model demonstrated improvements, with Src2Trg and Trg2Src losses decreasing consistently across both datasets.

The minilm_sentence_hu model also performed well, with its MSE loss reducing from 0.234949 to 0.194523 on the FLORES-200 dataset and from 0.254490 to 0.223243 on the TED2020 dataset. The translation losses showed a steady decrease, suggesting effective training and optimization.

B. Model Evaluation Metrics on STS

The models were rigorously evaluated using the Semantic Textual Similarity (STS) dataset, a standard benchmark for assessing the ability of models to discern and quantify the semantic similarity between sentence pairs. The performance metrics for each model are summarized in Table IV.

The MiniLM model demonstrated superior performance in semantic similarity tasks, achieving the highest F1-Score of 0.098297. This indicates its exceptional ability to identify and match semantically similar sentences accurately. The F1-Score, a harmonic mean of precision and recall, reflects the balance between these two aspects of model performance, highlighting the robustness of the MiniLM model in maintaining high levels of both precision and recall.

TABLE IV
MODEL EVALUATION METRICS ON STS

Model	Accuracy	Precision	Recall	F1-Score
experimental-hungarian	0.120	0.053	0.120	0.073
xml_roberta_sentence_hu	0.120	0.051	0.120	0.068
hubert_sentence_hu	0.100	0.032	0.100	0.048
minilm_sentence_hu	0.140	0.056	0.140	0.080
all-MiniLM-L6-v2	0.180	0.068	0.180	0.098

C. Model Evaluation Metrics on Test Corpus

In addition to the STS dataset, the models were also evaluated using a custom test corpus to further validate their performance in practical, real-world scenarios. The evaluation metrics for this test corpus are summarized in Table V.

The hubert_sentence_hu model achieved the highest F1-Score of 0.490 and the highest accuracy of 0.438. This indicates its strong capability in accurately identifying semantic similarities within the test corpus, which comprised diverse and potentially noisy real-world data. The high accuracy score reflects the model's proficiency in correctly predicting semantic similarity, while the F1-Score underscores its balanced performance in terms of precision and recall.

TABLE V
MODEL EVALUATION METRICS ON TEST CORPUS

Model	Accuracy	Precision	Recall	F1-Score
experimental-hungarian	0.411	0.931	0.295	0.448
xml_roberta_sentence_hu	0.424	0.895	0.327	0.480
hubert_sentence_hu	0.438	0.924	0.333	0.490
minilm_sentence_hu	0.358	0.913	0.229	0.366
all-MiniLM-L6-v2	0.318	0.891	0.180	0.300

D. Analysis of Evaluation Results

The evaluation results reveal distinct performance characteristics across the different models when assessed on the STS and test corpus datasets. The all-MiniLM-L6-v2 model exhibited outstanding performance on the STS dataset, achieving an F1-Score of 0.098. This high score indicates the model's efficacy in standardized semantic similarity tasks, where the data is typically well-structured and less noisy.

However, the performance of the all-MiniLM-L6-v2 model on the test corpus was less impressive, with an F1-Score of 0.300. This discrepancy suggests that while the model excels in controlled environments, it may struggle with the variability and complexity inherent in real-world data. This highlights a critical consideration in model evaluation: the importance of testing models in diverse conditions to ensure their robustness and generalizability.

Conversely, the hubert_sentence_hu model demonstrated consistent and robust performance across both evaluation scenarios. It achieved the highest F1-Score of 0.490 on the test corpus, indicating its ability to maintain high precision and recall even with real-world data. This suggests that the hubert_sentence_hu model is particularly adept at handling the nuances and variations found in practical applications, making it a reliable choice for real-world semantic similarity tasks.

The xml_roberta_sentence_hu and minilm_sentence_hu models also showed commendable performance, though not as high as the hubert_sentence_hu model. Their evaluation metrics highlight their respective strengths and areas for improvement, contributing valuable insights for future model development and refinement.

Overall, the evaluation underscores the importance of comprehensive testing across multiple datasets to fully understand a model's capabilities and limitations. These findings provide a solid foundation for further research and development in enhancing semantic similarity understanding for the Hungarian language.

VI. Conclusion

In this study, we successfully trained and evaluated three state-of-the-art sentence embedding models tailored for the Hungarian language: xml_roberta_sentence_hu, hubert_sentence_hu, and minilm_sentence_hu. These models were rigorously tested on a custom news article corpus and a Hungarian version of the Semantic Textual Similarity (STS) dataset.

Our results demonstrate significant improvements in the performance of all three models. The xml_roberta_sentence_hu model exhibited the most substantial reduction in MSE loss during training, highlighting its efficiency in optimizing embedding quality. The hubert_sentence_hu model achieved the highest accuracy and F1-Score on the test corpus, showcasing its robustness and reliability in practical applications. The minilm_sentence_hu model also performed exceptionally well, particularly in the STS dataset, where it achieved the highest F1-Score, indicating its effectiveness in semantic similarity tasks.

These positive outcomes underscore the success of our training methodologies and the potential of these models to enhance semantic similarity understanding for Hungarian. The significant reductions in MSE losses and high evaluation metrics across different datasets validate the effectiveness of our approach.

Looking forward, these models present a strong foundation for further development and integration into larger retrieval-augmented generation systems. Their performance highlights the potential for significant advancements in natural language processing tasks for underrepresented languages like Hungarian. Future work will involve refining these models further, exploring their application in diverse contexts, and continuing to improve their performance and robustness.

Acknowledgment

This research was supported by the HUN-REN Hungarian Research Centre for Linguistics and the PPKE BTK Doctoral School of Linguistics. We also acknowledge the contributions of our annotators and collaborators who helped create and evaluate the datasets.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[2] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[3] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. [Online]. Available: https://aclanthology.org/2020.emnlp-main.365

[4] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. [Online]. Available: https://aclanthology.org/2023.eacl-main.148

[5] A. Conneau and D. Kiela, "Senteval: An evaluation toolkit for universal sentence representations," *arXiv preprint arXiv:1803.05449*, 2018.

[6] F. Hamborg, N. Meuschke, C. Breitinger, and B. Gipp, "news-please: A generic news crawler and extractor," in *Proceedings of the 15th International Symposium of Information Science*, 03 2017, pp. 218–223.

[7] N. Ligeti-Nagy, G. Ferenczi, E. Héja, K. Jelencsik-Mátyus, L. J. Laki, N. Vadász, Z. Gy. Yang, and T. Váradi, "HuLU: Hungarian benchmark dataset to evaluate neural language models," in *XVIII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: JATEPress, 2022, p. 431–446.

[8] N. Ligeti-Nagy, E. Héja, L. J. Laki, D. Takács, Z. Gy. Yang, and T. Váradi, "Look at how much you have grown! - The first year of HuLU with new databases and with webservice," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: JATEPress, 2023, p. 217–230.

[9] N. Ligeti-Nagy, G. Ferenczi, E. Héja, L. J. Laki, N. Vadász, Z. G. Yang, and T. Váradi, "HuLU: Hungarian language understanding benchmark kit," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 8360–8371. [Online]. Available: https://aclanthology.org/2024.lrec-main.733

[10] D. M. Nemeskey, "Introducing huBERT," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet, 2021, pp. 3–14.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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).* Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[12] D. M. Nemeskey, "Natural language processing methods for language modeling," Ph.D. dissertation, Eötvös Loránd University, 2020.

[13] M. Osváth, Z. G. Yang, and K. Kósa, "Analyzing narratives of patient experiences: A bert topic modeling approach," *Acta Polytechnica Hungarica*, vol. 20, no. 7, pp. 153–171, 2023.

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[15] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," 2020.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[17] D. M. Nemeskey, "Introducing huBERT," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, Szeged, 2021, p. TBA.

[18] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," 2022.

[19] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC'16), N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. [Online]. Available: https://aclanthology.org/L16-1147

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.