AAdaM at SemEval-2024 Task 1: Augmentation and Adaptation for Multilingual Semantic Textual Relatedness

Miaoran Zhang[1] **Mingyang Wang**[2,3] **Jesujoba O. Alabi**[1] **Dietrich Klakow**[1]
[1]Saarland University, Saarland Informatic Campus
[2]Bosch Center for AI, [3]LMU Munich
mzhang@lsv.uni-saarland.de

Abstract

This paper presents our system developed for the SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian Languages. The shared task aims at measuring the semantic textual relatedness between pairs of sentences, with a focus on a range of underrepresented languages. In this work, we propose using machine translation for data augmentation to address the low-resource challenge of limited training data. Moreover, we apply task-adaptive pre-training on unlabeled task data to bridge the gap between pre-training and task adaptation. For model training, we investigate both full fine-tuning and adapter-based tuning, and adopt the adapter framework for effective zero-shot cross-lingual transfer. We achieve competitive results in the shared task: our system performs the best among all ranked teams in both subtask A (supervised learning) and subtask C (cross-lingual transfer).[1]

1 Introduction

Semantic Textual Relatedness (STR) measures the closeness of meaning between two linguistic units, such as a pair of words or sentences (Budanitsky, 1999; Mohammad and Hirst, 2012). For example, one can easily tell that *"I like playing games"* is more semantically related to *"The game is fun"* rather than *"The weather is good"*, which largely depends on their lexical semantic relation and topic consistency. Semantic Textual Similarity (STS), a closely related concept, indicates whether two units have a paraphrasing relation. The difference between these two concepts is clarified in Abdalla et al. (2023): while similar pairs are also related, the reverse is not necessarily true.

In stark contrast to the extensive research on STS (Gao et al., 2021; Chuang et al., 2022; Zhang et al., 2022; Seonwoo et al., 2023), exploration of STR lags behind and predominantly focuses on English (Marelli et al., 2014; Abdalla et al., 2023), mainly due to the lack of datasets. To close this gap, the SemEval-2024 Task 1: Semantic Textual Relatedness (Ousidhoum et al., 2024b) is proposed to encourage STR research on 14 African and Asian languages. The shared task consists of 3 subtasks: supervised (subtask A), unsupervised (subtask B), and cross-lingual (subtask C).

In this paper, we present our system AAdaM (Augmentation and Adaptation for Multilingual STR) developed for subtask A and C. Our system adopts a cross-encoder architecture which takes the concatenation of a pair of sentences as input and predicts the relatedness score through a regression head (Devlin et al., 2019). As the provided task data for non-English languages is relatively limited, we perform data augmentation for these languages via machine translation. To better adapt a pre-trained model to the STR task, we apply task-adaptive pre-training (Gururangan et al., 2020) which has shown effectiveness on many tasks (Xue et al., 2021; Wang et al., 2023). For subtask A, we explore full fine-tuning and adapter-based tuning (Houlsby et al., 2019) combined with previously mentioned techniques. Additionally, we use the adapter framework MAD-X (Pfeiffer et al., 2020) for cross-lingual transfer in subtask C.

We select the best model based on the performance on development sets for the final submission, and our system achieves competitive results on both subtasks. In subtask A, our system ranks first out of 40 teams on average, and performs the best in Spanish. In subtask C, our system ranks first among 18 teams on average, and achieves the best performance in Indonesian and Punjabi.

2 SemRel Dataset

To encourage STR research in the multilingual context, Ousidhoum et al. (2024a) introduce SemRel, a new STR dataset annotated by native speakers, covering 14 languages from 5 distinct lan-

[1]Our code: https://github.com/uds-lsv/AAdaM

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 800–810

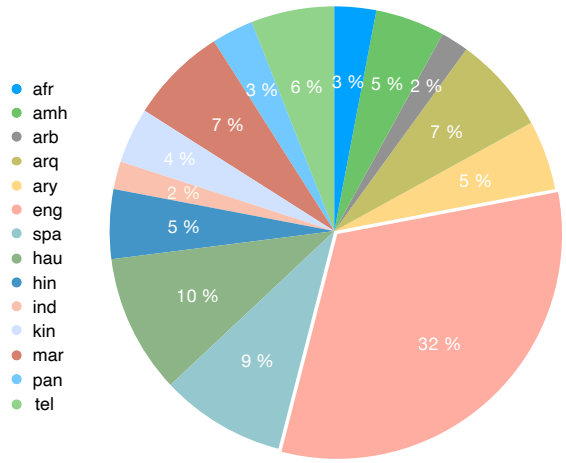


Figure 1: SemRel data distribution across languages.

guage families. These languages are mostly spoken in Africa and Asia, and many of them are under-represented in natural language processing resources. As shown in Figure 1, the data sizes vary widely from language to language constrained by the availability of resources. Notably, English data comprises 32% of the whole dataset and surpasses other languages by a large margin.

3 System Overview

Our system employs a *cross-encoder* architecture, which takes the concatenation of a pair of sentences as input and predicts the relatedness score through a regression head. Compared to bi-encoders (Reimers and Gurevych, 2019), which extract individual sentence representations and then compare them using cosine similarity, cross-encoders generally perform better, at the cost of increased inference latency (Humeau et al., 2020). We select cross-encoder because of its superior performance (see Appendix A), and leave the exploration of an efficient alternative as future work.

The core techniques underlying our system are (i) **data augmentation** using machine translation (§3.1), and (ii) **task-adaptive pre-training** on unlabeled task data (§3.2). We explore two training paradigms for supervised learning combined with the aforementioned techniques, i.e., **fine-tuning** and **adapter-based tuning** (§3.3), and the latter is also employed for cross-lingual transfer (§3.4).

3.1 Data Augmentation

Data augmentation (DA) serves as a widely used strategy to mitigate data scarcity in low-resource languages (Hedderich et al., 2021; Feng et al., 2021). Inspired by work on DA with machine translation (Hu et al., 2020; Amjad et al., 2020), we create additional training data for non-English languages by translating from various English sources, as illustrated below.

SemRel translation. As English data occupies a significant portion of the entire SemRel dataset, we perform augmentation by translating the English subset to other target languages.

STS-B translation. STS-B (Cer et al., 2017), a semantic similarity dataset, is highly relevant to STR, and we translate the STS-B training set in English to other target languages.

It worth noting that using translations as data augmentation yields a mixed data quality. For instance, the translation process may introduce artifacts that reduce data validity. Additionally, the concepts of "similarity" and "relatedness" are relevant but not equivalent, leading to a mismatch in their annotated scores. To leverage data in varied qualities, Zhu et al. (2023) shows that a two-phase approach is beneficial, in which the model is trained on noisy data first and then trained on clean data. Our training procedure follows this two-phase scheme: (i) training the model on augmented data as a *warmup*, and (ii) subsequently training the model on the original task data.

3.2 Task-Adaptive Pre-training

Pre-trained language models (PLMs) are trained on massive text corpora with self-supervision objectives for general purposes (Devlin et al., 2019; Liu et al., 2019). To better adapt PLMs to downstream tasks, Gururangan et al. (2020) propose task-adaptive pre-training (TAPT), i.e., continued pre-training on task-specific unlabeled data, and show that it can effectively improve downstream task performance. We integrate this strategy into our system, wherein we conduct masked language modeling (MLM) on unlabeled task data for a given target language before initiating any supervised training.

3.3 Fine-tuning vs. Adapter-based Tuning

Fine-tuning is the conventional approach to adapt general-purpose PLMs to downstream tasks. It updates all model parameters for each task, leading to inefficiency with the ever-increasing model scales and number of tasks. Recently, many works focus on introducing lightweight alternatives to improve parameter efficiency (Lester et al., 2021; Hu et al., 2022; He et al., 2022). For example, adapter-based

Model Tuning	TAPT	Warmup	arq	amh	eng	hau	kin	mar	ary	spa	tel
FINE-TUNING	✗	✗	52.96	87.70	83.07	78.91	68.59	85.23	88.26	73.83	84.90
	✗	SemRel	55.96	87.86	/	79.87	70.06	85.51	**88.59**	72.93	85.38
	✗	STS-B	62.05	88.50	**84.31**	79.86	69.78	86.48	86.97	73.33	85.15
	✓	✗	65.70	88.03	82.79	79.41	67.03	84.88	88.50	70.47	83.84
	✓	SemRel	66.74	85.58	/	**80.73**	71.29	85.74	87.01	73.37	**85.77**
	✓	STS-B	**68.25**	**88.72**	83.01	78.95	69.38	85.26	87.07	73.50	84.66
ADAPTER TUNING	✗	✗	55.44	87.01	82.96	78.23	70.45	84.62	86.43	72.62	84.51
	✗	SemRel	59.58	87.66	/	79.15	70.56	86.54	86.88	74.90	84.88
	✗	STS-B	62.83	87.63	82.97	80.29	82.01	87.18	87.53	74.18	84.17
	✓	✗	58.81	85.61	82.74	78.40	70.48	84.56	85.78	72.15	84.34
	✓	SemRel	58.47	87.57	/	79.78	71.67	**87.24**	87.35	**76.65**	85.69
	✓	STS-B	59.58	87.40	82.32	79.22	**73.04**	87.12	87.22	73.22	83.70

Table 1: Subtask A performance on development sets (Spearman's correlation $\times 100$). SemRel: warmup by training on SemRel translations; STS-B: warmup by training on STS-B translations. We underline the best performance of fine-tuning and adapter-based tuning, and **bold** the best performance across all variants.

tuning (Houlsby et al., 2019) only updates small modules known as adapters inserted between the layers of PLMs while keeping the remaining parameters frozen. In particular, it has shown impressive performance in cross-lingual transfer (Pfeiffer et al., 2020; Ansell et al., 2021; Pfeiffer et al., 2022).

We explore both fine-tuning and adapter-based tuning to compare their effectiveness on multilingual STR. For fine-tuning, we update all model parameters at each stage, namely the TAPT stage, the warmup stage and the final training stage using the original task data. For adapter-based tuning, we utilize the MAD-X framework (Pfeiffer et al., 2020) which consists of language-specific adapters and task-specific adapters. The language adapters are pre-trained with an MLM objective on unlabeled monolingual corpora. To this end, we collect open-source data from the Leipzig Corpus Collection (Goldhahn et al., 2012) for pre-training.[2] The task adapters are trained on labeled task-specific data (augmented or original), while keeping the language adapters fixed. Note that when applying TAPT, only language adapters are updated. In subtask A, we apply fine-tuning and adapter-based tuning in combination with TAPT and warmup techniques, and select the best model based on the performance on development sets.

3.4 Cross-lingual Transfer with Adapters

The high modularity of MAD-X enables efficient zero-shot cross-lingual transfer. During inference, we simply replace the source language adapter with the *target language adapter* while retaining the *source task adapter*. This task adapter has been trained on labeled data from the source language, without prior exposure to the target language.[3] A crucial challenge for cross-lingual transfer lies in source language selection, as improper sources may lead to negative results (Lange et al., 2021). To determine the best source language, we explore the following metrics to rank sources: (1) linguistic distance (Littell et al., 2017), (2) token overlap (Wu and Dredze, 2019), and (3) development set performance.[4] Results in Appendix C demonstrate that development set performance serves as the most reliable indicator of transfer performance. For subtask C, we select the optimal source from the adapters trained in subtask A based on their performance on development sets.

4 Experimental Setup

Model. Our backbone model is AfroXLMR-large-61L (Adelani et al., 2024), adapted from XLM-R (Conneau et al., 2020) through multilingual adaptive fine-tuning (Alabi et al., 2022). We use NLLB (`nllb-200-distilled-600M`) (Team et al., 2022) to translate from English resources to other languages as data augmentation.

Implementation. All experiments are conducted on a single NVIDIA A100 GPU with a batch size of 16. For MLM, we set the learning rate to 5e-5

[2]Details are provided in Appendix B.

[3]Note that when transferring from any other language to English, we ensure that the source task adapter has not been trained on augmented data translated from English resources, thereby eliminating the effect of data leakage.

[4]The existence of development sets is not realistic in the true zero-shot scenario, and we leave further discussion to the Limitations section.

Model	arq	amh	eng	hau	kin	mar	ary	spa	tel	Avg.↑
Overlap♢	40.	63.	67.	31.	33.	62.	63.	67.	70.	55.11
LaBSE♢	60.	85.	83.	69.	72.	88.	77.	70.	82.	76.22
PALI	67.88	**88.86**	**86.00**	**76.43**	81.34	**91.08**	**86.26**	72.38	86.43	**81.85**
king001	**68.23**	88.78	84.30	74.72	**81.69**	89.68	85.97	72.12	85.34	81.20
NRK	67.36	86.42	83.29	67.20	75.69	87.93	82.70	68.99	83.42	78.11
saturn	57.77	84.51	-	69.91	75.53	87.28	79.77	-	**87.34**	-
AAdaM (Ours)	66.23	86.71	84.84	72.36	77.91	89.43	83.50	**74.04**	84.77	79.98

Table 2: Subtask A performance on test sets (Spearman's correlation ×100). ♢: baseline results from Ousidhoum et al. (2024a). We **bold** the best performance across submitted systems.

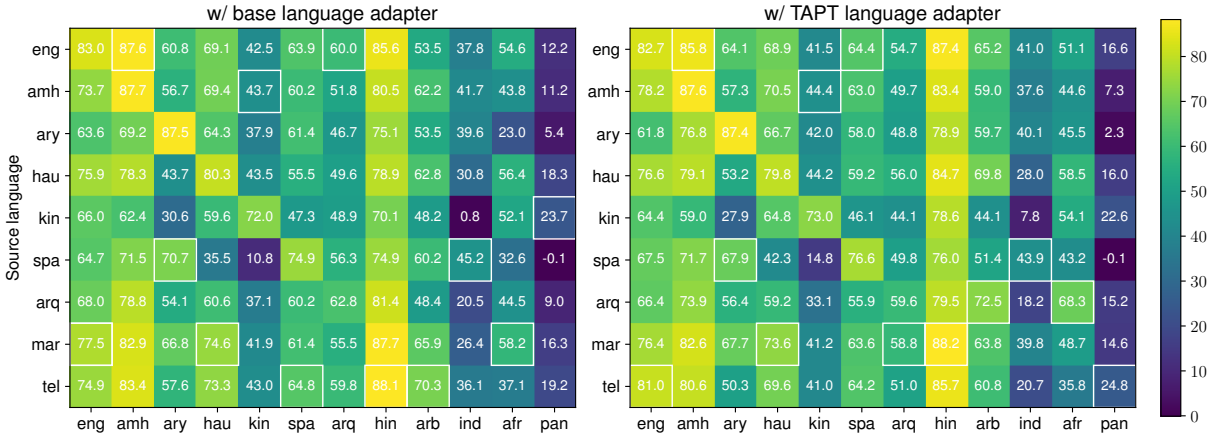


Figure 2: Subtask C performance on development sets (Spearman's correlation ×100) using different types of language adapters. Boxes highlight the optimal performances for each target language, and we select the best source for final submission.

and train models for 10 epochs. For fine-tuning, we conduct a grid-search of learning rate from {2e-5, 5e-5} on SemRel development sets and train models for 6 epochs. For adapter-based tuning, we select the optimal learning rate from {1e-4, 2e-4, 5e-5} and train adapters for 15 epochs.

5 Results and Analysis

5.1 Subtask A: Supervised Learning

In Table 1, we compare the performance on development sets using fine-tuning and adapter-based tuning along with various techniques. Fine-tuning achieves the best performance in most languages (6 out of 9), which is unsurprising as it optimizes the entire parameter space. Notably, adapter-based tuning demonstrates comparable performance to fine-tuning in Hausa (hau) and Telugu (tel), while even surpassing it in Kinyarwanda (kin), Marathi (mar) and Spanish (spa). Looking at the effectiveness of TAPT and warmup, we observe that they provide benefits in most cases compared to using no techniques at all. Nonetheless, the improvements are sometimes marginal, particularly in languages such as Amharic (amh), English (eng), and Moroccan Arabic (ary), where the baseline performances are already relatively strong compared to other languages.

In our final submission, we selected the best model for each language based on the performance of development sets. As shown in Table 2, our approach largely improves the baseline results (Ousidhoum et al., 2024a), especially for Algerian Arabic (arq), Kinyarwanda (kin), and Moroccan Arabic (ary). In comparison to several top-performing submitted systems, we achieve the best performance in Spanish (spa). There were a total of 40 final submissions in subtask A, and our system ranks first on average in the official leaderboard.[5]

5.2 Subtask C: Cross-lingual Transfer

In subtask C, we replace source language adapters from subtask A with target language adapters. We analyze two groups of language adapters: base language adapters trained only on Leipzig corpora and TAPT language adapters further trained on unlabeled task data. The cross-lingual transfer results

[5]PALI and king001 also achieved competitive performance; however, they are not ranked in the official leaderboard due to missing system descriptions.

Model	afr	arq	amh	eng	hau	hin	ind	kin	arb	ary	pan	spa	Avg.↑
Overlap♢	71.	40.	63.	67.	31.	53.	55.	33.	32.	63.	-27.	67.	45.67
LaBSE♢	79.	46.	84.	80.	62.	76.	47.	57.	61.	40.	-5.	62.	57.42
king001	81.00	**61.44**	**87.83**	-	**73.35**	**84.39**	37.58	62.99	65.68	**81.96**	-	**70.76**	-
UAlberta	80.57	44.13	81.60	-	67.85	82.78	44.90	63.58	**67.15**	60.22	-1.74	57.16	-
ustcctsu	74.87	41.44	70.90	78.40	47.63	65.80	46.02	45.41	46.87	61.32	-24.79	68.51	51.87
umbclu	**82.23**	12.63	4.30	78.75	45.69	15.52	51.53	48.36	3.54	-3.75	-7.75	60.89	32.66
AAdaM (ours)	81.39	55.07	86.29	**79.37**	72.88	83.86	**52.80**	**64.99**	65.32	60.03	**15.53**	62.05	**64.97**

Table 3: Subtask C performance on test sets (Spearman's correlation ×100). ♢: baseline results from Ousidhoum et al. (2024a). We **bold** the best performance across submitted systems.

on development sets are shown in Figure 2. We observe a discrepancy in the optimal source languages selected with two types of adapters, indicating a behavior shift after applying TAPT. Furthermore, the performance for target languages shows high sensitivity to the choice of source language. For example, using Spanish (spa) as the source language for Indonesian (ind) performs significantly better than using Kinyarwanda (kin), showcasing the importance of careful source language selection. When examining each target language, we find that in the case of Amharic (amh), the cross-lingual transfer performance is comparable to its supervised learning performance. However, it remains a challenge for a few languages, such as Indonesian (ind) and Punjabi (pan).

The results for test sets are shown in Table 3. Compared to LaBSE (Feng et al., 2022), a multilingual sentence embedding model, our cross-lingual transfer approach achieves better performance on most languages, especially for Algerian Arabic (arq), Hausa (hau), Moroccan Arabic (ary), and Punjabi (pan). However, our system is surpassed by the simple word overlap baseline in Indonesian (ind), Moroccan Arabic (ary) and Spanish (spa). This highlights the need for nuanced investigation of data distributions across various languages. Subtask C received 18 submissions in total, and we perform the best in the official leaderboard. In particular, we achieve the best performance in Indonesian (ind) and Punjabi (pan), which seem harder for other teams. For Punjabi (pan), where most teams get negative correlation scores, our method maintains its effectiveness.

5.3 Analysis

We partition ground-truth relatedness scores, ranging from 0 to 1, to different levels for fine-grained analysis. Figure 3 shows the detailed model performance for several under-performing languages.

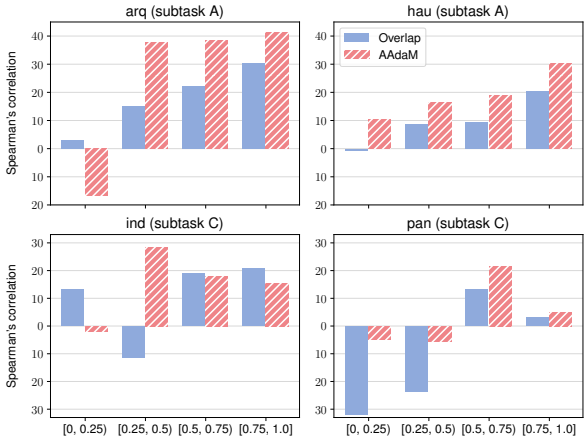


Figure 3: Performance on test sets (Spearman's correlation ×100) in different relatedness levels.

Although our evaluation scores on the entire test sets are all positive, some subsets exhibit negative correlations, particularly those with lower relatedness scores. Moreover, AAdaM largely lags behind the simple word overlap baseline for Algerian Arabic (arq) and Indonesian (ind) within the 0 to 0.25 range. These observations highlight the complexity of capturing nuanced relationships within specific categories, possibly affected by the data annotation procedure and unbalanced learning.

6 Conclusion

In this paper, we introduce our multilingual STR system, AAdaM, developed for the SemEval-2024 Task 1, which achieves competitive results in both subtask A and subtask C. We see noticeable improvements by using data augmentation and task-adaptive pre-training, and demonstrate that adapter-based tuning is an effective approach for supervised learning and cross-lingual transfer. Despite these strengths, our fine-grained analysis reveals that capturing nuanced semantic relationships remains a challenge, highlighting the need for further granular investigation and modeling improvements.

Limitations

Although our approach has demonstrated impressive performance, relying on development sets for source language selection undermines its practical value in the true zero-shot setting. While linguistic (dis)similarity (Littell et al., 2017) is a commonly used estimator for cross-lingual transfer performance, it alone does not explain many transfer results (Lauscher et al., 2020). Philippy et al. (2023) survey different factors that impact cross-lingual transfer performance, finding contradictory conclusions from previous studies. In future work, we plan to scrutinize the interplay among various factors, and select the optimal source language without relying on post-hoc evaluation.

Acknowledgements

We thank Vagrant Gautam and Badr M. Abdullah for their proofreading and anonymous reviewers for their feedback. Miaoran Zhang received funding from the DFG (German Research Foundation) under project 232722074, SFB 1102. Jesujoba O. Alabi was supported by the BMBF's (German Federal Ministry of Education and Research) SLIK project under the grant 01IS22015C.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Maaz Amjad, Grigori Sidorov, and Alisa Zhila. 2020. Data augmentation using machine translation for fake news detection in the Urdu language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2537–2542, Marseille, France. European Language Resources Association.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander Budanitsky. 1999. Lexical semantic relatedness and its application in natural language processing. Technical report, technical report CSRG-390, Department of Computer Science, University of Toronto.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Lukas Lange, Jannik Strötgen, Heike Adel, and Dietrich Klakow. 2021. To share or not to share: Predicting sets of sources for model transfer learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8744–8753, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M Mohammad and Graeme Hirst. 2012. Distributional measures as proxies for semantic relatedness.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yeon Seonwoo, Guoyin Wang, Changmin Seo, Sajal Choudhary, Jiwei Li, Xiang Li, Puyang Xu, Sunghyun Park, and Alice Oh. 2023. Ranking-enhanced unsupervised sentence representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15783–15798, Toronto, Canada. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. NLNDE at SemEval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 488–497, Toronto, Canada. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. MCSE: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969, Seattle, United States. Association for Computational Linguistics.

Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023. Weaker than you think: A critical look at weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14229–14253, Toronto, Canada. Association for Computational Linguistics.

A Model and Architecture Selection

In our preliminary study, we examine the capacity of different pre-trained models with or without any training. To assess their out-of-the-box effectiveness, we extract contextual embeddings for pairs of sentences from various multilingual models, and use the cosine similarity to predict the semantic relatedness score. The multilingual models include:

- **sentence transformers**: mpnet-base-v2[6] and LaBSE (Feng et al., 2022)
- **general-purpose models**: XLMR-large (Conneau et al., 2020), AfroXLMR-large (Alabi et al., 2022), AfriBERTa-large (Ogueji et al., 2021), AfroXLMR-large-61L and AfroXLMR-large-75L (Adelani et al., 2024)

Additionally, we add two simple baselines for comparison: word overlap[7] and fastText (Mikolov et al., 2018). For both fastText vectors and contextual embeddings, we employ mean pooling to get sentence embeddings.

In Table 4, we can see that sentence transformers achieve superior performance in most languages when no training is conducted. This observation is not unsurprising, as they have been trained for sentence embeddings that can better capture the semantic relationships. However, this trend shifts upon fine-tuning the models on task data with either bi-encoder or cross-encoder architecture. Notably, with the cross-encoder architecture, AfroXLMR-large-61L achieves comparable performance to LaBSE. To satisfy the requirement in subtask C, for which the pre-trained model should not be trained on any relatedness or similarity datasets, we adopt AfroXLMR-large-61L as our backbone model with the cross-encoder architecture for all our experiments.

[6] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[7] https://github.com/semantic-textual-relatedness/Semantic_Relatedness_SemEval2024/blob/main/STR_Baseline.ipynb

B Pre-training Data Collection

To pre-train language adapters, we collect open-source corpora from the Leipzig Corpus Collection and use the recent data derived from news and wikipedia domains. Data statistics are shown Table 5. As the SemRel data spans over diverse domains, there is a potential risk of domain mismatch between the pre-training data and task data, which needs a further investigation.

C Source Language Selection

To determine the best source language for cross-lingual transfer, we explore three metrics to estimate the transfer performance:

Linguistic distance. We use the average of six distances obtained from the URIEL Database (Littell et al., 2017) to measure the similarity between a pair of languages. These distances include syntactic, phonological, inventory, geographic, genetic, and featural distances. A lower distance indicates that the two languages are more similar, potentially facilitating more effective transfer.

Token overlap. We follow (Wu and Dredze, 2019) to measure how many tokens are shared in the source training set and the target test set. A higher token overlap indicates that more tokens were encountered during training in the source language, potentially transferring more supervision from the source to the target.

Development set performance. As small development sets are available in the shared task, we use their performance as an indicator of the transfer performance on test sets, assuming that they share a similar data distribution.[8]

In Figure 4, we show the metric values across different source languages, along with the best source languages identified by distinct metrics. After post-hoc evaluation following the release of test sets, we find that the performance of the development set indeed serves as the most reliable indicator, as the

[8] When training is allowed, it might be more advantageous to use small development sets for training directly rather than source selection, which needs to be further explored.

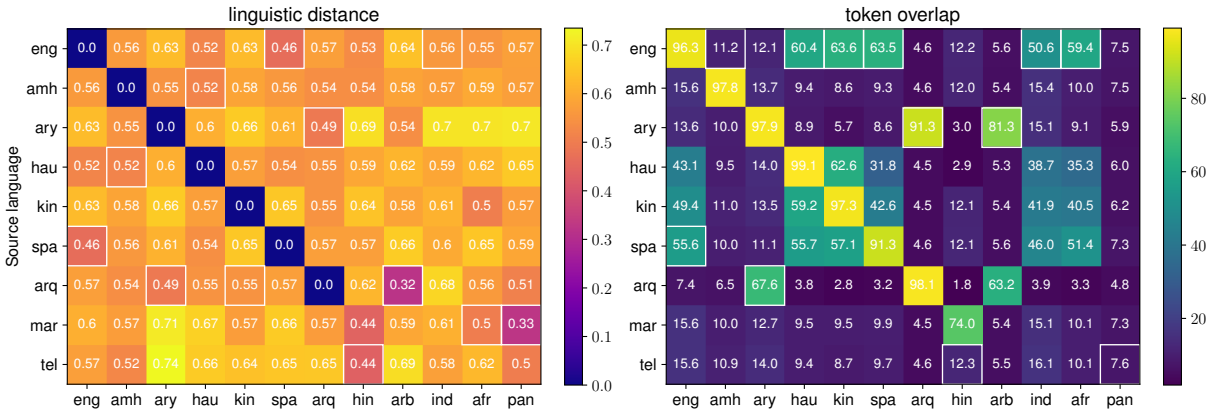
Model	eng	amh	arq	ary	spa	hau	mar	tel	Avg.↑
Baselines w/o training:									
Overlap	56.57	63.28	44.00	**53.76**	58.67	38.86	57.52	60.61	54.16
FastText	55.69	60.64	**44.27**	22.12	57.47	9.19	59.23	69.39	47.25
mpnet-base-v2	**81.94**	69.94	26.35	34.40	56.58	30.86	72.43	56.33	53.60
LaBSE	72.14	**76.49**	40.80	38.58	**63.11**	**41.51**	**73.83**	**75.99**	**60.31**
XLMR-large	39.53	42.07	27.91	4.15	47.59	7.34	40.51	56.36	33.18
AfroXLMR-large	16.55	39.82	20.30	-0.46	30.42	8.13	35.94	30.74	22.68
AfriBERTa-large	53.12	69.23	16.04	13.36	56.68	35.14	20.84	9.73	34.27
AfroXLMR-large-61L	44.10	52.96	32.15	0.35	51.07	17.62	37.66	47.17	35.39
AfroXLMR-large-75L	22.61	37.93	29.38	-2.39	43.58	13.86	32.13	40.42	27.19
Bi-encoders w/ supervised training:									
mpnet-base-v2	85.07	80.43	56.73	75.51	65.29	58.62	81.53	74.49	72.21
LaBSE	84.45	82.59	59.49	78.29	69.02	68.94	**83.97**	76.35	75.39
AfroXLMR-large-61L	82.81	74.61	40.02	66.58	66.65	66.51	38.51	65.73	62.68
Cross-encoders w/ supervised training:									
mpnet-base-v2	80.26	75.04	60.25	80.31	64.92	53.66	65.36	68.54	68.54
LaBSE	86.13	84.75	**60.75**	**82.55**	67.23	69.31	81.10	77.25	**76.13**
AfroXLMR-large-61L	**86.65**	**84.88**	46.61	81.56	**69.08**	**74.65**	75.55	**80.94**	74.99

Table 4: Performance of 10-fold cross-validation on training sets (Spearman's correlation $\times 100$). For each language, we **bold** the best performance achieved in *w/o training* and *w/ supervised training* settings.

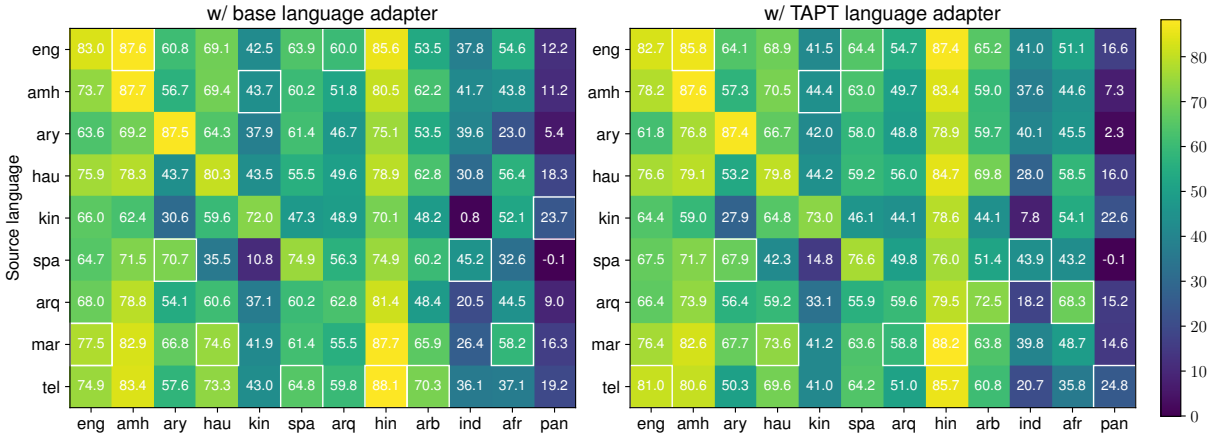
Language	Family / Subfamily	Domain	Corpus Size
English (eng)	Indo-Europoean / Germanic	News, Wikipedia	1.2M
Afrikaans (afr)	Indo-Europoean / Germanic	News, Wikipedia	68k
Amharic (amh)	Afro-Asiatic / Semitic	Community, Wikipedia	250k
Modern Standard Arabic (arb)	Afro-Asiatic / Semitic	News, Wikipedia	110k
Algerian Arabic (arq)	Afro-Asiatic / Semitic	News	244k
Moroccan Arabic (ary)	Afro-Asiatic / Semitic	News	564k
Spanish (spa)	Indo-Europoean / Italic	News, Wikipedia	444k
Hausa (hau)	Afro-Asiatic / Chadic	Community, Wikipedia	564k
Hindi (hin)	Indo-European / Indo-Iranian	News, Wikipedia	472k
Indonesian (ind)	Austronesian / Malayic	News, Wikipedia	92k
Kinyarwanda (kin)	Niger-Congo / Atlantic–Congo	Community	320k
Punjabi (pan)	Indo-European / Indo-Iranian	Wikipedia	412k
Marathi (mar)	Indo-European / Indo-Iranian	News, Wikipedia	856k
Telugu (tel)	Dravidian / South-Central	News, Wikipedia	756k

Table 5: Data statistics for pre-training corpora collected from the Leipzig Corpus Collection.

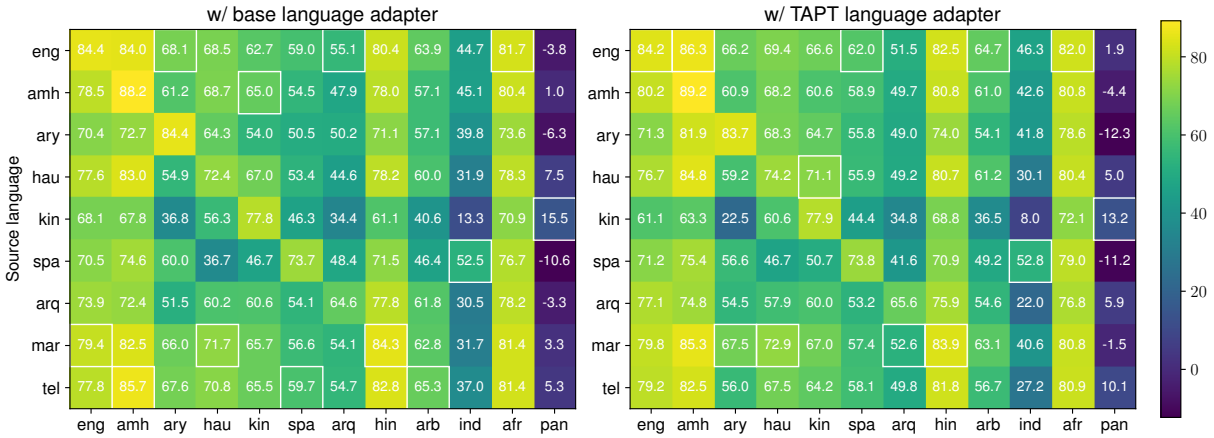
optimal source languages it selected closely align with the ground truth selections.



(a) Left: Linguistic distances between source and target languages. The smallest distance for each target language is highlighted with a box. Right: Token overlaps between source and target languages. The highest overlap for each target language is highlighted with a box. The corresponding source languages are predicted as the best sources for cross-lingual transfer.



(b) Performance on development sets (Spearman's correlation ×100) using different types of language adapters. Boxes are used to highlight the optimal performances for each target language, and the corresponding source languages are predicted as the best sources for cross-lingual transfer.



(c) Performance on test sets (Spearman's correlation ×100) using different types of language adapters. Boxes are used to highlight the optimal performances for each target language, and the corresponding source languages are the ground-truth best sources for cross-lingual transfer.

Figure 4: Comparison of different source language selection methods.