USTCCTSU at SemEval-2024 Task 1: Reducing Anisotropy for Cross-lingual Semantic Textual Relatedness

Jianjian Li[1] * **Shengwei Liang**[1] * **Yong Liao**[1,2] †
Hongping Deng[2] **Haiyang Yu**[2] †
1. University of Science and Technology of China, CCCD Key Lab of MCT
2. Institute of Dataspace
{sa22221088,sewell}@mail.ustc.edu.cn

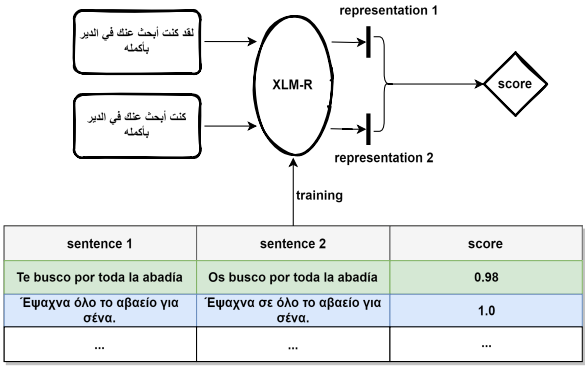
Abstract

Cross-lingual semantic textual relatedness task is an important research task that addresses challenges in cross-lingual communication and text understanding. It helps establish semantic connections between different languages, crucial for downstream tasks like machine translation, multilingual information retrieval, and cross-lingual text understanding. Based on extensive comparative experiments, we choose the $XLM\text{-}R_{base}$ as our base model and use pre-trained sentence representations based on whitening to reduce anisotropy. Additionally, for the given training data, we design a delicate data filtering method to alleviate the curse of multilingualism. With our approach, we achieve a **2nd** score in Spanish, a **3rd** in Indonesian, and multiple entries in the top ten results in the competition's track C. We further do a comprehensive analysis to inspire future research aimed at improving performance on cross-lingual tasks.

1 Introduction

Semantic textual relatedness (STR) encompasses a broader concept that takes into account various commonalities between two sentences. This includes factors such as being on the same topic, expressing the same viewpoint, originating from the same period, one sentence elaborating on or following from the other, and more. SemEval is an international workshop on semantic evaluation. In track C of SemEval-2024 task 1: Cross-lingual (Ousidhoum et al., 2024b), participants are to submit systems,which are developed without the use of any labeled semantic similarity or semantic relatedness datasets in the target language and with the use of labeled datasets (Ousidhoum et al., 2024a) from at least one other language.

Various methods were proposed to address the task of textual relatedness. One common approach



sentence 1	sentence 2	score
Te busco por toda la abadía	Os busco por toda la abadía	0.98
Έψαχνα όλο το αβαείο για σένα.	Έψαχνα σε όλο το αβαείο για σένα.	1.0
...	...	...

Figure 1: The description of cross-lingual semantic textual relatedness task.

is based on feature engineering, where the syntactic, semantic, and structural features of text, such as word frequency, TF-IDF, and word embeddings, are extracted. Machine learning algorithms are then employed for relatedness determination. Another popular approach is based on deep learning methods, such as Convolutional Neural Networks (LeCun et al., 1989), Recurrent Neural Networks (Graves and Graves, 2012) , and self-attention mechanisms (Vaswani et al., 2017). These methods can capture semantic relationships and contextual information within the text, and they are trained on large-scale datasets to enhance model performance and generalization ability.

However, there are two challenges in track C of SemEval-2024 task 1:

- Compared with static word representation such as Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014), the pre-trained language models (PLM) can obtain sentence representation for different sentence in different contexts, thereby solving different problems. However, the vectors of BERT-based PLM models have limitations: **BERT-based models always induces a non-smooth anisotropic semantic space of sentences,**

* Equal contribution and shared co-first authorship.
† Corresponding author.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 881–887

which harms its performance of semantic similarity (Gao et al., 2019; Li et al., 2020), which can lead to a challenge that sentences are strikingly similar while using the cosine similarity metric.

- Participants are not allowed to utilize labeled datasets in the target language for training. Instead, they must use labeled data in different languages as the training set to train the model and provide predictions in the target language. However, **multilingual pre-trained models suffer from the curse of multilingualism** (Conneau et al., 2020), that is, the overall performance of both monolingual and cross-lingual baselines declines when adding more languages to training data over a certain point. Hence, it is essential to investigate which additional languages would be inefficient as the training dataset for the target language.

In this paper, we used whitening techniques (Su et al., 2021), which maps vectors to standard orthogonal bases, to transform the word vector representations from anisotropic to isotropic, and surprisingly, we found that whitening significantly improves the accuracy of judging semantic similarity. Given the absence of labeled data in the target language, it is difficult to determine which other language would yield better prediction results when used as training data. Therefore, we proposed that removing certain language categories from the training data for a specific target language contributed to improving performance.

We conducted extensive experiments to demonstrate the effectiveness of the method we employed. As a result, our submitted outcomes achieved a **2nd** score in Spanish and a **3rd** score in Indonesian in track C of SemEval-2024's task 1. Additionally, we obtained multiple top-ten rankings in the competition.

2 Background

The task of semantic text relatedness covers several specific subtasks, including semantic similarity, semantic matching, textual entailment, semantic relation classification, and text pair ranking. Previous work has proposed various methods for these specific tasks, such as: Lexical and syntactic-based methods (Gamallo et al., 2001; Pakray et al., 2011): These methods rely on lexical and syntactic rules, such as word vector matching, lexical overlap, and syntactic tree matching. However, these methods often fail to capture higher-level semantic relationships. Feature engineering-based machine learning methods (Chia et al., 2021; Fan et al., 2019): These methods involve using manually designed features, such as bag-of-words models (Zhang et al., 2010), tf-idf weights, and syntactic features, followed by using machine learning algorithms like support vector machines and random forests for prediction.

While these methods have improved performance to some extent, they still have limitations in capturing complex semantic relationships. Neural network-based models: These models use neural networks to learn representations of text and capture semantic relationships between texts through training data. This includes methods that fine-tune pre-trained language models (e.g., BERT (Kenton and Toutanova, 2019) and GPT2 (Radford et al., 2019) etc.), as well as approaches that employ Siamese networks, LSTM, CNN, and other architectures for text encoding and matching. Transfer learning and multi-task learning (Pilault et al., 2020; Wu et al., 2020): These methods leverage knowledge from pre-trained models on related tasks to improve the performance of semantic textual relatedness tasks through transfer learning (Koroleva et al., 2019). Multi-task learning combines multiple related tasks in training to enhance the model's generalization ability and effectiveness. Application of external knowledge resources: Researchers have also attempted to incorporate external knowledge resources such as word embeddings, semantic knowledge graphs, and multilingual data to enhance the model's understanding of semantic relationships.

For cross-lingual semantic similarity tasks, mapping texts from different languages into a shared semantic space for similarity calculation is necessary. To address this, researchers have proposed various cross-lingual representation learning methods. Among them, unsupervised alignment methods like unsupervised machine translation (Lample et al., 2017) and cross-lingual pre-training models (Liang et al., 2020) can learn the correspondences between multiple languages and map texts to a shared vector space.

However, (Conneau and Lample, 2019) and (Wang et al.) mentioned that vector representations based on the Transformer models exhibit anisotropy, which means that the vectors are unevenly distributed and clustered in a narrow cone-shaped space. Therefore, both Bert-flow (Li et al.,

2020) and Bert-whitening (Su et al., 2021) aim to address the same issue, which is the anisotropy and uneven distribution of sentence embeddings.

3 System Overview

3.1 Framework Overview

In this section, we will introduce our proposed method for STR task which has three main modules.

- **PLM Encoder** We adopted the pretrained language model XLM-RoBERTa-base ($XLM\text{-}R_{base}$) (Conneau et al., 2020) for initial sentence encoding, which combines two powerful models: Transformer and RoBERTa. $XLM\text{-}R_{base}$ demonstrates strong multilingual capabilities and a deep understanding of semantics, surpassing some monolingual pre-training models. After conducting a series of tests on mBERT (Pires et al., 2019), XLM (Conneau and Lample, 2019), and $XLM\text{-}R_{base/large}$, we selected $XLM\text{-}R_{base}$ as the encoder due to its superior performance.

- **Whitening Module** After obtaining the sentence vectors of two utterances using $XLM\text{-}R_{base}$, we could have directly calculated the cosine similarity between the two vectors, but the sentence vectors after $XLM\text{-}R_{base}$ show anisotropy between them and are distributed in a conical space, resulting in a high cosine similarity. Therefore, we introduce the Whitening module to change the distribution of the sentence vector space so that its distribution has various anisotropies, amplifying the differences between the vectors and stimulating the performance of $XLM\text{-}R_{base}$ on the semantic text similarity reading task.

- **Data Filtering** The authors of (Conneau et al., 2020) mention the curse of multilingualism, where adding more languages leads to an improvement in cross-lingual performance for low-resource languages up to a certain point, after which the overall performance of both monolingual and cross-lingual baselines declines. In the task of cross-lingual semantic text similarity, to maximize the exploration of the positive impact of other languages on the target language, we propose a new dataset selection method. As the influence between languages is mutual, we utilize the unlabeled data of the target language to detect the impact of each language in track A, excluding the target language, and infer its influence on the target language. This allows us to select the training dataset optimally. This approach helps eliminate interference from certain languages on the target language and avoids the curse of multilingualism.

3.2 PLM Encoder

Through a simple test and comparative analysis of different multilingual pre-training models, we found that $XLM\text{-}R_{base}$ outperforms mBERT. $XLM\text{-}R_{base}$ is a cross-lingual pre-training model based on the BERT architecture, an improvement and extension of the original XLM model. The goal of $XLM\text{-}R_{base}$ is to enhance the performance and effectiveness of multilingual text processing. $XLM\text{-}R_{base}$ utilizes larger-scale pre-training data and more sophisticated training methods to enhance the model's representation capabilities. It undergoes deep learning on a large amount of unsupervised data using RoBERT (Liu et al., 2019) technology. This enables $XLM\text{-}R_{base}$ to better understand and capture the semantic and grammatical features between different languages. Compared to the original XLM, $XLM\text{-}R_{base}$ has made several improvements. Firstly, it introduces a dynamic masking mechanism that allows the model to better perceive contextual information. Secondly, $XLM\text{-}R_{base}$ emphasizes cross-lingual consistency learning through adversarial training, enabling better alignment and sharing of model parameters. This enables $XLM\text{-}R_{base}$ to provide more accurate representations of texts in cross-lingual tasks. Compared to mBERT, $XLM\text{-}R_{base}$ employs larger-scale pre-training data, covers more languages, and incorporates improvements through RoBERTa technology. This enables $XLM\text{-}R_{base}$ to better learn and capture the semantic and grammatical features between different languages, thereby enhancing the model's representation capabilities and performance.

3.3 Whitening Module

Due to the existence of anisotropy among the vectors obtained from the initial encoding by $XLM\text{-}R_{base}$, cosine similarity cannot accurately measure the semantic similarity between sentences. Therefore, we chose to use whitening to map the original vector space to an isotropic space, where the vectors are transformed into vectors in a standard

orthogonal bases. The principle is as follows:

Suppose we have a set of sentence vectors $S = \{s_1, s_2, \ldots, s_n\}$, the set of vectors can be transformed into a set of vectors with isotropy (i.e., zero mean and a covariance matrix of the identity matrix) through the following transformation $\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_n\}$.

$$\tilde{s}_i = (x_i - \mu)\mathbf{W} \tag{1}$$

If we want to make the set $\tilde{S}$ have a zero mean, we need to:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{2}$$

The next step is to calculate $\mathbf{W}$. The covariance matrix of S:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{s_i} - \mu)^\top (s_i - \mu) \tag{3}$$

The covariance matrix of $\tilde{S}$:

$$\tilde{\Sigma} = \mathbf{W}^\top \Sigma \mathbf{W} \tag{4}$$

If we want to transform $\tilde{\Sigma}$ into the identity matrix I, we need to:

$$\tilde{\Sigma} = \mathbf{W}^\top \Sigma \mathbf{W} = \mathbf{I} \tag{5}$$

Then:

$$\Sigma = (\mathbf{W}^\top)^{-1}\mathbf{W}^{-1} = (\mathbf{W^{-1}})^\top \mathbf{W}^{-1} \tag{6}$$

Since Σ is a positive definite symmetric matrix as the covariance matrix, it can be decomposed using Singular Value Decomposition (SVD), yielding:

$$\Sigma = \mathbf{U}\Lambda\mathbf{U}^\top \tag{7}$$

By combining equations (6) and (7), we obtain:

$$(\mathbf{W^{-1}})^\top \mathbf{W}^{-1} = \mathbf{U}\Lambda\mathbf{U}^\top = \mathbf{U}\sqrt{\Lambda}\sqrt{\Lambda}\mathbf{U}^\top \tag{8}$$

Then:

$$(\mathbf{W^{-1}})^\top \mathbf{W}^{-1} = (\sqrt{\Lambda}\mathbf{U}^\top)^\top \sqrt{\Lambda}\mathbf{U}^\top \tag{9}$$

Therefore, we can obtain $\mathbf{W^{-1}} = \sqrt{\Lambda}\mathbf{U}$, and finally obtain $\mathbf{W}$ as follows:

$$\mathbf{W} = \mathbf{U}\sqrt{\Lambda^{-1}} \tag{10}$$

3.4 Data Filtering

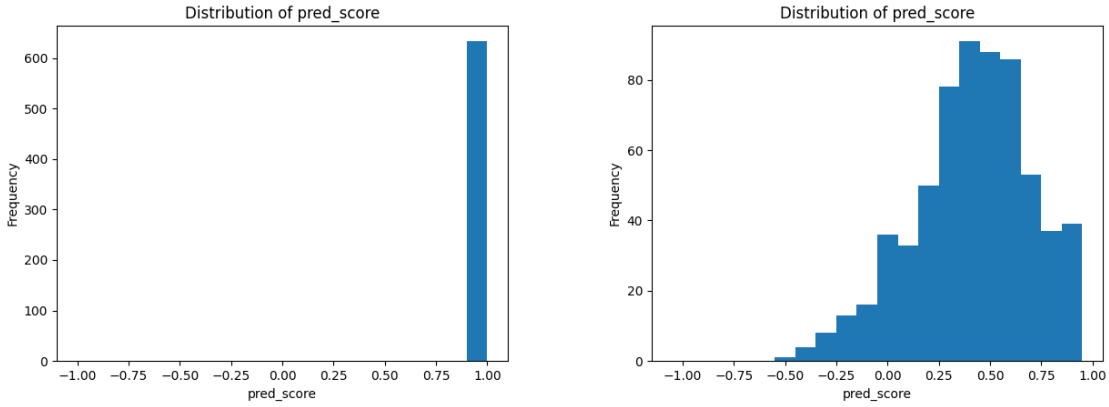
Our experiments have shown that when selecting training data for the target language, using a mixture of multiple languages often yields better results than using a single language. The authors of the $XLM\text{-}R_{base}$ paper mentioned that incorporating more languages improves the cross-lingual performance of low-resource languages up to a certain point. Beyond that point, the overall performance of both monolingual and cross-lingual benchmarks starts to decline. Additionally, we believe that there is interdependence between languages. For example, if including text from language A in training set to compute whitening parameters leads to a decrease in the prediction performance for language B, we expect that the opposite would hold true as well.

Therefore, inspired by this insight, we used the text in the target language as the dataset and individually tested the labeled training data provided in track A for different languages. For example, if the target language is identified by T, we use the text of T for whitening, and test the performance on language $Test_A$, $Test_B$, $Test_C$, $Test_D$,... one by one. If the prediction performance of $Test_A$ decreases after using T compared to not using T (measured by the Spearman correlation (Myers and Sirois, 2004) between the gold labels and predicted labels obtained using language $Test_A$), then $Test_A$ is excluded from target language's training set.

In the case of the Spanish, using the training set without data filtering (1,000 each of all data except Spanish) resulted in a final spearman coefficient of 0.6375; using the training set with data filtering (1000 each of kin and ind) resulted in a final spearman coefficient of 0.6886. Although the training data for about ten languages were reduced, the results were are significantly improved.

4 Experimental Setup

We use the 12 labeled training data from (Ousidhoum et al., 2024a) as training data and the test data from track C as test data. We observe that the amount of data for each language is concentrated around 1,000, so we take 1,000 as the boundary, use oversampling to make up for less than 1,000, and use randomization to take out 1,000 for more than 1,000 to ensure that sentence pairs of different similarities are involved. In finding the training set combinations for the target languages, we compute



(a) without whitening (b) with whitening

Figure 2: The results of model without whitening and with whitening.

the μ and $\mathbf{W}$ of whitening using the text data of the target languages in track C. We predict the training data one by one for each language, and compute the spearman coefficients using the gold labels and the predicted labels of the training data, and compare the results with the data without any whitening (i.e., the prediction result of the base model) to evaluate whether the target language enhances a certain language in the training data or not, and if it does not, it is excluded from the train data. Eventually, the remaining language data is used as a training set to predict the target language.

The hyperparameters are set as follows: we choose to freeze the pretrained model $XLM\text{-}R_{base}$ while setting the topk parameter of whitening to 256. The rubric we used was the spearman coefficient, calculated using the methodology provided by the competition officials.

5 Results

The official competition used the spearman coefficients to evaluate the results, and Table 1 gives the results of the spearman coefficients for both Indonesian (ind) and Spanish (esp) languages throughout the experiment. There is a big difference in the multilingual ability of different model bases. We chose $XLM\text{-}R_{base}$, which performs better, and we can see that the overall results are improved after using the whitening module to transform the vector space; $XLM\text{-}R_{base}$ with whitening is better than baseline, and we got a good ranking in track C of SemEval-2024 task 1, in which we ranked second in esp and third in ind.

As can be seen from Table 1, the whitening module improves the STR task more significantly,the

	ind-test	**esp-test**
Baseline	0.4700	0.6200
mBERT	0.4390	0.5971
$XLM\text{-}R_{base}$	0.4390	0.5907
$XLM\text{-}R_{large}$	0.4267	0.6003
mBERT-whitening	0.4471	0.6411
$XLM\text{-}R_{base}$-whitening	0.4746	**0.6886**
$XLM\text{-}R_{large}$-whitening	**0.4845**	0.6648

Table 1: The spearman coefficient of different models and baseline.

baseline is given by (Ousidhoum et al., 2024a). In order to further verify whether whitening works, we counted the cosine similarity distribution statistics of the data without whitening processing and after whitening. Figure 2 gives two cosine similarity statistics. The left side is the cosine similarity statistics without whitening. The cosine similarity of all utterance pairs is concentrated between 0.9 and 1.0, indicating that the vector space is anisotropic. In contrast, after adding whitening, the whole distribution tends to be normal, which indicates that whitening plays a role in mapping the vectors to an isotropic space, amplifying the differences between statements.

6 Conclusion

We use $XLM\text{-}R_{base}$ with whitening and propose a dataset filtering method that exploits the positive correlation of linguistic interactions, achieving good rankings in SemEval-2024 task 1 track C. We verifies that whitening performs well on utterance characterization as well as STR task. Besides, the proposed dataset filtering method is more efficient

and can alleviate the multilingual curse problem in cross-language problems to some extent.

In the future, we will further study this positive correlation of language interactions, and we hope that this correlation can become more detailed, not only in terms of inter-language correlations but also in terms of the domain of the text. We also hope that this correlation can be better utilized in dataset preprocessing, not only to eliminate poorly performing languages but to further improve the combination of datasets that can be directly selected to correspond to the optimal solution.

Acknowledgments

We want to express gratitude to the anonymous reviewers for their hard work and kind comments, which will further improve our work in the future. This work is funded by national key research and development program under grant 2021YFC3300500-02.

References

Zheng Lin Chia, Michal Ptaszynski, Fumito Masui, Gniewosz Leliwa, and Michal Wroczynski. 2021. Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing & Management*, 58(4):102600.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Cheng Fan, Yongjun Sun, Yang Zhao, Mengjie Song, and Jiayuan Wang. 2019. Deep learning-based feature engineering methods for improved building energy prediction. *Applied energy*, 240:35–45.

Pablo Gamallo, Caroline Gasperin, Alexandre Agustini, and Gabriel P Lopes. 2001. Syntactic-based methods for measuring word similarity. In *International Conference on Text, Speech and Dialogue*, pages 116–125. Springer.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv e-prints*, pages arXiv–1907.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Anna Koroleva, Sanjay Kamath, and Patrick Paroubek. 2019. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics*, 100:100058.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

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. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12.

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. *Preprint*, arXiv:2402.08638.

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.

Partha Pakray, Sivaji Bandyopadhyay, and Alexander Gelbukh. 2011. Textual entailment using lexical and syntactic similarity. *International Journal of Artificial Intelligence and Applications*, 2(1):43–58.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jonathan Pilault, Amine Elhattami, and Christopher Pal. 2020. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. *arXiv preprint arXiv:2009.09139*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. Improving neural language generation with spectrum control.

Sen Wu, Hongyang R Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52.