MaiNLP at SemEval-2024 Task 1: Analyzing Source Language Selection in Cross-Lingual Textual Relatedness

Shijia Zhou[1,*] **Huangyan Shan**[1,*] **Barbara Plank**[1,2] **Robert Litschko**[1,2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
{zhou.shijia, Shan.Huangyan}@campus.lmu.de {bplank, rlitschk}@cis.lmu.de

Abstract

This paper presents our system developed for the SemEval-2024 Task 1: Semantic Textual Relatedness (STR), on Track C: Cross-lingual. The task aims to detect semantic relatedness of two sentences in a given target language without access to direct supervision (i.e. zero-shot cross-lingual transfer). To this end, we focus on different source language selection strategies on two different pre-trained languages models: XLM-R and FURINA. We experiment with 1) single-source transfer and select source languages based on typological similarity, 2) augmenting English training data with the two nearest-neighbor source languages, and 3) multi-source transfer where we compare selecting on all training languages against languages from the same family. We further study machine translation-based data augmentation and the impact of script differences. Our submission achieved the first place in the C8 (Kinyarwanda) test set.

1 Introduction

The task of semantic textual relatedness (STR) has a long-standing tradition in NLP (e.g., Mohammad, 2008). It consists of predicting a score that reflects the closeness in semantic meaning between two given sentences. For example, consider the following examples extracted from the actual shared task data (Abdalla et al., 2023) shown in Figure 1. For English, the annotators scored the first pair higher than the second sentence pair. Similarly, for Afrikaans the annotators scored the first example higher than the second one. As further described in Abdalla et al. (2023), all sentence pairs were annotated manually in a pairwise fashion to obtain semantic textual relatedness (STR) scores between 0 (completely unrelated) and 1 (maximally related).

While previous work has largely focused on English, the SemEval-2024 shared task 1 (Ousidhoum et al., 2024b) aims to extend the language coverage. It proposes datasets to evaluate the relatedness of sentence pairs for a total of 14 languages, including low-resource tail languages such as Kinyarwanda (kin) or Marathi (mar) (Abdalla et al., 2023) (see §2.1). The shared task includes three subtracks, each with a focus on supervised, unsupervised and cross-lingual STR, respectively. In this paper, we focus on Track C, *cross-lingual STR*. In this track, the goal is to develop a system to predict STR scores *without* access to any labeled data for the target language (importantly, also no target development data). That is, Track C requires the development of a regression model for 12 target languages, without relying on any labeled datasets in the target language (or pre-trained language model fine-tuned on other STR tasks). Instead the cross-lingual task allows to utilize training datasets from at least one other language from the other tracks (which includes training data of up to 9 languages). Returning to our running example in Figure 1, the task is to develop a system for example for Afrikaans as target by transferring knowledge from one or more source languages (which may include English).

Pair	STR	Sentence Pair
eng-25	0.88	"It is better known as a walk." "It is also known as a walk ."
eng-31	0.30	"But, of course, it's not that simple" "However, this is not for me."
afr-87	0.72	"ols totdat dit n bal vorm." "Dit moet n stywe bal deeg vorm."
afr-78	0.09	"Stel jou voor jou kind skryf elke week n opstel." "Washington is ook n fietsryer-vriendelike stad."

Figure 1: Examples from the dev sets for Semantic Textual Relatedness (STR). eng: English, afr: Afrikaans.

Previous work on multilingual NLP has illustrated the *curse of multilinguality* (Conneau et al., 2020), that is, diminishing returns for training a

[*] Both authors contributed equally.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1842–1853

single system on many languages due to language interference. This shared task has a focus on low-resource languages and languages typologically distant to English, a setup in where cross-lingual transfer has shown to be particularly challenging (Lauscher et al., 2020). Motivated by these two aspects, we set out to study the use of fewer but more relevant *source* languages for a given target language. More specifically, we aim to find good "donor language(s)" (Malkin et al., 2022) and compare those to baselines that either only use English, or a multi-source model trained on all source languages (except the target). We aim to answer the following research questions: **RQ1** To what extent does knowledge transfer from source languages improve STR models? **RQ2** Do multilingual STR models exhibit language interference (Wang et al., 2020), i.e., performance drops when training data from heterogeneous languages are combined? **RQ3** To what extent do script differences play a role in STR ("script gap"), and can we narrow the script gap by using a foundation model specialized to align transliterated data and data written in different scripts? **RQ4** Can we further improve the transfer performance by relying on machine translation to augment existing training data?

To study RQ1, we make use of typological information available in language vectors. For RQ2, we opt for a multi-source approach, that combines the training data for all languages (except the target). To study the impact of scripts (RQ3), we make use of transliteration, and further compare a BERT-based model to FURINA (Liu et al., 2024), a recently proposed language model that aims to better align languages across scripts. Finally for RQ4, we investigate the use of machine translation (MT) for data augmentation. We apply ours methods to 12 target languages in Track C. The specific details about languages are presented in §2.1.

2 Background

2.1 STR Task Setup and Datasets

The STR task (Ousidhoum et al., 2024b) aims to measure the extent to which two linguistic elements share semantic proximity (Ousidhoum et al., 2024a). These elements may be associated through various means, such as conveying similar ideas, originating from the same historical period, complementing each other's meaning, and so forth. It offers 3 tracks to follow: supervised (Track A), unsupervised (Track B), cross-lingual (Track C). In Track C, participants must provide systems developed without relying on any labeled datasets specifically tailored for semantic similarity or relatedness in the target language. Instead, they are required to employ labeled dataset(s) from at least one other language.

The STR task involves 14 monolingual datasets for Afrikaans (`afr`), Amharic (`amh`), Modern Standard Arabic (`arb`), Algerian Arabic (`arq`), Moroccan Arabic (`ary`), English (`eng`), Spanish (`esp`), Hausa (`hau`), Hindi (`hin`), Indonesian (`ind`), Kinyarwanda (`kin`), Marathi (`mar`), Punjabi (`pan`), and Telugu (`tel`). Among these, Track A and Track C comprise 9 and 12 languages respectively (see Table 1). In the training datasets, each instance consists of a sentence pair and is assigned a golden STR score as judged by native speakers. The score ranges between 0 and 1, with higher values indicating greater relatedness between the sentence pairs. For details on the data collection, we refer the reader to the shared task overview paper (Ousidhoum et al., 2024a).

As per requirement, we designate the 9 languages in Track A as source languages and those in Track C as the 12 target languages. An overview of the resulting train/dev/test data statistics for the 14 languages is provided in Table 1.

2.2 Evaluation Metric

The evaluation metric used in this shared task is Spearman's rank correlation coefficient. It evaluates the strength and direction of the monotonic relationship between two variables with a range from -1 to 1. In the context of our task, as previously mentioned, the scoring has been adjusted to range between 0 and 1. We use the evaluation script provided by the organizers (Ousidhoum et al., 2024b).

2.3 Baselines

The organizers fine-tuned LaBSE (Feng et al., 2022) on the English training set to get baselines for all target languages except English (cf. §3.1). For English, they fine-tuned LaBSE on Spanish as a baseline. Since the test dataset for Spanish has not been made publicly available, all models aimed at Spanish evaluation are conducted solely on their respective validation datasets. In order to ensure a more equitable comparison with other findings, we reproduce the baseline LaBSE model utilizing the methodology provided by the organizers. It yields a baseline score of 0.687 on the Spanish validation

	eng	esp	afr	hin	pan	amh	arb	arq	ary	hau	ind	kin	mar	tel	total
Train	5,500	1,562	-	-	-	992	-	1,261	924	1,736	-	778	1,200	1,170	15,123
Dev	249	139	375	288	242	95	32	97	70	212	144	222	293	130	2,588
Test	2,600	140	375	968	634	171	595	583	425	594	360	222	-	-	7,667

Table 1: STR Dataset statistics. Indo-European lanuguages including esp, afr, hin, ind, pan and mar: 10,424 train instances; 1,811 dev instances; 5,357 test instances. Afro-Asiatic languages including hau, amh, arb, ary and arq: 3,921 train instances; 411 dev instances; 2,197 test instances. Out of 14 languages, 5 languages including amh, hin, arb, arq, ary are in non-latin script, all the rest of languages are in latin script.

dataset.

3 Methods

We opt for two RoBERTa-based (Liu et al., 2019) models for the regression task trained with a mean-squared error (MSE) loss. More specifically, we use the XLM-RoBERTa base model, and FURINA (Liu et al., 2024), which is a XLM-R derivative based on Glot-500 (ImaniGooghari et al., 2023), further detailed below. We adopt a multi-source approach that involves individually fine-tuning a model for each target language in Track C. This fine-tuning process utilizes the training datasets from all languages available in Track A, explicitly excluding the dataset of the test language itself. For baseline comparisons, we use XLM-RoBERTa (Conneau et al., 2020) and FURINA (Liu et al., 2024) models fine-tuned solely on English datasets.

3.1 Model Selection

XLM-RoBERTa. The multilingual masked language model XLM-RoBERTa (XLM-R) (Conneau et al., 2020) pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages has shown superior performance compared to Multilingual BERT (mBERT) (Devlin et al., 2019) across a range of cross-lingual benchmarks. In the experiment, we utilize the base version of XLM-R.[1] XLM-R has seen all SemRelEval languages except for Algerian Arabic (arq), Moroccan Arabic (ary), Kinyarwanda (kin) at pre-training time.

FURINA. FURINA (Liu et al., 2024) covers 511 low-resource languages. It was fine-tuned on Glot500-m (ImaniGooghari et al., 2023). The training data consists of 5% of Glot500-m's pretraining sentences in original script as well as their corresponding Latin transliterations. At pre-training time FURINA has been exposed to all SemRelEval languages except for Algerian Arabic (arq).

LaBSE. The organizers provide cross-lingual baselines for each target language by fine-tuning Language-agnostic BERT Sentence Embeddings (`LaBSE`) (Feng et al., 2022), which supports 109 languages. LaBSE was pre-trained using Translation language modeling (TLM) (Conneau and Lample, 2019), which included bilingual translation sentence pairs for training. The bilingual corpus is constructed from web pages using a bitext mining system, filtered by a pre-trained contrastive data-selection scoring model, and manually curated to create a high-quality collection of 6 billion translation pairs. Out of those, LaBSE has been exposed to different amounts of parallel data (`eng-xxx`) from SemRelEval languages. The largest amount of parallel text involves Spanish with over 375M sentence pairs (`eng-esp`), followed by Indonesian with over 250M sentence pairs (eng-ind), followed by Hindi and Arabic (`eng-{hin, arb}`) with over 125M language pairs. All other languages (`afr`, `pan`, `amh`, `haus`, `tel`, `kin`, `mar`) appear in the TLM training corpus with less than 125M sentence pairs.

3.2 Source Language Selection

Single-Source Transfer. In our first approach, we follow the standard single-source zero-shot cross-lingual transfer setup and fine-tune pre-trained language models on English data ($\texttt{XLM-R}_{\texttt{eng}}$, $\texttt{Furina}_{\texttt{eng}}$). This is a common evaluation approach adopted in standard natural language understanding and generation benchmarks (Liang et al., 2020; Ruder et al., 2023). However, English has been shown to not always be the best source language (Turc et al., 2021). To investigate if this also true for SemRelEval, we further experiment with selecting for each test language its closest (i.e., most similar) source language. Here, we measure language similarity according to typological features from the lang2vec library (Littell et al., 2017).

[1] https://huggingface.co/FacebookAI/xlm-roberta-base

K-nearest-neighbor languages. In this approach we augment the English training dataset with the datasets of k languages that are closest to the target language, dubbed kNN. To determine suitable source languages for each target language, we assess language similarity by calculating the cosine similarity between language vectors learned by a multilingual neural MT model provided by Malaviya et al. (2017). We specifically use the cell_state language vectors, which are computed by encoding all sentences in a given language and then computing the average hidden cell state of the encoder LSTM.[2] These vectors can be seen as language embeddings encoding latent typology features (Östling and Tiedemann, 2017; Yu et al., 2021). With our kNN-models we aim for a good balance between large amounts of training instances (English) and positive transfer from similar languages.

Multi-Source Transfer. The STR dataset contains languages from different language families. To investigate whether training a single model on a disverse set of languages leads to negative interference (Wang et al., 2020) we compare two multi-source models. In the first model, dubbed MS-All, we fine-tune XLM-R and Furina on the concatenation of all training sets from Track A (excluding the target language). Inspired by previous work on combining *multiple related* source languages (Snæbjarnarson et al., 2023; Lim et al., 2024), we further evaluate multi-source models trained on languages from the same language family (MS-Fam).

3.3 Other Approaches

Machine Translation. For the purpose of data augmentation and balance of languages, we translate selected languages into each other using NLLB (Costa-jussà et al., 2022), ensuring that each language contributes equally to the training dataset. Taking Kinyarwanda as an example, we select Hausa and Spanish as the two languages closest to it, based on dense language vector similarity as outlined above (kNN), along with English, as training dataset. We translate among these three languages mutually, thus tripling the size of the training dataset while ensuring a balanced representation of all languages.

Transliteration. Additionally, we attempt to further facilitate multilingual transfer learning by standardizing script across languages. Utilizing the tool Uroman[3] (Hermjakob et al., 2018), which was also used by FURINA (Liu et al., 2024), we transliterate the train and test datasets of languages written in non-Latin scripts, including both the original datasets and the translated datasets, into Latin script. We evaluate the models fine-tuned on Romanized training data on the Romanized test dataset. This attempt only involves non-Latin script languages (amh, arb, ary, arq, hin).

4 Experimental Setup

The detailed setting are listed in Appendix A. As baseline, we exclusively train a model on the English dataset ($\texttt{XLM-R}_{\texttt{eng}}$, $\texttt{Furina}_{\texttt{eng}}$) and assess its performance across all target languages. Subsequently, for each target language, we fine-tune a multi-source model: if the target language is not within the 9 training datasets, we train on the union of all $n = 9$ training languages. Otherwise we train a multi-source model on $n - 1 = 8$ source languages, excluding the target ($\texttt{XLM-R}_{\texttt{MS-All}}$, $\texttt{Furina}_{\texttt{MS-All}}$). Following this, we explore whether it is helpful to prune certain languages, retaining only English and the two closest to the target languages according to lang2vec (Littell et al., 2017)[4] as source languages ($\texttt{XLM-R}_{\texttt{L2V}}$, $\texttt{Furina}_{\texttt{L2V}}$). Due to the reduction in the training set, which significantly decreased the size of the data, we attempted to expand the dataset through cross-translation ($\texttt{XLM-R}_{\texttt{L2V-Aug}}$, $\texttt{Furina}_{\texttt{L2V-Aug}}$; cf. §3.2).

5 Results and Discussion

Our main results are presented in Table 2 and are discussed in the following section.

Single-source versus multi-source transfer. We first compare the performance of a zero-shot STR model trained on English ($\texttt{XLM-R}_{\texttt{eng}}$, $\texttt{Furina}_{\texttt{eng}}$) against a multi-source model trained on the concatenation of all available languages from Track A ($\texttt{XLM-R}_{\texttt{MS-All}}$, $\texttt{Furina}_{\texttt{MS-All}}$). Our results reveal that knowledge transfer from multiple source languages (**RQ1**) improves STR models, affirming the potential of multi-source training to enhance cross-lingual capabilities. On average, both MS-All models outperform their single-source counterparts by 0.02 and 0.09 respectively. This is expected since

[2] https://github.com/chaitanyamalaviya/lang-reps/

[3] https://github.com/isi-nlp/uroman

[4] We compare the similarity of languages based on three criteria: lang_cell_states, lang_vecs and language typological vectors

	Indo-European					Afro-Asiatic					Other		
	eng	esp	afr	hin	pan	amh	arb	arq	ary	hau	ind	kin	avg
LaBSE (baseline)	0.80	0.69	0.79	0.76	-0.05	**0.84**	**0.61**	0.46	0.40	0.62	**0.47**	0.57	0.67
$\text{Furina}_{\text{eng+esp+hau}}$	-	-	0.74	0.70	**0.09**	0.73	0.40	0.27	0.57	-	0.32	0.68	-
Models based on XLM-R (Conneau et al., 2020)													
$\text{XLM-R}_{\text{eng}}$	-	0.67	**0.81**	0.80	-0.02	0.81	0.60	0.50	0.60	0.64	0.42	0.46	0.71
$\text{XLM-R}_{\text{MS-All}}$	**0.84**	0.63	0.80	**0.82**	-0.01	0.80	0.56	0.59	0.82	0.66	0.42	0.69	**0.73**
$\text{XLM-R}_{\text{MS-Fam}}$	0.82	0.71	**0.81**	**0.82**	0.00	0.69	0.44	0.37	**0.83**	0.66	-	-	0.68
$\text{XLM-R}_{\text{kNN}}$	-	0.59	**0.81**	0.78	-	0.75	0.57	-	0.50	0.62	0.45	0.41	0.69
$\text{XLM-R}_{\text{kNN+MT}}$	-	0.64	0.80	0.78	-	0.77	0.54	-	0.55	0.62	0.36	0.55	0.70
$\text{XLM-R}_{\text{kNN+TL}}$	-	-	-	0.66	-	0.37	0.45	-	0.52	-	-	-	-
Models based on Furina (Liu et al., 2024)													
$\text{Furina}_{\text{eng}}$	-	0.54	0.79	0.70	-0.14	0.74	0.37	0.45	0.59	0.63	0.44	0.53	0.62
$\text{Furina}_{\text{MS-All}}$	0.83	0.59	0.79	0.76	-0.02	0.81	0.49	**0.61**	**0.83**	0.65	0.35	**0.78**	0.71
$\text{Furina}_{\text{MS-Fam}}$	0.83	**0.72**	0.79	0.77	0.02	0.66	0.42	0.55	0.82	**0.68**	-	-	0.71
$\text{Furina}_{\text{kNN}}$	-	0.59	0.80	0.72	-	0.74	0.43	-	0.57	0.63	0.46	0.68	0.67
$\text{Furina}_{\text{kNN+MT}}$	-	0.56	0.78	0.75	-	0.74	0.44	-	0.57	0.59	0.37	0.64	0.67
$\text{Furina}_{\text{kNN+TL}}$	-	-	-	0.67	-	0.72	0.44	-	0.56	-	-	-	-

Table 2: Spearman's rank correlation of zero-shot transfer experiments on SemRelEval 9 test languages. The organizers decided to keep the test set for Spanish private, we therefore report the performance on the validation set. We exclude English from the average result (avg). **bold**: Best result for each language. Languages not covered by all L2V features are excluded from the average (eng, pan, arq, ind, kin). For our kNN-variants we opt for $k = 2$.

the multi-source training dataset is with 15,123 instances almost three times larger than the English dataset with 5,500 instances (cf. Table 1). When trained solely on English data, FURINA performs substantially worse than XLM-R. However, this performance gap narrows when transitioning from single-source to multi-source training.

Transfer from language families. After showing that models trained on all languages outperform the single-source baseline, we now investigate the effect of training on languages from the same family as source languages. Here we experiment with two multi-source models specialized only on Indo-European and Afro-Asiatic languages respectively (MS-Fam). Importantly, for each target language we train a multi-source model on all other languages in the same language family.[5] On Indo-European languages, we find that $\text{XLM-R}_{\text{MS-Fam}}$ and $\text{Furina}_{\text{MS-Fam}}$ yield similar results with much less training data (i.e., 4,913 fewer instances belonging to other language families). For Spanish, our models show performance gains of +0.8 and +0.13 for XLM-R and FURINA respectively, when compared to models trained on all languages. This underscores the presence of language interference (Wang et al., 2020) in multilingual STR models when the training data from dissimilar languages are combined (**RQ2**). On Afro-Asiatic languages, we observe average performance drops of -0.09 and -0.06 for XLM-R and FURINA when moving from MS-All to MS-Fam. We hypothesize that this can be attributed to the amount of training data available. In fact, there are 28% fewer training instances for all Afro-Asiatic languages than for English (5,500).

Transfer from nearest language neighbors. We now investigate the transfer performance when training STR models on their two closest languages according to cosine similarity of language cell state vectors, i.e. learned language vectors presented in (Malaviya et al., 2017). As mentioned earlier, we add English due to its large scale as a third training language. Our submitted system, $\text{Furina}_{\text{eng+esp+hau}}$, is trained on the two closest training languages of Kinyarwanda (kin) and has been ranked first place on the shared task leaderboard. Applying the same approach for each test language ($\text{XLM-R}_{\text{kNN}}$, $\text{Furina}_{\text{kNN}}$) shows mixed results. This indicates that the strong performance on kin can be attributed to the fact that, contrary to XLM-R, kin has been seen by Furina during pretraining.

Transliteration and cross-translation. The STR dataset contains six test languages in non-Latin scripts: Hindi (hin), Punjabi (pan), Amharic

[5] Indonesian and Kinyarwanda are the only SemRel languages in their family, we therefore cannot evaluate multi-source for those languages.

	Indo-European					Afro-Asiatic					Other		
XLM-R	eng	esp	afr	hin	pan	amh	arb	arq	ary	hau	ind	kin	avg
MIN	0.78	0.57	0.74	0.71	-0.14	0.73	0.47	0.39	0.40	0.40	0.31	0.41	0.58
XLM-R$_{eng}$	-	0.67	0.81	0.80	-0.02	0.81	0.60	0.50	0.60	0.64	0.42	0.46	0.71
XLM-R$_{kNN}$	-	0.68	0.74	0.72	-	0.75	0.57	-	0.40	0.63	0.49	0.43	0.64
L2V-Pho	0.78	0.67	0.80	0.80	-0.03	0.78	0.51	0.58	0.55	0.58	0.39	0.47	0.67
L2V-Syn	0.78	0.67	0.83	0.80	-0.03	0.74	0.51	0.58	0.55	0.61	0.39	0.43	0.67
L2V-Inv	0.82	0.63	0.80	0.80	-0.03	0.79	0.51	0.58	0.55	0.58	0.33	0.45	0.67
L2V-Fam	0.82	0.67	0.83	0.80	-0.03	0.79	0.51	0.58	0.55	0.595	-	-	0.68
L2V-Geo	0.78	0.57	0.80	0.80	-0.03	0.75	0.47	0.56	0.55	0.63	0.31	0.45	0.66
L2V-LRN	-	0.68	0.74	0.72	-	0.79	0.57	-	0.54	0.40	0.49	0.43	0.63
MAX	0.82	0.69	0.83	0.80	0.04	0.79	0.61	0.63	0.74	0.66	0.49	0.65	0.73

Table 3: Single-source transfer results in terms of spearman correlation. The language selection is based on the cosine similarity of different typological features obtained from lang2vec (L2V). We additionally report the lower (MIN) and upper bound (MAX) obtained from selecting the best and worst source language. Languages not covered by all L2V features are excluded from the average: eng, pan, arq, ind, kin. For L2V-Phon, both tel and mar are closest to hin. For L2V-Fam amh and arq are the closest languages, we report their average score (0.595). In single-source transfer with XLM-R$_{kNN}$ we use $k = 1$ and do not combine the selected language with eng training data.

(amh), Standard Arabic (arb), Algerian Arabic (arq), and Moroccan Arabic (ary). Zero-shot cross-lingual transfer of models fine-tuned on English performs worse for Arabic scripts than for amh and hin. Punjabi shows the lowest results by a large margin. When fine-tuned on on multiple source languages (MS-All), XLM-R improves the performance on four out of six languages while Furina yields improvements on all five languages. We find that (1) there is no clear winner between XLM-R and FURINA when applied on text written in different scripts, and (2) romanizing all training and test languages did not improve zero-shot cross-lingual transfer for STR (**RQ3**).

Next, we investigate the impact of augmenting the training data with translated data. The varied outcomes of augmenting data indicate that while machine translation can enhance transfer performance for certain languages. Performance drops in others may stem from shifts in label semantics and the degree of relatedness between original and translated sentence pairs (**RQ4**). Appendix C (Table 14) shows an example where MT fails to capture nuanced differences between closely, but not perfectly related sentences, leading to near-identical translations and inconsistent labels.

Single-source transfer results. We now select the most similar source languages based on different typological features obtained from the lang2vec (L2V) library. We obtain L2V vectors for Phonology (Pho), Syntax (Syn), Inventory (Inv), Family (Fam), Geography (Geo) and learned (LRN) features.

Table 3 shows our results for XLM-R.[6] Overall, a careful selection of a single-source language is crucial for zero-shot cross-lingual transfer. There is a substantial gap between the worst possible result (0.58) and the best possible result (0.73). On average, English is the most effective source language with a correlation of 0.71. A closer analysis reveals that English is the best language only for half of the target languages, despite being the language with the largest training dataset (cf. Table 13 in Appendix). Interestingly, the best possible single-source language selection (MAX) results into the same performance as XLM-R$_{MS-All}$ (cf. Table 2).

6 Conclusion

In this paper, we investigate source language selection for cross-lingual transfer for Semantic Textual Relatedness (STR). We evaluate three different language selection strategies: single-source, multi-source transfer and transfer from English and two nearest language neighbors. We find that the transfer performance crucially depends on the size of the training dataset and the linguistic proximity to the test language. We further show that script differences cause high variance transfer performance and MT-based data augmentation can lead to shifts in label semantics. Fine-tuning FURINA on eng, esp, and hau, we achieve first place in the SemEval-2024 Task 1, Track C8 (kin).

[6] FURINA results can be found in Appendix Table 12.

Acknowledgements

This research is supported by the ERC Consolidator Grant DIALECT 101043235.

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.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, volume 32, pages 7059–7069. Curran Associates, Inc.

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.

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.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. 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.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Seong Hoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. *arXiv preprint arXiv:2402.13562*.

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.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schütze. 2024. Translico: A contrastive learning framework to address the script barrier in multilingual pretrained language models.

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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Saif Mohammad. 2008. *Measuring semantic distance using distributional profiles of concepts*. University of Toronto.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. 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. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Dian Yu, Taiqi He, and Kenji Sagae. 2021. Language embeddings for typology and cross-lingual transfer learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7210–7225, Online. Association for Computational Linguistics.

A Hyperparameters

We employed identical hyperparameters across all variants of XLM-R and FURINA. We train our models for at most 30 epochs with a batch size of 32 and a learning rate of 2e-5 and use AdamW (Loshchilov and Hutter, 2017) with a weight decay of 1e-3. We evaluate the dev set performance every 200 steps and stop early based on the spearman correlation on the validation set (patience counter: 8, patience threshold: 1e-4).

B Language Similarities

Table 4 shows for each test language its two closest source languages (`kNN`) according to cell state vectors from (Malaviya et al., 2017) and learned vectors from lang2vec (`L2V-LRN`) (Littell et al., 2017). We find both language vectors lead to similar results. Here, we further show the selected languages for our multi-source model (`MS-Fam`), which outperforms both `L2V-LRN` and `kNN`.

In Table 5-11 we show cosine similarities between all train and test languages according to different typological features extracted from L2V and learned vectors from (Malaviya et al., 2017). We use the similarities to select source languages for our `kNN` and single-source model variants.

Model Variant	Source languages	Target language	# Train Instances	FURINA	XLM-R
Based on cell state vectors (kNN) (Malaviya et al., 2017)					
1	esp, kin	afr hau	7840	0.80 0.63	0.81 0.62
2	esp, hau	ind kin	8798	0.46 0.68	0.45 0.41
3	kin, hau	amh esp	8014	0.74 0.59	0.75 0.59
4	amh, hau	ary	8228	0.57	0.50
5	kin, amh	arb	7270	0.44	0.57
6	amh, esp	hin	8054	0.72	0.78
avg	-	-	-	0.58	0.58
Based on learned lang2vec vectors (L2V-LRN) (Littell et al., 2017)					
1	esp, kin	afr arb hau ind	7840	0.80 0.46 0.63 0.44	0.81 0.60 0.62 0.39
2	esp, hau	kin	8798	0.68	0.41
3	kin, hau	amh esp	8014	0.74 0.59	0.75 0.59
4	amh, hau	hin	8228	0.74	0.79
5	kin, amh	ary	7270	0.52	0.55
avg	-	-	-	0.59	0.60
Based on language familis features (MS-Fam)					
1	esp, mar, tel	eng	3932	0.83	0.82
2	eng, mar, tel	esp	7370	0.72	0.71
3	eng, esp, mar, tel	afr hin pan	9432	0.79 0.77 0.02	0.81 0.82 -0.00
4	arq, ary, hau	amh	3921	0.66	0.69
5	amh, arq, ary, hau	arb	4913	0.42	0.44
6	amh, ary, hau	arq	3652	0.55	0.37
7	amh, arq, hau	ary	3989	0.82	0.83
8	amh, arq, ary	hau	3117	0.68	0.66

Table 4: Model variants based on language vectors, language cell state vectors and language families. All variants include eng for training.

	amh	ary	esp	hau	kin
afr	0.75	0.57	0.83	0.79	0.82
amh	-	0.62	0.66	0.69	0.71
ary	0.62	-	0.54	0.61	0.49
arb	0.76	0.73	0.73	0.73	0.79
esp	0.66	0.54	-	0.76	0.82
hau	0.69	0.61	0.76	-	0.84
hin	0.80	0.73	0.74	0.72	0.71
ind	0.71	0.65	0.83	0.76	0.76
kin	0.71	0.49	0.82	0.84	-

Table 5: Cosine similarities between source languages (columns) and target languages (rows). Language vectors are obtained from lang2vec: **kNN** (cell_state vectors) (Malaviya et al., 2017). We exclude four languages for which we cannot obtain feature vectors: arq, mar, tel, eng.

	amh	ary	esp	hau	kin
afr	0.07	-0.05	0.23	0.07	0.22
amh	-	-0.01	0.00	0.07	0.05
ary	-0.01	-	-0.06	-0.03	0.06
arb	0.07	-0.05	0.13	-0.03	0.11
esp	0.00	-0.06	-	0.22	0.23
hau	0.07	-0.03	0.22	-	0.19
hin	0.13	-0.01	0.06	0.07	0.06
ind	0.00	0.05	0.11	0.06	0.09
kin	0.05	0.06	0.23	0.19	-

Table 6: Cosine similarities between source languages (columns) and target languages (rows). Language vectors are obtained from lang2vec: `L2V-LRN` (Littell et al., 2017). We exclude four languages for which we cannot obtain **`L2V-LRN`** features: arq, mar, tel, eng.

	amh	ary	esp	hau	kin	arq	mar	tel	eng
afr	0.86	0.70	0.76	0.87	0.85	0.73	0.80	0.80	0.82
amh	-	0.73	0.80	0.82	0.78	0.76	0.95	0.84	0.76
ary	0.73	-	0.73	0.67	0.73	0.97	0.77	0.69	0.70
arb	0.85	0.90	0.76	0.77	0.76	0.93	0.80	0.71	0.73
esp	0.80	0.73	-	0.73	0.78	0.76	0.84	0.74	0.86
hau	0.82	0.67	0.73	-	0.82	0.69	0.77	0.77	0.78
hin	0.82	0.75	0.82	0.75	0.82	0.77	0.87	0.87	0.78
ind	0.76	0.70	0.76	0.78	0.85	0.73	0.80	0.80	0.91
kin	0.78	0.73	0.78	0.82	-	0.76	0.82	0.82	0.85
arq	0.76	0.97	0.76	0.69	0.76	-	0.80	0.71	0.73
eng	0.76	0.70	0.86	0.78	0.85	0.73	0.80	0.80	-
pan	0.95	0.77	0.84	0.77	0.82	0.80	1.00	0.89	0.80

Table 7: Cosine similarities between source languages (columns) and target languages (rows). Language vectors are obtained from lang2vec: **`L2V-Phon`** (Littell et al., 2017).

	amh	ary	esp	hau	kin	arq	mar	tel	eng
afr	0.62	0.66	0.73	0.71	0.55	0.67	0.62	0.56	0.85
amh	-	0.59	0.63	0.57	0.51	0.60	0.72	0.77	0.59
ary	0.59	-	0.81	0.72	0.63	0.93	0.50	0.48	0.73
arb	0.61	0.87	0.75	0.64	0.64	0.85	0.49	0.50	0.64
esp	0.63	0.81	-	0.74	0.59	0.81	0.56	0.52	0.82
hau	0.57	0.72	0.74	-	0.65	0.78	0.52	0.34	0.75
hin	0.74	0.67	0.68	0.57	0.46	0.65	0.83	0.78	0.62
ind	0.45	0.73	0.66	0.67	0.52	0.74	0.36	0.32	0.73
kin	0.51	0.63	0.59	0.65	-	0.64	0.39	0.38	0.49
arq	0.60	0.93	0.81	0.78	0.64	-	0.49	0.47	0.74
eng	0.59	0.73	0.82	0.75	0.49	0.74	0.56	0.52	-
pan	0.71	0.68	0.70	0.59	0.49	0.67	0.79	0.75	0.61

Table 8: Cosine similarities between source languages (columns) and target languages (rows). Language vectors are obtained from lang2vec: **`L2V-Syn`** (Littell et al., 2017).

	amh	ary	esp	hau	kin	arq	mar	tel	eng
afr	0.65	0.56	0.62	0.61	0.69	0.61	0.67	0.68	0.69
amh	-	0.76	0.74	0.83	0.80	0.73	0.73	0.64	0.70
ary	0.76	-	0.62	0.70	0.70	0.85	0.63	0.57	0.65
arb	0.72	0.83	0.65	0.70	0.71	0.98	0.64	0.60	0.73
esp	0.74	0.62	-	0.67	0.68	0.64	0.66	0.66	0.64
hau	0.83	0.70	0.67	-	0.76	0.72	0.64	0.59	0.62
hin	0.66	0.69	0.57	0.62	0.69	0.77	0.72	0.77	0.71
ind	0.88	0.75	0.76	0.79	0.82	0.77	0.74	0.68	0.76
kin	0.80	0.70	0.68	0.76	-	0.72	0.65	0.63	0.69
arq	0.73	0.85	0.64	0.72	0.72	-	0.65	0.62	0.71
eng	0.70	0.65	0.64	0.62	0.69	0.71	0.76	0.67	-
pan	0.71	0.60	0.69	0.65	0.71	0.66	0.82	0.78	0.77

Table 9: Cosine similarities between source languages (columns) and target languages (rows). Language vectors are obtained from lang2vec: **L2V-Inv** (Littell et al., 2017).

	amh	ary	esp	hau	kin	arq	mar	tel	eng
afr	0.00	0.00	0.11	0.00	0.00	0.00	0.15	0.00	0.50
amh	-	0.40	0.00	0.17	0.00	0.43	0.00	0.00	0.00
ary	0.40	-	0.00	0.16	0.00	0.94	0.00	0.00	0.00
arb	0.46	0.87	0.00	0.18	0.00	0.93	0.00	0.00	0.00
esp	0.00	0.00	-	0.00	0.00	0.00	0.12	0.00	0.10
hau	0.17	0.16	0.00	-	0.00	0.17	0.00	0.00	0.00
hin	0.00	0.00	0.11	0.00	0.00	0.00	0.46	0.00	0.13
ind	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
kin	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00
arq	0.43	0.94	0.00	0.17	0.00	-	0.00	0.00	0.00
eng	0.00	0.00	0.10	0.00	0.00	0.00	0.14	0.00	-
pan	0.00	0.00	0.12	0.00	0.00	0.00	0.50	0.00	0.14

Table 10: Cosine similarities between source languages (columns) and target languages (rows). Language vectors are obtained from lang2vec: **L2V-Fam** (Littell et al., 2017).

	amh	ary	esp	hau	kin	arq	mar	tel	eng
afr	0.97	0.91	0.90	0.96	0.99	0.91	0.92	0.92	0.87
amh	-	0.95	0.95	0.98	0.99	0.96	0.97	0.96	0.94
ary	0.95	-	1.00	0.98	0.94	1.00	0.88	0.87	0.99
arb	0.99	0.95	0.96	0.97	0.97	0.97	0.98	0.97	0.96
esp	0.95	1.00	-	0.98	0.94	1.00	0.90	0.89	1.00
hau	0.98	0.98	0.98	-	0.99	0.98	0.90	0.90	0.96
hin	0.97	0.89	0.91	0.90	0.94	0.91	1.00	1.00	0.91
ind	0.89	0.77	0.79	0.81	0.87	0.79	0.96	0.96	0.79
kin	0.99	0.94	0.94	0.99	-	0.95	0.94	0.94	0.92
arq	0.96	1.00	1.00	0.98	0.95	-	0.90	0.90	0.99
eng	0.94	0.99	1.00	0.96	0.92	0.99	0.90	0.89	-
pan	0.96	0.90	0.91	0.91	0.93	0.92	1.00	1.00	0.92

Table 11: Cosine similarities between source languages (columns) and target languages (rows). Language vectors are obtained from lang2vec: **L2V-Geo** (Littell et al., 2017).

	Indo-European					Afro-Asiatic					Other		
FURINA	eng	esp	afr	hin	pan	amh	arb	arq	ary	hau	ind	kin	avg
MIN	0.34	0.38	0.48	0.35	-0.19	0.68	0.04	0.00	0.28	0.33	0.22	0.23	0.36
Furina$_{eng}$	-	0.54	0.79	0.70	-0.14	0.74	0.37	0.45	0.59	0.63	0.44	0.53	0.62
Furina$_{L2V\text{-}kNN}$	-	0.62	0.71	0.35	-	0.73	0.42	-	0.28	0.64	0.42	0.68	0.53
L2V-Pho	0.76	0.56	0.79	0.77	0.03	0.76	0.46	0.48	0.63	0.43	0.43	0.68	0.63
L2V-Syn	0.76	0.56	0.80	0.78	0.03	0.74	0.39	0.48	0.63	0.54	0.34	0.68	0.63
L2V-Inv	0.78	0.38	0.79	0.76	0.03	0.76	0.46	0.48	0.63	0.43	0.22	0.23	0.60
L2V-Fam	0.78	0.64	0.80	0.78	0.03	0.76	0.46	0.48	0.63	0.485	-	-	0.65
L2V-Geo	0.76	0.47	0.79	0.78	0.03	0.73	0.04	0.53	0.63	0.64	0.30	0.23	0.58
L2V-LRN	-	0.62	0.71	0.35	-	0.76	0.42	-	0.59	0.33	0.42	0.54	0.54
MAX	0.79	0.64	0.81	0.78	0.06	0.76	0.53	0.55	0.77	0.66	0.45	0.78	0.71

Table 12: Single-source transfer results in terms of spearman correlation. The language selection is based on the cosine similarity of different typological features obtained from lang2vec (L2V). We additionally report the lower and upper bound (MIN, MAX) when choosing the worst and best possible donor language for each test language. Languages that are not covered by all L2V features are excluded from the average (eng, pan, arq, ind, kin).

	afr	amh	ary	arb	esp	hau	hin	ind	kin	arq	eng	pan
MIN (XLM-R)	esp	esp	amh	amh	ary	esp	esp	tel	arq	amh	esp	ary
MIN (Furina)	amh	esp	amh	amh	amh	esp	amh	amh	amh	amh	amh	ary
kNN	esp	kin	amh	kin	kin	kin	amh	esp	hau	-	-	-
L2V-Pho	hau	mar	arq	arq	eng	amh	mar+tel	eng	eng	ary	esp	mar
L2V-Syn	eng	tel	arq	ary	eng	arq	mar	arq	hau	ary	esp	mar
L2V-Inv	kin	hau	arq	arq	amh	amh	tel	amh	amh	ary	mar	mar
L2V-Fam	eng	arq	arq	arq	mar	amh+arq	mar	-	-	ary	mar	mar
L2V-Geo	kin	kin	arq	amh	arq	kin	mar	tel	amh	esp	esp	mar
L2V-LRN	esp	hau	kin	kin	kin	esp	amh	esp	esp	-	-	-
MAX (XLM-R)	eng	eng	eng	mar	hau	eng	eng	esp	mar	eng	mar	amh
MAX (Furina)	mar	arq	eng	eng	mar	mar	mar	ary	mar	mar	hau	kin

Table 13: Each cell shows a given test language and lang2vec (L2V) feature the closest source language used for single source transfer in Table 3 and Table 12. We further show the closest languages according to cell-state vectors obtained from a multilingual MT system (kNN) (Malaviya et al., 2017), see §3.2 for details. MIN and MAX show the source language for which best transfer and worst transfer performance is achieved.

Pair	**Sentence Pair**
esp-182	"Un hombre está saltando a una pared baja." "Un hombre está saltando a un muro bajo."
translated	"A man is jumping into a low wall." "A man is jumping into a low wall"

Table 14: An example from Spanish training dataset with its English translation, the label is 0.80.

C Translation quality.

We reviewed some machine-translated examples and noticed that subtle differences in the original language can be lost during translation. As shown in Table 14, the two translated sentences, apart from punctuation, share no differences while the label assigned is 0.8. This undoubtedly has the potential to interfere with the model's learning process for the STR task.