Massively Cross-Language Understanding

Anonymous ACL submission

Abstract

Natural Language Inference (NLI) is one of the standard tasks researchers use to benchmark the language understanding capability of language models. Traditionally, NLI has the premise and hypothesis in the same language, with existing datasets covering 15 languages in a monolingual setting. A cross-language variation, where they are in different languages, is a mostly unexplored task that tests the capabilities of models to understand and correlate text from different languages at once. In this work, we 1) create a cross-language entailment dataset built on existing entailment datasets and expand it to 93 languages, 2) test and provide baselines for the cross-language reasoning capability of large masked language models, and 3) investigate the cross-lingual transfer ability of our dataset. Overall, we found that models perform worse in a cross-language setting than they do monolingually, with performance degrading as we scale up the number of languages. Finally, we show that using our dataset achieves greater cross-lingual transfer than monolingual data does. This work sheds light on the challenges and opportunities for enhancing the cross-language reasoning abilities of language models and invites further exploration of this task.

1 Introduction

Natural Language Inference (NLI) is a task that looks at reasoning about hypotheses given premises. A hypothesis sentence is classified as "entailing" if it necessarily follows from the premise, "contradiction" if it necessarily *does not* follow, and "neutral" otherwise. The concept of entailment draws from the linguistic study of semantics and is a strict relation; the hypothesis is not entailed if it is very likely to be true given the premise, only if it must be so. This task is used to examine the ability of models to understand and reason about language. Most existing datasets for NLI are entirely in English, with monolingual premises and hypotheses (Bowman et al., 2015; Khot et al., 2018). One popular benchmark NLI dataset is MNLI (Williams et al., 2018), which consists of 433k pairs drawn from 10 different genres of text. This is not the largest available dataset, but the greater variety in styles of text compared to alternatives promotes more robust models.

The largest multilingual dataset, with pairs in multiple languages, is XNLI (Conneau et al., 2018), which provides 7.5k English pairs in the style of MNLI, along with human translations in 14 other languages. Training for XNLI typically uses MNLI, and machine translates that corpus into other languages as needed. Evaluation is done monolingually on each of the 15 languages, with average accuracy often reported as an overall score. To our knowledge, there is no large NLI dataset in which the premise and hypothesis are in different languages, a setting which we will refer to as Cross-Language Inference (CLI).

XNLI is a standard benchmark for cross-lingual transfer, where a model is trained for a task in one language and evaluated on multiple others. There is no equivalent standard for cross-language capabilities, where models must process text in multiple languages at once. There are many tasks where such reasoning ability is necessary such as cross-language information retrieval (CLIR) and cross-language question answering. This is particularly important in low-resource settings, where answering a query in a low-resource language may require consulting sources in other languages due to data sparsity. Models that struggle with cross-language reasoning may in turn be less suited to such tasks. Cross-language entailment can be viewed as proxy to compare the overall cross-language ability of different models, and thus their suitability for these problems.

We introduce CLI-93, a cross-language entailment silver dataset built from machine translation of MNLI. CLI-93 extends the entirety of MNLI ($\sim 430k$ examples) to be cross-language over 93

languages, all contained in XLM-RoBERTa (XLM-R), a large pretrained multilingual masked language model (Conneau et al., 2020). Our dataset adds 79 languages to those in XNLI. Additionally, we introduce CLI-15, which uses the same methodology, but only expands MNLI to the 15 languages in XNLI. CLI-15 is presented as a less challenging variation of the dataset that maintains compatibility with the gold XNLI test set for evaluation purposes. We perform an investigation into translation quality to ensure that the data quality is acceptable. We also train baseline models to set basic benchmarks for these datasets. Finally, we investigate the cross-lingual transfer that models can achieve using our dataset. Overall, our contributions are that we a) significantly increase the number of languages that have NLI data, many of which are low-resource and under-studied, b) show that working in a cross-language setting is more difficult than dealing with monolingual tasks for current models, and c) show that our datasets can be used to achieve greater cross-lingual transfer than purely monolingual data of the same size. Both datasets will be released through HuggingFace.

2 Related Works

SNLI (Bowman et al., 2015) is an English-only NLI dataset consisting of 570k premise-hypothesis pairs. It was created by using an open-source image caption dataset for premises, and having Amazon Mechanical Turk workers write hypotheses for them corresponding to each label. Each entry has five annotator judgements that are used to assign a gold label. MNLI is a successor dataset that uses a similar procedure, but draws premises from a greater variety of sources. There are fewer examples, but the data is of higher quality.

Non-English resources for NLI beyond XNLI (Conneau et al., 2018) are limited. TERRa (Shavrina et al., 2020) is a dataset for entailment in Russian, consisting of $\sim 6k$ pairs. There are also datasets on the order of 10k pairs for Dutch (Wijnholds and Moortgat, 2021) and Portuguese (Fonseca et al., 2016). For multilingual data, Agić and Schluter (2018) create a dataset by manually translating 1,332 English pairs into 4 other languages. Kumar Upadhyay and Kumar Upadhya (2023) produce multilingual training data by using newer translation models to translate MNLI into the 14 XNLI languages than the original XNLI corpus provided. To our knowledge these, along with XNLI, are the only large-scale datasets covering multiple languages.

There is some existing work on cross-language entailment, but the datasets are significantly smaller in both size and language coverage. CLTE-2013 (Negri et al., 2013) is an older dataset covering English paired with Spanish, Italian, French, and German. The dataset consists of 1500 pairs for each. Khanuja et al. (2020) introduce a dataset for code-switched English and Hindi. They collect 400 premises and 2240 hypotheses where both are in code-mixed Hindi-English. Our datasets are over 100 times larger than existing alternatives and cover a significantly wider variety of languages.

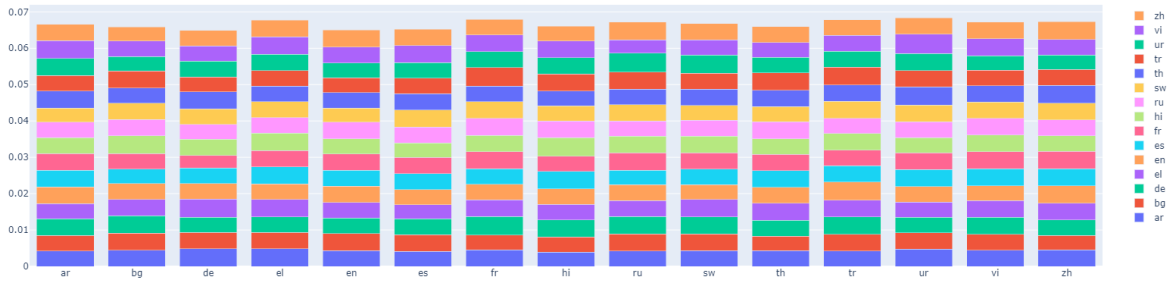
3 Datasets

To understand the extent of the cross-language capabilities of current models, we perform baseline experiments on mBERT and XLM-RoBERTa (Devlin et al., 2019). See Table 3 for a summary of all results. We use Google Translate (Wu et al., 2016) for our machine translation, accessed through translate-shell [1], a command line tool that allowed for the translation to be done directly in a Python script. The translations were done between April and July 2023. Pairs for our datasets are generated by following a simple procedure: for each entry in MNLI, two languages are chosen at random. The premise is translated to the first, and the hypothesis the second. The original English sentences and pairID are kept to maintain easy interoperability with MNLI. The 93 languages of CLI-93 cover a diverse range of language families and typologies, and include numerous low-resource languages, such as Sundanese and Pashto. These languages are also covered by large multilingual models, including full coverage in XLM-R (Conneau et al., 2020). MNLI consists of three splits: train, validation_matched, and validation_mismatched. The difference between the two validation sets is that mismatched draws from genres of text not included in the training data. We translate all three of these for our CLI datasets.

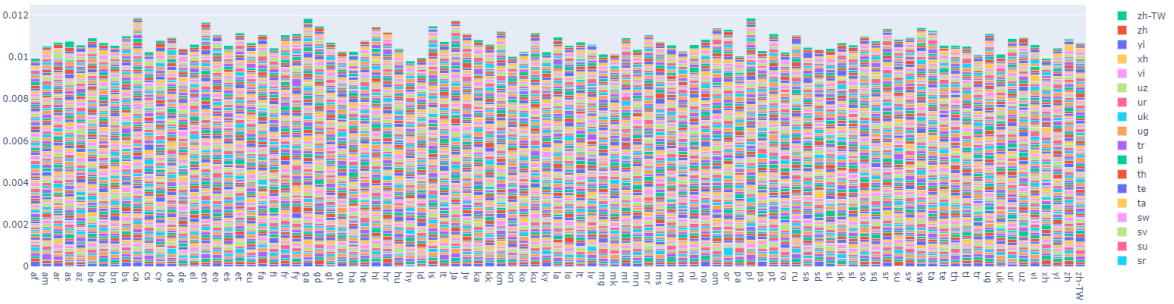
3.1 Translation Quality

To evaluate the quality of our translation pipeline, we use the test set of XNLI which is human-translated. We translate the premise and hypothesis of the English test set (5010 pairs) using our script and calculate the BLEU score (Papineni et al.,

[1] translate-shell can be viewed here



(a) Distribution of language pairs in CLI-15 train. Both individual languages and pairs show a roughly balanced distribution. The random selection of pairs introduces some variance but there is no drastic under-representation of language pairs.



(b) Distribution of language pairs in CLI-93 train. The gap between the best and worst-represented languages are about 0.2%, and there is again no egregious over or under-representations of language pairs.

Figure 1: Pairwise distribution of languages in the CLI-15 (a) and CLI-93 (b) training sets

2002a) using the dataset's translations as the reference. Table 1 shows the results for all 14 non-English languages. Only two scores were below 30: Urdu and Thai. Thai is a notably challenging language for translation systems due to issues with segmentation and tokenization (Lyons, 2020). Urdu is a low-resource language and the score we see is in line with other published results on the language (Tiedemann and de Gibert, 2023).

3.2 Human Annotation

As a second check on both translation quality and the preservation of entailment relations, we perform a human annotation. We recruited 10 annotators for 5 languages, with one annotator working on two languages. The annotators were university students who indicated that they were proficient in reading and understanding both English and one of the CLI-93 languages. Requests for annotators were posted through university channels. The participants were informed that they were annotating data for a new multilingual entailment dataset. The annotators were compensated with 20$ or an item of equivalent value. Each annotator was given a short description of the task and 100 randomly selected examples to label. The examples were drawn from the CLI-93 training set, where either the premise or hypothesis was in the annotator's second language, and the other was replaced by the English MNLI entry. The labeling set was split 50-50 on English being in the premise or the hypothesis and randomly shuffled. Each annotator for the same language labeled the same examples, although this was not guaranteed to be the case for different languages. Annotation was done remotely and upon submitting their work, annotators were asked for comments on the text they read, and if they used any translation software or Large Language Models for their labeling. Three annotators completed less than 75% of their assigned data, and two amongst them indicated that they used a language model. We remove these annotators from the results presented in Table 2. Overall, the annotators recovered the correct label for 82.37% of

fr	es	de	el	bg	ru	tr
49.78	56.28	42.21	51.30	48.22	31.15	30.37
ar	vi	th	zh	hi	sw	ur
33.98	51.47	13.8	47.73	33.43	32.63	23.98

Table 1: using our MT pipeline on the human translations in XNLI (Conneau et al., 2018) calculated using SacreBLEU (Papineni et al., 2002b; Post, 2018). Default settings are used for all languages except zh, which uses the Chinese tokenizer. Scores are above 30 for all languages except Thai, which has issues with tokenization, and Urdu, a low-resource language that MT systems generally struggle with.

the examples. This is in line with Conneau et al. (2018), which had two bilingual annotators label 100 examples each in English and French and saw that the English label was recovered 85% of the time and the French was for 83% of the examples. Annotator comments on the text ranged from saying it was "understandable, but not great", to "generally good". The coverage of languages was unfortunately biased by the availability of bilingual speakers, but the results suggest that the translations mostly maintained the same entailment relations as the original English.

3.3 Data Distribution

The language pairs in our datasets were chosen randomly, so the distribution of pairs should be roughly even. To verify that this expectation held in the actual data and no language was over or underrepresented, we perform an investigation into the distribution of languages in both CLI-15 and 93. Figure 1 shows the distribution of language pairs in the training splits of CLI-15 and CLI-93. See section A.2 in the appendix for similar charts on the two validation splits, and details on individual language distribution in the premises and hypotheses. The CLI-15 training set is mostly balanced, as expected, with each individual language covering 6-7% of the data and pairs being roughly evenly distributed. The matched and mismatched validation sets are a little noisier, likely due to their smaller size, but there are no glaring discrepancies. The mismatched set is more evenly distributed than validation_matched. The distribution patterns of CLI-93 are similar, although the massive increase in language pairs means that every pair makes up a significantly smaller portion of the overall dataset. In the train split, individual languages make up between 1-1.2% of the overall data, with some variation in the pairs. Again, the two validation sets are noisier, even more so than with CLI-15 due to the increase in potential pairs. Individual languages make up between 0.8-1.4% of the data. Overall, the datasets are not perfectly balanced, but there are no immediately concerning skews either.

4 Baseline Experiments

4.1 mBERT

Multilingual-BERT (mBERT) is the multilingual version of BERT which was trained on text from 104 languages (Devlin et al., 2019). We fine-tune mBERT-base (110M parameters) on CLI-15 and CLI-93. We use a learning rate of 2e-5 for these experiments and train for 5 epochs with a batch size of 32. These experiments are intended as a general baseline, and a hyperparameter sweep may yield better results. The test set of MNLI is private, so we use a 70-30 dev-test split from the validation_matched set. We also use the validation_mismatched set, which draws sentences from different sources than the training data, as a second, larger test set. The languages in mBERT do not fully overlap with those in CLI-93, so for the CLI-93 experiment, we remove any examples containing one of the 19 languages not covered by the model. This results in a loss of 16,504 examples from the training set, 270 from validation, 97 from test, and 404 from validation_mismatched. mBERT achieves 69.6% accuracy on test and 70.2% on validation_mismatched for CLI-15. On CLI-93, performance drops to 66.9% on test and 68.1% on validation_mismatched.

4.2 XLM-RoBERTa

XLM-RoBERTa (XLM-R) is a state of the art multilingual masked language model that covers over 100 languages (Conneau et al., 2020). XLM-R has been trained on a large amount of web data across

Language	# of Annotators	Agreement with CLI-93
es	1	76
hi	3	83.83
te	2	80
zh	1	83.5
gu	1	88
overall	8	82.37

Table 2: Annotation statistics. We show the languages annotated, the number of annotators, and the percent of the time they chose the same label as our dataset. Overall, annotators successfully recovered the label $\sim 82\%$ of the time, which is comparable to a similar analysis in Conneau et al. (2018)

Model	Trained On	CLI-15 Matched	CLI-15 Mismatched	CLI-93 Matched	CLI-93 Mismatched
mBERT	CLI-15	69.6	70.2	-	-
mBERT	CLI-93	-	-	66.9	68.1
XLM-R	multilingual-mnli	67.6	68.1	-	-
XLM-R	CLI-15	76.2	77.1	-	-
XLM-R	CLI-93	-	-	69.4	70.7

Table 3: We report model accuracy on matched and mismatched sets after fine-tuning XLM-R and mBERT on our datasets. Due to the MNLI test set being private, our evaluation sets are a 30% split of validation_matched (2945 examples) and the full validation_mismatched (9816). The matched version of the dataset contains examples from the same source as the training set and mismatched comes from different sources. The Trained On column tells which dataset the model was fine-tuned on. multilingual-mnli is a dataset with the examples in MNLI translated into monolingual pairs in 15 languages. We find that XLM-R outperforms mBERT, especially on CLI-15 and both models see a drop in performance on CLI-93. We also find that multilingual-mnli does not perform as well as training on cross-language data.

different languages. For our next experiment, we fine-tune XLM-R-base (270M parameters) on CLI-15. The learning rate used is 5e-6 and all other hyperparameters are unchanged. The resulting model outperforms mBERT with an accuracy of 76.2% on test, and 77.1% on validation_mismatched. Running the same procedure on CLI-93 dropped performance drastically to 69.6% and 70.7%, respectively. The gap between these results and those of mBERT is much smaller than it was for CLI-15. The model is highly sensitive to hyperparameters, as using a higher learning rate of 2e-5, broke training entirely and produced a 33% model that could only predict entailment.

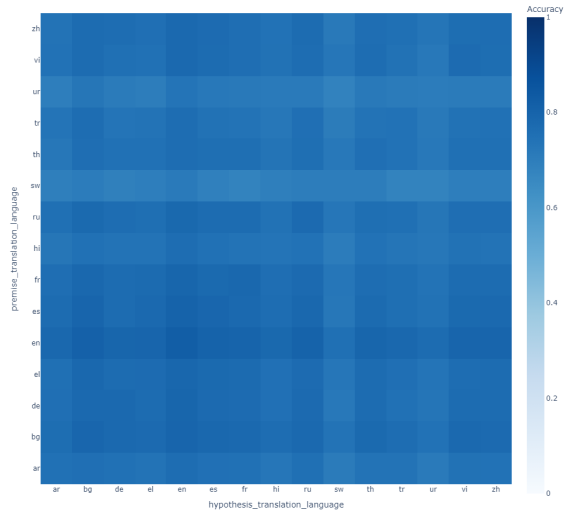
4.2.1 multilingual-mnli

In order to test whether cross-language capabilities can be acquired through monolingual training in multiple languages, we fine-tune XLM-R on data in this setting, which we refer to as multi-mono. We prepare a dataset which consists of MNLI pairs in the CLI-15 languages with premise and hypothesis in the same language. We use HuggingFace XNLI[2] , which provides MNLI examples machine translated to these languages, and partition the $\sim 393k$ pairs of the data into 15 parts, with the examples in each being in the same language. The data is then shuffled. This creates a balanced training set for the 15 languages that is the same size as our cross-language data. The resulting XLM-R model significantly underperforms the CLI-15 one, only reaching 67.6% on test and 68.1% on validation_mismatched.

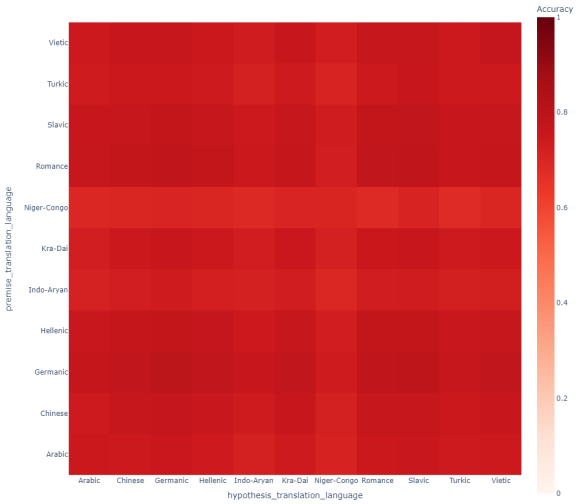
5 Results and analysis

Our experiments find that XLM-R performs worse in a cross-language setting compared to a monolingual one. On CLI-15, where the languages are the same as XNLI, the model is 3.4 points worse than the reported XNLI average performance (Conneau et al., 2020) for XLM-R, dropping further when the number of languages increases. mBERT, which is

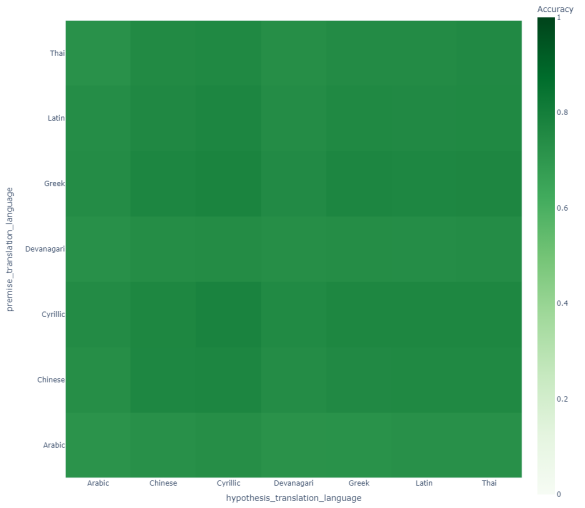
[2]https://huggingface.co/datasets/xnli



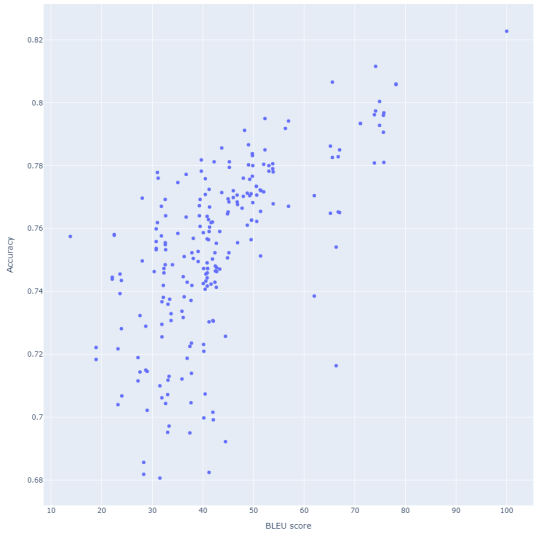
(a) A heatmap of the accuracy of different language pairs. The low-resource Swahili and Urdu consistently underperform. We also see symmetry showing that position in premise or hypothesis doesn't have an influence on performance



(b) A heatmap of the accuracy of different language family pairs. Many have only one representative language so the inferences we can draw are limited, but Niger-Congo and Indo-Aryan are the weakest-performing.



(c) A heatmap of the accuracy of different language script pairs. The weakest scripts are Devanagari (Hindi), and Arabic (Urdu, Arabic).



(d) A scatterplot of average BLEU score vs accuracy on language pairs. We can see a loose increase in accuracy as the BLEU score increases but the highest pairs are all pairs with English (BLEU 100). Looking at the other languages, there doesn't seem to be a correlation

Figure 2: Breakdown of the gradual fine-tuned CLI-15 model by language, family, script, and translation quality.

typically worse in multilingual tasks also underperforms compared to XLM-R in our cross-language one. The consistent drop in CLI-93 across models may also be due to a "curse of multilinguality" effect (Conneau et al., 2020), where adding more languages lowers performance on the highest resource ones. mBERT does see a much smaller drop going from CLI-15 to CLI-93, but its score may be slightly inflated as the removed languages are likely to be lower-resource, so the model benefits from an easier evaluation set. Additionally, we find that training with monolingual pairs in multiple languages is not enough to learn cross-language inference; cross-language data is needed. The XLM-R model trained on multi-mono data not only underperformed a cross-language XLM-R,

but also scored lower than mBERT. Models only trained in multi-mono settings may not be ideal for downstream cross-language tasks.

5.1 Analysis on Gold XNLI test set

To further investigate the effect of CLI-93 and CLI-15 on different languages, families and scripts, we created permutations between all language pairs, including monolingual pairs, in the XNLI test set. Our final test set had 5010 sentence pairs for each of the permutations of the CLI-15 languages for a total of 1127250 pairs. To group the languages by family and script we use the classifications of Fan et al. (2020). This section describes results on the CLI-15 XLM-R model but the broad trends were similar for CLI-93 and figures can be seen in the appendix 7.

Position of Language: To see if there is any performance difference for a language based on if it is in the premise or the hypothesis, we examine differences in score by position. We saw some small variance in the scores of languages being in the premise vs. being in the hypothesis but the effect was very small with an average difference of about 0.803% in scores for when a language is in the premise vs when it is in the hypothesis.

Language Script Pairs: Figure 2c shows the scores of different pairs of scripts. Arabic-Arabic was the worst performing pair. On the other end, we see pairs with Greek and Cyrillic performing well. Latin does not stand out even though it contains English and Spanish, a few of the top performing languages, because Swahili is also Latin script, showing the steep decline in performance on low-resource languages.

Language Pairs: Figure 2a shows the scores of different language pairs. Pairs with Swahili and Urdu, the lowest resource languages in the data, are consistently among the weakest. Unsurprisingly, the strongest performing language is English. All the languages saw higher scores when paired with English. We also saw that monolingual pairs were in the top 50% of the scores. Below we show the highest and lowest performing pairs. English-English scored highest with 82.2% and Turkish-Swahili was lowest with a score of 68.1%.

Language Family Pairs: Figure 2b shows the scores of different language family pairs. We can see that Niger-Congo is the worst performing, which makes sense as Swahili, one of the weakest languages, is the only representative in this data. Unsurprisingly again, the strongest family is Germanic, whose representatives are English and German, two very high-resource languages. Unfortunately, this analysis is limited by the sparsity of languages in each family.

BLEU Score: We plot the average BLEU score for each pair against accuracy in 2d. The BLEU scores are taken from Table 1, which were done by comparing our translation pipeline to human-translated XNLI sentences. The chart seems to show a slight correlation between higher BLEU scores and higher accuracy, however most of the higher BLEU scores come from pairs with English (100 BLEU by default). When considering this, the plot separates into two sections, one with English, and one without and neither seems to show a correlation between accuracy and BLEU.

5.2 Cross Lingual Transfer

Cross-lingual transfer, where a model is evaluated on a task in a language that was not in the training data, is an important technique for low-resource NLP. To evaluate the degree of transfer possible with our datasets, we experiment with a "CLI-14", where all sentences in one held-out language are replaced with the original English. We fine-tune XLM-R models on this data and evaluate on the combined XNLI validation and test sets for the held out language. Importantly, the evaluation here is purely monolingual; we want to see if cross-language training can teach a model the task in an unseen language. These experiments were run using Spanish, Swahili, and Urdu to compare the effects of low vs high-resource and the presence of another language in the same family (French for Spanish, Hindi for Urdu, none for Swahili) in the training data. As a baseline, we also train a model on MNLI, and evaluate it on each held-out language, effectively repeating the cross-lingual transfer section of Conneau et al. (2020) on a subset of the languages.

Table 4 shows the accuracy of XLM-R trained on different variants of CLI-15 and tested on XNLI test and validation sets. CLI-14 shows impressive cross-lingual transfer, especially with Urdu (ur), where we see an increase of 4.6 points over the MNLI baseline and just 1.5 points underperformance from the CLI-15 baseline. Transfer for Spanish was good across all models, with CLI-14 getting almost the same results as MNLI, and only 1 point worse than CLI-15. Spanish is a very

experiment	xnli-ur	xnli-sw	xnli-es
CLI-14-ur	68.9	-	-
CLI-14-sw	-	65.9	-
CLI-14-es	-	-	78.0
CLI-15	70.4	70.0	79.2
MNLI	64.3	64.8	77.9

Table 4: Results for fine-tuning XLM-R on CLI-14, where one of the CLI-15 languages is replaced by English. CLI-14-ur is CLI-15 but every Urdu sentence is replaced by the original MNLI sentence in English. Accuracy is reported on the combined XNLI monolingual validation and test sets for the held-out language. CLI-15 refers to training a model on the full data, while MNLI is the traditional cross-language transfer technique of training only on English. We find that CLI-14 uses the same amount of data to achieve greater transfer than MNLI in all three languages and comes close to CLI-15 in two.

high-resource language so this performance may reflect the model's increased capabilities with it. Another factor that may benefit both languages is the presence of high-resource neighbors in Hindi and French, meaning that some training was done on a related language. Swahili, which has no language family neighbor, sees an increase of 1.1 points over MNLI but falls 4.1 points below CLI-15.

Overall, these results show that our datasets can support cross-lingual transfer. For all three languages we experiment with, training on the same amount of data with CLI-14 outperforms the baseline MNLI model, and by an especially wide margin for Urdu. Meanwhile, on Urdu and Spanish, the results come very close to training with the language included, although we don't see this for Swahili. This suggests that a competitive monolingual entailment model for a new language can be learned without explicit training examples using cross-language data when there are related languages in the training data.

6 Conclusions and Future Work

Our initial results show that CLI is a challenging task and leaves the door open for further work in many directions. There is much room for improvement in the dataset itself, and the creation of a small gold evaluation set, in the style of XNLI would be helpful to counteract concerns of error propagation with machine translation. Our annotation efforts also left many languages and families underexplored, due to the difficulty of finding bilingual annotators. Finally, there is room for improvement in monolingual NLI. One annotator commented that the English text they looked at, which was drawn directly from MNLI, was of low quality, and sometimes unintelligible. Our own analysis on MNLI, especially the mismatched set, also found that the quality of both text and labeling was lacking. Better monolingual data that better captures the nuances of natural language would propagate improvement to the cross-language space as well.

There is also room for deeper investigation on the modeling side. The two models we use for benchmarking are both encoders using masked language modeling. Other architectures, such as encoder-decoder Sequence2Sequence models, like mT5 (Xue et al., 2021), may give different results. Including cross-language tasks in the pretraining objectives of BERT-like models may also improve cross-language ability. Finally, the use of CLI as a pretraining task may help with downstream cross-language tasks, such as information retrieval or question answering.

In this work, we introduce CLI-15 and CLI-93, silver datasets that are the largest existing resources for cross-language inference, and cover 79 languages that prior multilingual entailment datasets did not include. We perform both automatic and human-guided analysis on the translations, both of which suggest that the quality of the data is mostly preserved. We run a variety of baseline experiments, showing that this is a challenging task for modern models, and that there is much room for improvement, which in turn has implications for their suitability for downstream tasks. We also show that the cross-language setting can support greater cross-lingual transfer than a purely monolingual approach does. Finally we discuss potential directions for future work on this problem, both for datasets and models.

7 Limitations

Much of our in-depth analysis is limited to the 15 languages of CLI-15, due to the availability of gold evaluation data in XNLI. There is room for more examination on 79 additional languages contained in CLI-93. Google Translate is regularly updated and our translations may not be fully replicable as a result. Additionally, while we took steps to check that the quality of the data is acceptable, it is still machine translated and may not be fully accurate. Our annotation efforts were also limited by the participants we had available, and don't fully represent the languages in the dataset. Finally our baseline models are intended as a general guideline, and do not cover the full range of architectures or optimizations possible for this task.

References

Željko Agić and Natalie Schluter. 2018. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 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, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16, Marseille, France. European Language Resources Association.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Ankit Kumar Upadhyay and Harsit Kumar Upadhya. 2023. Xnli 2.0: Improving xnli dataset and performance on cross lingual understanding (xlu). In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–6.

Seamus Lyons. 2020. A review of thai–english machine translation. *Machine Translation*, 34.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2013. Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin

Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Jörg Tiedemann and Ona de Gibert. 2023. The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327, Toronto, Canada. Association for Computational Linguistics.

Gijs Wijnholds and Michael Moortgat. 2021. SICK-NL: A dataset for Dutch natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1474–1479, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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

A Appendix

A.1 Languages Covered

Table 5 and Table 6 below list the languages covered by the CLI-15 and CLI-93 respectively. These are the languages covered by XLM-R with the exception of Breton and Romanized variations of languages already covered. The languages of CLI-15 are a subset of the CLI-93 languages which have gold XNLI annotations.

A.2 Language Analysis

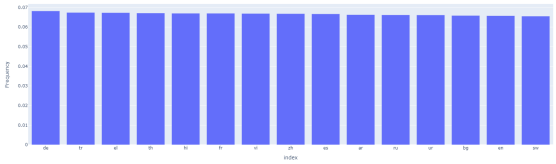
Figures 3 and 4 show the distribution of languages in the premise and hypothesis of train, validation, and validation mismatched sets. We see that the training set is balanced while the validation mismatched is not completely balanced and we see a little drop which is slightly more pronounced in the validation set. Figures 5 and 6 show pairwise distributions in the matched and mismatched validation sets

A.3 CLI-93 model Performance Analysis

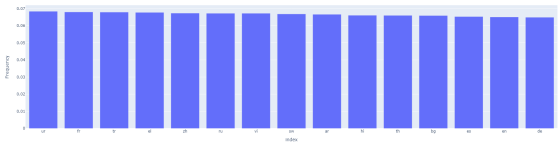
Figure 7 shows the performance of the CLI-93 model on different language, family and script pairs. Overall, the trends mostly resemble those of the equivalent CLI-15 results in the main paper.

Language	Code
Arabic	ar
Thai	th
Urdu	ur
Chinese (Simplified)	zh
French	fr
Spanish	es
German	de
Greek	el
Bulgarian	bg
Russian	ru
Turkish	tr
Vietnamese	vi
Hindi	hi
Swahili	sw

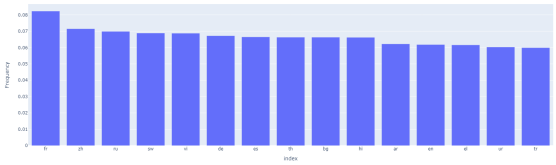
Table 5: Language codes and language for the CLI-15 datasets



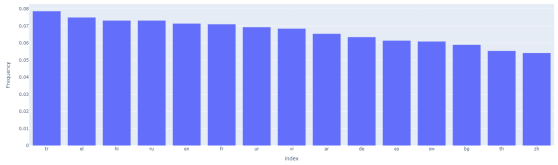
(a) CLI-15 train hypothesis distributions



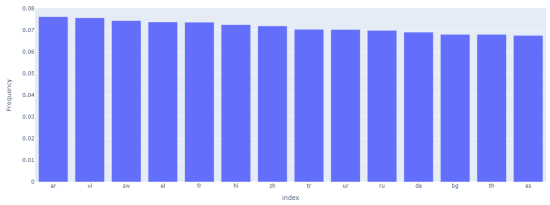
(b) CLI-15 train premise distributions



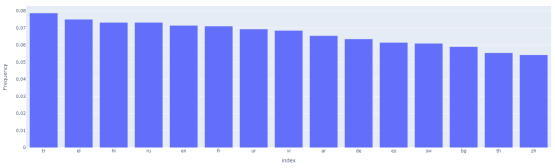
(c) CLI-15 validation hypothesis distributions



(d) CLI-15 validation premise distributions



(e) CLI-15 validation mismatched hypothesis distributions

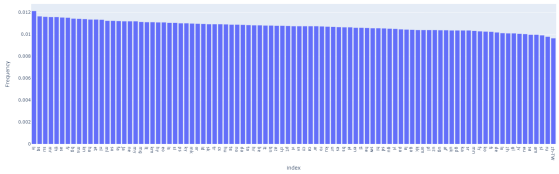


(f) CLI-15 validation mismatched premise distributions

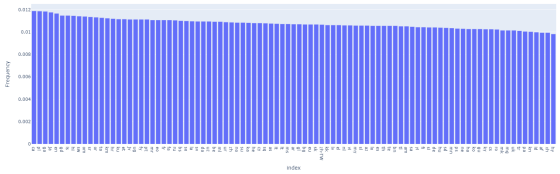
Figure 3: Distribution of languages in CLI-15

Language	Code	Language	Code
Afrikaans	af	Kurdish (Kurmanji)	ku
Albanian	sq	Kyrgyz	ky
Amharic	am	Lao	lo
Arabic	ar	Latin	la
Armenian	hy	Latvian	lv
Assamese	as	Lithuanian	lt
Azerbaijani	az	Macedonian	mk
Basque	eu	Malagasy	mg
Belarusian	be	Malay	ms
Bengali	bn	Malayalam	ml
Bosnian	bs	Marathi	mr
Bulgarian	bg	Mongolian	mn
Burmese	my	Nepali	ne
Catalan	ca	Norwegian	no
Chinese (Simplified)	zh	Oriya	or
Chinese (Traditional)	zh-TW	Oromo	om
Croatian	hr	Pashto	ps
Czech	cs	Persian	fa
Danish	da	Polish	pl
Dutch	nl	Portuguese	pt
English	en	Punjabi	pa
Esperanto	eo	Romanian	ro
Estonian	et	Russian	ru
Filipino	tl	Sanskrit	sa
Finnish	fi	Scottish Gaelic	gd
French	fr	Serbian	sr
Galician	gl	Sindhi	sd
Georgian	ka	Sinhala	si
German	de	Slovak	sk
Greek	el	Slovenian	sl
Gujarati	gu	Somali	so
Hausa	ha	Spanish	es
Hebrew	he	Sundanese	su
Hindi	hi	Swahili	sw
Hungarian	hu	Swedish	sv
Icelandic	is	Tamil	ta
Indonesian	id	Telugu	te
Irish	ga	Thai	th
Italian	it	Turkish	tr
Japanese	ja	Ukrainian	uk
Javanese	jv	Urdu	ur
Kannada	kn	Uyghur	ug
Kazakh	kk	Uzbek	uz
Khmer	km	Vietnamese	vi
Korean	ko	Welsh	cy
Western Frisian	fy	Xhosa	xh
Yiddish	yi		

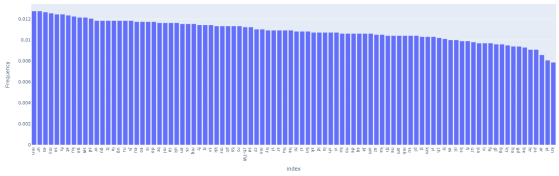
Table 6: Language codes and language for the CLI-93 datasets



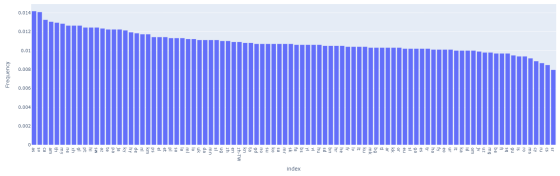
(a) CLI-93 train hypothesis distributions



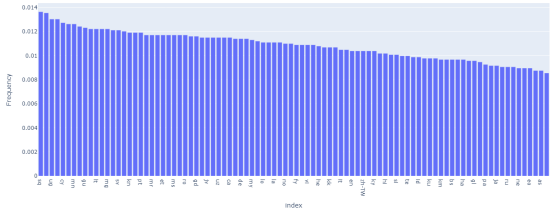
(b) CLI-93 train premise distributions



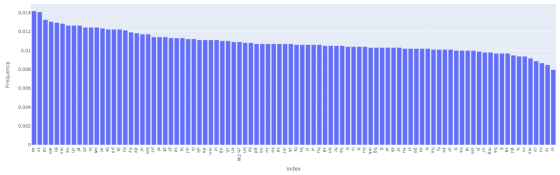
(c) CLI-93 validation hypothesis distributions



(d) CLI-93 validation premise distributions

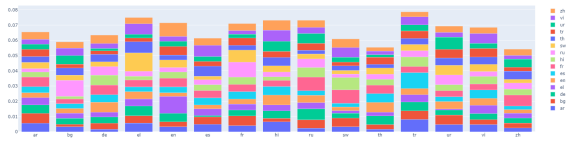


(e) CLI-93 validation mismatched hypothesis distributions

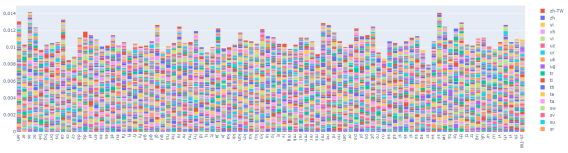


(f) CLI-93 validation mismatched premise distributions

Figure 4: Distribution of languages in CLI-93

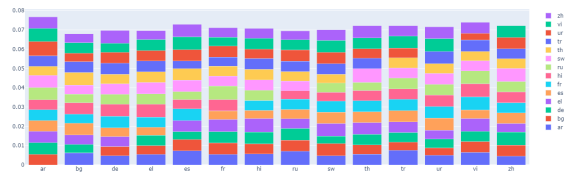


(a) CLI-15 validation pairwise distributions

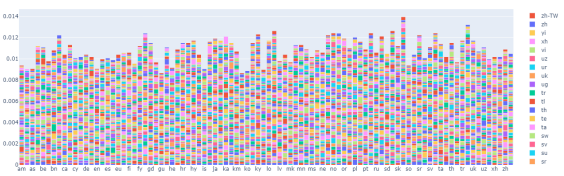


(b) CLI-93 validation pairwise distributions

Figure 5: Pairwise distribution of languages in the CLI-15 and CLI-93 validation set

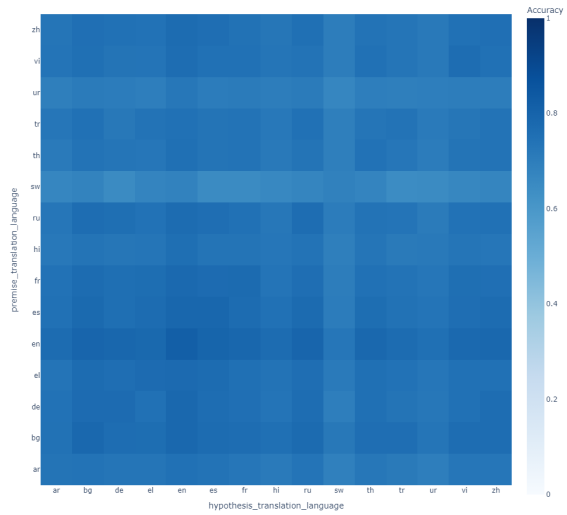


(a) CLI-15 validation mismatched pairwise distributions

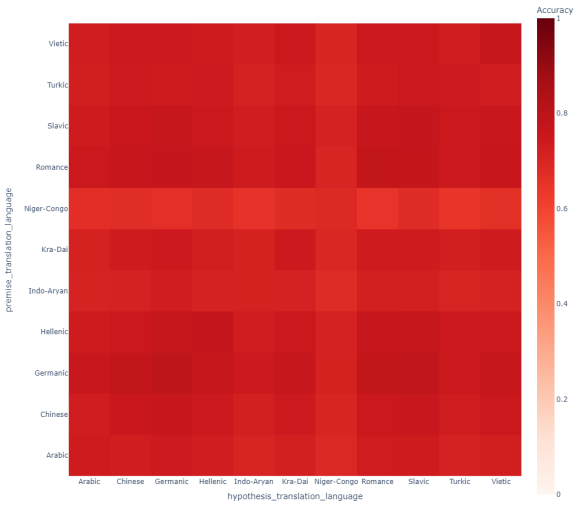


(b) CLI-93 validation mismatched pairwise distributions

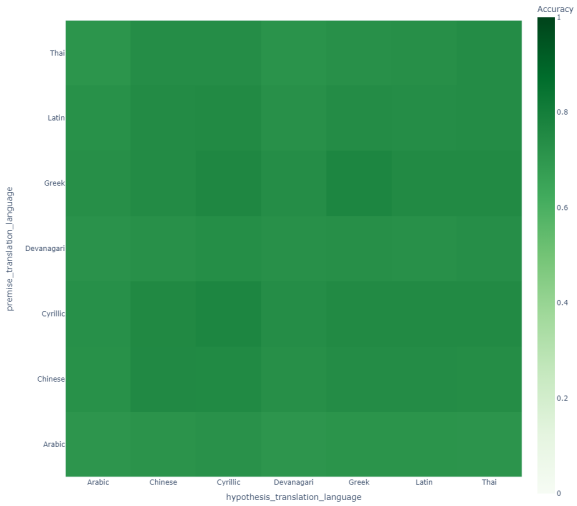
Figure 6: Pairwise distribution of languages in the CLI-15 and CLI-93 mismatched validation set



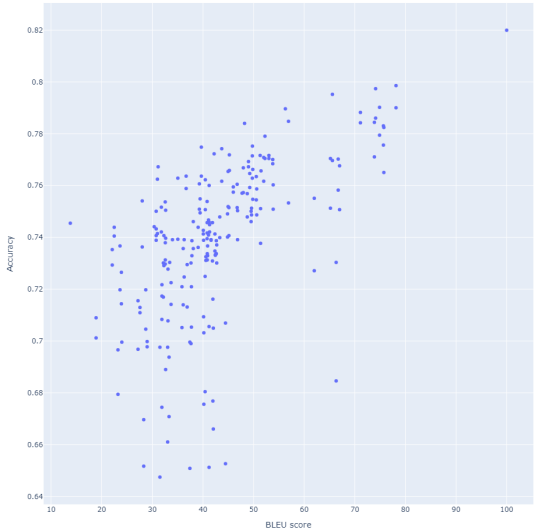
(a) Scores grouped by languages



(b) Scores grouped by language family



(c) Scores grouped by script



(d) Scatter plot of average BLEU score vs accuracy on language pairs

Figure 7: Heatmaps of the average scores of the CLI-93 XLM-R model on the XNLI test set