silp_nlp at SemEval-2024 Task 1: Cross-lingual Knowledge Transfer for Mono-lingual Learning

Sumit Singh, Pankaj Kumar Goyal and **Uma Shanker Tiwary**
Indian Institute of Information Technology, Allahabad
{sumitrsch, pankajgoyal02003}@gmail.com
ust@iiita.ac.in

Abstract

Our team, silp_nlp, participated in all three tracks of SemEval2024 Task 1: Semantic Textual Relatedness (STR). We created systems for a total of 29 subtasks across all tracks: nine subtasks for track A, 10 subtasks for track B, and ten subtasks for track C. To make the most of our knowledge across all subtasks, we used transformer-based pre-trained models, which are known for their strong cross-lingual transferability. For track A, we trained our model in two stages. In the first stage, we focused on multi-lingual learning from all tracks. In the second stage, we fine-tuned the model for individual tracks. For track B, we used a unigram and bigram representation with suport vector regression (SVR) and eXtreme Gradient Boosting (XGBoost) regression. For track C, we again utilized cross-lingual transferability without the use of targeted subtask data. Our work highlights the fact that knowledge gained from all subtasks can be transferred to an individual subtask if the base language model has strong cross-lingual characteristics. Our system ranked first in the Indonesian subtask of Track B (C7) and in the top three for four other subtasks.

1 Introduction

The importance of semantic relatedness in language has been long recognized. Applications include sentence representation, question answering, and text summarization (Abdalla et al., 2023). Sentences can be related through either paraphrasal or entailment relations, or through broader commonalities such as shared topics, viewpoints, temporal origins, and logical connections.

This shared task (Ousidhoum et al., 2024b) aims to expand the scope of significant research in natural language processing (NLP) by incorporating 14 languages. The focus of the research is on semantic similarity and has predominantly been conducted in English. The languages included in the task are Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu.

The task requires predicting the degree of semantic textual relatedness (STR) between pairs of sentences in multiple languages. The task is to rank these sentence pairs based on their level of relatedness, with scores ranging from 0 (completely unrelated) to 1 (maximally related). This task is divided into the following three tracks.

Track A: Supervised Challenge was developing a system trained on labelled datasets provided for the 11 subtasks. Publicly available related datasets could be utilized, but no additional dataset could be used in this work.

Track B: Unsupervised The challenge was developing a system without using labelled datasets to measure semantic relatedness or similarity with the text units longer than two words only. We took advantage of unigram and bigram features with SVM regression and XGBoost.

Track C: Cross-Lingual The challenge was to develop a system without labelled datasets in the target language and with the use of labelled dataset(s) in at least one other language (subtask). All datasets of track A other than the targeted subtask are utilized for similarity prediction in this work.

Our system utilizes cross-lingual learning by implementing multi-stage training methods similar to those used in (Wang et al., 2022), (He et al., 2022), and (Singh and Tiwary, 2023) for track A. In the first stage, we selected pre-trained cross-lingual language models that cover the languages used in our task and fine-tuned them on a combined dataset of all subtasks in track A for five epochs. This created a model checkpoint that had knowledge of multiple languages relevant to our task. In the second stage, we fine-tuned the model checkpoint generated in the first stage for each track individually.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1197–1203

For track C, we created a dataset by combining all the data from track A, except for the targeted subtask, and fine-tuned it with the language model in a supervised manner. We used unigram and bigram bag-of-words representation with SVM-based regression (SVR) and XGBoost for each subtask in track B.

Our team achieved the best model for the Telugu and Marathi subtasks of track A by using the MuRIL large model (Khanuja et al., 2021). In the first stage, we fine-tuned the model for all three tracks (English, Telugu, and Marathi) and then fine-tuned the model checkpoints for Telugu and Marathi.

For all track A languages, we fine-tuned XLM-R for subtasks in the 1st stage, and we fine-tuned the checkpoint generated in stage one for all monolingual tracks in the 2nd stage.

For Track B, only monogram or bigram representation is allowed for supervised training. We obtained comparable results using both unigram and bigram representations in combination with SVR and XGBoost.

In Track C, we used all training data from Track A except for the current subtask data since the use of the same language data was not allowed in Track C. We adopted a cross-lingual transfer approach, where MuRIL gave the best result for the Hindi subtask, while XLM-R predicted the best results for the other subtasks.

The results for each subtask are presented in Table 2, 3 and 4, along with the baseline results. In addition to being multilingual, the key challenges of the task were the presence of many low-resource languages that lack proper pre-trained models for learning and the limited availability of training examples for some subtasks of Track A (see Table 1). To address these challenges, we utilized language models (LMs) with cross-lingual transferability, along with a two-stage training strategy. Our code can be found here[1].

2 Related Work

The Semantic Textual Similarity (STR) task 2015 (Agirre et al., 2015) had three subtasks. The findings showed that the UMBC PairingWords system achieved the best score by semantically differentiating distributionally similar terms (Han et al., 2015). In the subsequent STR task (Cer et al., 2017), there are seven tasks that concentrate on multilingual and cross-lingual pairs. Additionally, one sub-track will delve into MT quality estimation data. The team ECNU (Tian et al., 2017) achieved the highest score using ensembles of well-performing feature-engineered models with deep learning methods. These models used random forest (RF), gradient boosting (GB), and XGBoost (XGB) regression methods. However, statistical and machine learning models were not the best, as transformer-based models gained attention after (Devlin et al., 2019). These models are pre-trained on large amounts of data and fine-tuned for various downstream tasks. Researchers have created the sentence transformer (Reimers and Gurevych, 2019) architecture for finding similar sentences. It is useful when given multiple sentences corresponding to a sentence, and we need to find the most similar one. However, fine-tuning the sentence transformer with the downstream task requires proper alignment between the dataset on which the sentence transformer is pre-trained and the dataset of the downstream task. In this task, there are multiple subtasks associated with multiple languages. Therefore, motivated by the performance of cross-lingual transformer-based models, we have used transformer-based language models that have strong cross-lingual transferability. It has been seen that cross-lingual transferability has advantages in various NLP tasks (Singh and Tiwary, 2023; Wang et al., 2022).

3 Data

This shared task provided fourteen sets of monolingual data (Ousidhoum et al., 2024a). There were nine languages for track A, each with training, validation, and testing data. For tracks B and C, only validation and testing data were provided for the Afr, Arb, Hin, Ind, and Pan languages. The training, validation, and testing data distribution for all languages is tabulated in Table 1. The English language had over 5,500 training examples, while other languages had comparatively fewer data provided.

4 Methodology

Our system has chosen robust cross-lingual transfer models such as XLM-R (Conneau et al., 2020), which is pre-trained on over 100 languages, and MuRIL (Khanuja et al., 2021), which is pre-trained on all Indic and English languages, for track A and C. We have followed a two-stage training approach

[1] https://github.com/singhsumit1/Semeval-Semantic_textual-relatedness.git

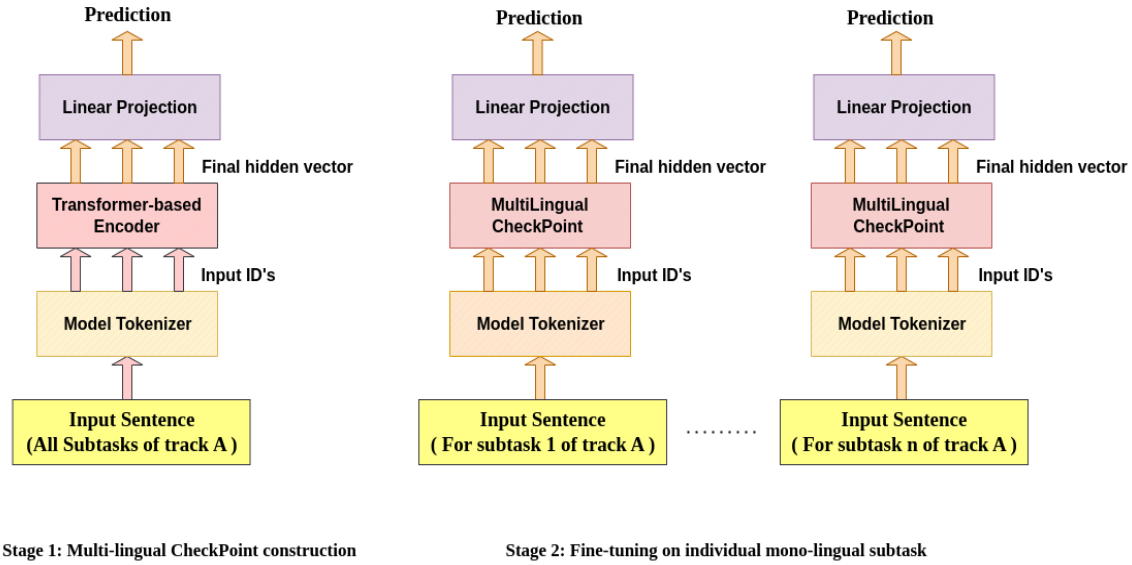

Figure 1: General Architecture of two-stage training.

Data	afr	amh	arb	arq	ary	eng	esp	hau	hin	ind	kin	mar	pan	tel
Train	-	992	-	1,262	925	5,500	1,562	1,763	-	-	778	1,155	-	1,146
Dev	375	95	32	92	70	250	140	212	288	144	102	293	242	130
Test	375	171	595	584	427	2,500	600	603	968	360	222	298	634	297
Total	750	1,258	627	1,938	1,422	8,250	2,299	2,578	1,256	504	1,102	1,746	876	1,573

Table 1: The datasets varied in the number of instances within their training, development, and test sets. Languages such as afr, arb, hin, ind, and pan lacked training data and were exclusively utilized in unsupervised and cross-lingual contexts.

for track A. The detailed methodology for each track is explained in the following subsections.

For tracks A and C, we combined both sentences of an example data with a special token of LM (</sep> for XLM-R and Roberta, and [SEP] for MuRIL) in the preprocessing stage. Model tokenizer tokenizes the combined sentence into tokens and generates token IDs and attention masks. For other subtasks, we utilized XLM-R and MuRIL with the two-stage training approach.

4.1 Methodology for track A

4.1.1 Two-stage training

In the initial stage of our project, we conducted multi-lingual training by utilizing the training data of all subtasks of track A in a selected LM that supports them. As shown in Fig. 1, our model used the annotations of all subtasks of track A in this stage. We performed experiments using various hyperparameters across five epochs, selecting the best multilingual checkpoint based on the average validation data loss. In the second stage, we fine-tuned the multilingual checkpoint from the first stage and utilized it as an initial model for fine-tuning each monolingual subtask in track A. We trained each mono-lingual track with different hyper-parameters in the second stage and selected the final model based on the validation data loss of the corresponding subtask.

4.1.2 Model Architecture

Figure 1 shows that the model tokenizer first tokenizes the input sentence. To improve GPU utilization, the tokenizers are set to a length of 92. Next, the language model generates word embeddings for each token. The embedding of the first token (which is <s> for the XLM-R and [SEP] for MuRIL) is passed through a linear layer, which projects it into logits, a vector of size one that represents the predicted similarity. Finally, we apply the Mean Square Error (MSE) loss function to calculate the difference between the prediction and the ground truth.

4.2 Methodology for track B

For track B, the sentences were converted into unigram and bigram representation and Support Vec-

Language	A1 arq	A2 amh	A3 eng	A4 hau	A5 kin	A6 mar	A7 ary	A8 esp	A9 tel
XLM-R (one-stage)	0.49	0.49	0.78	0.68	0.23	0.86	0.63	0.58	0.78
MuRIL (one-stage)	-	-	0.77	-	-	0.856	-	-	0.838
XLM-R (two-stage)	0.59	0.84	0.84	0.72	0.49	0861	0.81	0.66	0.789
MuRIL (two-stage)	-	-	-	-	-	0.862	-	-	0.842
Baseline_Score	0.6	0.85	0.83	0.69	0.72	0.88	0.77	0.7	0.82
Rank	7	8	6	5	15	13	10	17	6

Table 2: Results of all subtasks of track A has been tabulated for both the settings one-stage and two-sage with both the LMs XLM-R and MuRIL.

tor Regression[2] (SVR) (Fu et al., 2016; Liu et al., 2017).

4.2.1 Unigrams /Bigrams embeddings

Both sentences in the examples are combined and transformed into a vector. To create the vector, each sentence is indexed based on the presence of unigrams/bigrams, and the corresponding index value is filled with the count of unigrams/bigrams. The resulting vector is then fed into the SVR model along with the label values for training.

4.3 Methodology for track C

The methodology followed in track C is similar to the first stage of track A, with the difference that the combined dataset includes all subtasks except for the targeted subtask. For instance, to build a system for the English (eng) subtask, all data from the subtasks of track A, except for the eng subtask, is collected. The model architecture is also similar to that of track A but with only one stage involved.

5 Experimental setup

5.1 Track A and Track C

We achieved our best score using the MuRIL setup for the Telugu and Marathi subtasks, while the XLM-R setup worked best for the other track. During the training process, we experimented with different learning rates (5e-6, 2e-5, 5e-5, 8e-5, and 1e-4) and batch sizes (16, 32, and 64) in both stages. We selected the best model based on validation loss after five epochs of training.

For track A and track C, we used the setups outlined in Fig. 1. We implemented our task using the xxxTokenClassification class defined in (Wolf et al., 2020) for regression, where xxx refers to the selected model. We set the number of labels to one. The other hyperparameters for achieving the best results with both language models are listed in Table 5.

[2]Support Vector Regression

5.2 Track B

For subtasks which are training data available in track, we have generated monogram and bigram embedding and performed supervised learning with support vector regression (SVR) and gradient Boosting regression (XGBoost) with the Scikitlearn[3] library.

Evaluation metrics Results are Pearson correlation coefficient, which shows the similarity between two sentences.

6 Results and Analysis

The table below shows the Pearson score with official rank for Track A, Track B and Track C, along with the baseline score. Please refer to Table 2, 3 and 4 for more details. Our system performed exceptionally well in the Indonesian (ind) subtask of track B (B7), achieving 1st rank with a Pearson score of 0.53%. We secured 3rd rank in the three subtasks: B15, B10 and C10.

Track A: Two-stage MuRIL setup achieved the best scores for Telugu and Marathi subtasks, while two-stage XLM-R setup achieved the best score for all other subtasks.

Comparison between one-stage and two-stage methods with XLM-R LM. A comparison has been illustrated in Fig. 2. Based on the average performance of all subtasks in track A, it can be inferred that the two-stage strategy outperforms the one-stage strategy. The average score for all subtasks using the two-stage strategy was 0.73, while the average score for all subtasks using the one-stage strategy was 0.61.

Comparison between MuRIL and XLM-R for the Telugu and Marathi Table 2 shows that for Telugu and Marathi, MuRIL performed better than XLM-R. Two-stage MuRIL produces a 0.05 higher score for Telugu subtask than two-stage XLM-R. For Marathi, the Two-stage MuRIL produces slightly more than the Two-stage XLM-R.

[3]scikit-learn.org/stable

Language	B1 afr	B2 arq	B3 amh	B4 eng	B5 hau	B6 hin	B7 ind	B8 kin	B9 arb	B10 ary
Unigram SVR	-	0.4	0.31	0.39	0.41	-	-	0.47	-	0.68
Bigram SVR	-	0.4	0.64	0.32	0.39	-	-	0.35	-	0.55
Unigram XGBoost	-	0.3	0.4	0.28	0.35	-	-	0.38	-	0.72
Bigram XGBoost	-	0.31	0.4	0.33	0.33	-	-	0.37	-	0.72
Dice Loss	0.73	0.44	0.69	0.74	0.42	0.57	0.53	0.36	0.31	0.6
Baseline_Score	0.74	0.43	0.72	0.68	0.16	0.93	0.68	0.74	0.56	0.27
Rank	5	4	5	10	3	6	1	6	6	3

Table 3: All the results for subtasks of Track B have been displayed. For subtasks B1, B7, B8, and B9, labelled data was not provided, so only the Pearson scores predicted by the Dice loss are shown. For the other subtasks, the Pearson scores are displayed for unigram SVR, bigram SVR, unigram XGBoost, bigram XGBoost, and dice loss.

Language	C1 afr	C2 arq	C3 amh	C4 eng	C5 hau	C6 hin	C7 ind	C9 arb	C10 ary	C12 esp
Cross-lingual (XLM-R)	0.7468	0.3867	0.8048	0.7372	0.6428	0.7476	0.4716	0.4267	0.6732	0.5691
Cross-lingual (MuRIL)	-	-	-	-	-	0.8008	-	-	-	-
Baseline_Score	0.79	0.46	0.84	0.8	0.64	0.76	0.47	0.61	0.4	0.62
Rank	7	6	5	6	4	5	5	6	3	9

Table 4: Table shows the results of all subtasks of track C. MuRIL LM support Hindi (C6) subtask of the track C therefore only Pearson score of C6 given for the MuRIL LM.

track B Pearson score of track B tabulated in Table 3. It is clear from Table 3 that SVR performed better than XGBoost. The performance of SVR with unigram and bigram is not straightforward. Results showed that for B4, B5, B8, and B10, bigram embeddings perform better than unigram embeddings. However for the B3 unigram performed better.

track C Pearson score of the track C also showed in Table 4. Table 4 shows that for Hindi (C6), MuRIL performed better than XLM-R. All the other subtasks of this track are only predicted by the XLM-R in cross-lingual settings.

7 Conclusion

In this paper, we utilized multi-lingual track knowledge for the STR shared task to enhance the performance of monolingual models. Our team achieved first rank in the B7 subtask and third rank in the B5, B10, and C10 subtasks. We demonstrate that two-stage fine-tuning can help the monolingual models learn from the training data of all languages, leading to better performance. The results of track C illustrate the effectiveness of cross-lingual learning in a zero-shot scenario.

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.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

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.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Cheng Fu, Bo An, Xianpei Han, and Le Sun. 2016. ISCAS_NLP at SemEval-2016 task 1: Sentence similarity based on support vector regression using multiple features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 645–649, San Diego, California. Association for Computational Linguistics.

Lushan Han, Justin Martineau, Doreen Cheng, and Christopher Thomas. 2015. Samsung: Align-and-differentiate approach to semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 172–177, Denver, Colorado. Association for Computational Linguistics.

Jianglong He, Akshay Uppal, Mamatha N, Shiv Vignesh, Deepak Kumar, and Aditya Kumar Sarda. 2022. Infrrd.ai at SemEval-2022 task 11: A system for named entity recognition using data augmentation, transformer-based sequence labeling model, and EnsembleCRF. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1501–1510, Seattle, United States. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Wenjie Liu, Chengjie Sun, Lei Lin, and Bingquan Liu. 2017. ITNLP-AiKF at SemEval-2017 task 1: Rich features based SVR for semantic textual similarity computing. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 159–163, Vancouver, Canada. 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.

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.

Sumit Singh and Uma Tiwary. 2023. Silp_nlp at SemEval-2023 task 2: Cross-lingual knowledge transfer for mono-lingual learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1183–1189, Toronto, Canada. Association for Computational Linguistics.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 191–197, Vancouver, Canada. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

8 Appendix

8.1 Details of Hyper-parameters

Table 5 shows the details of Hyper-parameters of best models for MuRIL and XLM-R setups with two-stage setup for Track A.

8.2 Comparison of Results of Track A

A comparison of subtasks of track A with one-stage XLM-R and two-stage XLM-R are shown in Fig. 2. For all the subtasks two-stage architecture performed better than one-stage architecture.

Hyper parameters	MuRIL setup	XLM-R setup
Baseline language model for first stage	google/MuRIL-large-cased	XLM-Roberta-large
Loss function	MSE	MSE
Hidden size for language model	1024	1024
Learning rate for language models	5e-05	5e-05
First-stage training epochs	5	5
Second-stage training epochs	5	5
Batch size	64	64
Dropout rate	0.1	0.1
Optimizer	AdamW	AdamW

Table 5: Hyper-parameters for MuRIL and XLM-R setups with two-stage setup for Track A.

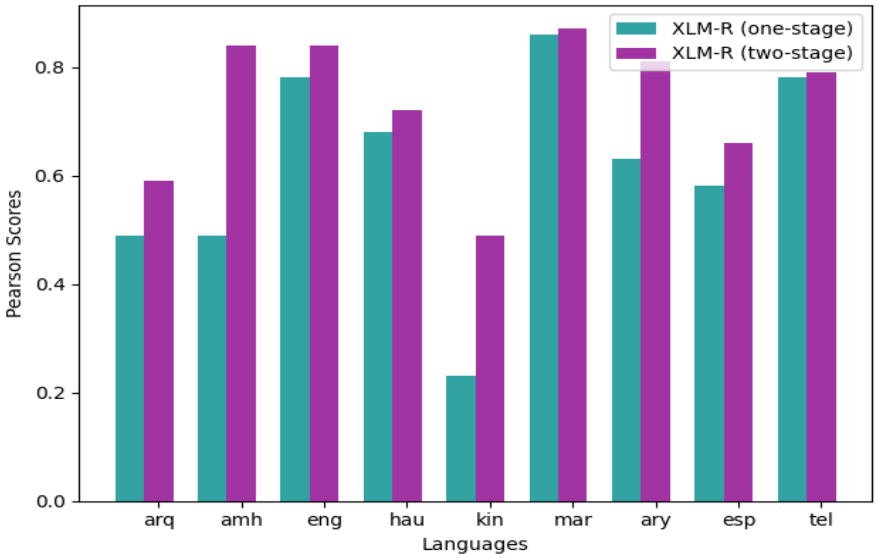

Figure 2: A comparison of subtasks of track A with one-stage XLM-R and two-stage XLM-R.