C-XNLI: Croatian Extension of XNLI Dataset

Leo Obadić, Andrej Jertec, Marko Rajnović, Branimir Dropuljić
RealNetworks, Inc.
{lobadic,anjertec,mrajnovic,bdropuljic}@realnetworks.com

Abstract

Comprehensive multilingual evaluations have been encouraged by emerging cross-lingual benchmarks and constrained by existing parallel datasets. To partially mitigate this limitation, we extended the Cross-lingual Natural Language Inference (XNLI) corpus with Croatian. The development and test sets were translated by a professional translator, and we show that Croatian is consistent with other XNLI dubs. The train set is translated using Facebook's 1.2B parameter `m2m_100` model. We thoroughly analyze the Croatian train set and compare its quality with the existing machine-translated German set. The comparison is based on 2000 manually scored sentences per language using a variant of the Direct Assessment (DA) score commonly used at the Conference on Machine Translation (WMT). Our findings reveal that a less-resourced language like Croatian is still lacking in translation quality of longer sentences compared to German. However, both sets have a substantial amount of poor quality translations, which should be considered in translation-based training or evaluation setups.

1 Introduction

Natural language processing has developed rapidly in recent years. Models are starting to achieve human-like performance, but most of these achievements are concentrated on only a small fraction of the world's 7000+ languages. This is to be expected due to the nature of linguistic annotation, which is not only tedious, subjective, and costly, but also requires domain experts, which are in decline (Lauscher et al., 2020).

There are two main approaches commonly used to handle that problem from the models' perspective. The first approach relies on cross-lingual transfer, where the model is pretrained to learn multilingual representations (Conneau et al., 2020; Pires et al., 2019), while the other approach relies heavily on Machine Translation (MT) systems to translate the text from a low-resource language to a high-resource language (or vice versa). Both approaches can be easily evaluated on cross-lingual benchmarks such as XTREME (Hu et al., 2020) or XGLUE (Liang et al., 2020). They consist of cross-lingual datasets grouped by task to allow comprehensive evaluation. Unfortunately, XTREME covers 40 languages and XGLUE only 19.

Since none of these benchmarks include Croatian language in any of their datasets, and Cross-lingual Natural Language Inference (XNLI; Conneau et al., 2018) corpus is included in both, we decided to extend XNLI with Croatian (C-XNLI). The task is to classify whether a premise contradicts, entails, or is neutral to the hypothesis. XNLI's development and test sets are crowd-sourced in English and human-translated into 14 languages, while MultiNLI's (Williams et al., 2018) training set is used for training. It also consists of machine-translated sets required for the translate-train and translate-test paradigms.

Our Croatian extension is created in the same manner as its XNLI parent. The development and test sets are translated by a professional translator. Since XNLI provides translate-train, translate-dev and translate-test sets, we opted for Facebook's 1.2B parameter `m2m_100` MT model (Fan et al., 2020) to create our own translations.

It has been shown that MT models still suffer from errors like mistranslations, non-translations and hallucinations (Freitag et al., 2021; Raunak et al., 2021), which motivated us to analyze the quality of our dataset. For this purpose, we sampled 2000 sentences per language in both Croatian and German, and evaluated the translations using a variant of the Direct Assessment (DA) score proposed in the Multilingual Quality Estimation dataset (MLQE; Fomicheva et al., 2022).

To summarize, our contributions are the following: (1) we create and analyze the Croatian exten-

Findings of the Association for Computational Linguistics: ACL 2023, pages 2258–2267
July 9-14, 2023

sion of XNLI and provide baseline models, (2) we create Quality Estimation (QE) datasets for Croatian and German to evaluate the quality of machine-translated sentences from the translate-train sets, and (3) we quantify the textual overlap between hypothesis and premise and analyze its impact on baseline models.

2 Datasets

2.1 C-XNLI

In creating the dataset, we follow the same procedure as Conneau et al. (2018). We hired a native Croatian professional translator to translate the English development (2490 samples) and test (5010 samples) sets of the XNLI dataset into Croatian. Premises and hypotheses were given to the translator separately to ensure that the premises did not provide context for the hypotheses. The English training set, derived from MultiNLI and containing 392,702 samples, was translated into Croatian using a selected MT model. We considered a total of eight models and opted for Facebook's multilingual `m2m_100` model with 1.2B parameters because of its highest BLEU score (Papineni et al., 2002) on the FLORES dataset (Guzmán et al., 2019), as shown in Table 1. All of `m2m_100` and `mbart` models are available on fairseq[1] (Ott et al., 2019), whereas `opus` models are available on Helsinki-NLP[2] (Tiedemann, 2020; Tiedemann and Thottingal, 2020) and are evaluated by Marian-NMT (Junczys-Dowmunt et al., 2018).

model name	BLEU
m2m_100_1.2B	27.81
opus_sla	25.73
opus_hr	25.64
m2m_100_615M	23.74
mbart50_en2m	23.72
m2m_100_418M	22.95
mbart50_m2m	22.66
m2m_100_175M	15.67

Table 1: Translation scores on Croatian part of FLORES devtest set for each model.

2.2 DA Scores

To evaluate the quality of the system used to translate English to Croatian, we compare the generated translations with the available translations from a high-resource language. We score a sample of Croatian and German translations from the train set and compare the results. The sentences were sampled using a semantic similarity-based metric that correlates with translation quality (Cer et al., 2017) to flatten the original distribution of scores and analyze samples of diverse quality. A cosine score between the multilingual sentence representations from both LASER (Artetxe and Schwenk, 2019) and SBERT (Reimers and Gurevych, 2019) were used to measure semantic similarity between the source and translated sentences. These models are commonly used at the Conference on Machine Translation (WMT) for QE task (Specia et al., 2021, 2020). The SBERT we used is a multilingual variant trained on the paraphrase dataset which has slightly better performance than the models trained on similarity tasks (Reimers and Gurevych, 2020).

By utilizing a histogram of cosine scores with a bin size of 0.05, we adopted a circular sampling approach to randomly select one premise from each bin until a total of 50 premises were obtained. Similarly, we followed the same procedure for hypotheses, alternating between SBERT and LASER cosine scores. Furthermore, we implemented an additional criterion to ensure the inclusion of all premises and hypotheses that share a common premise. This entire process was repeated until we reached a 1000 samples each, for both SBERT and LASER cosine scores (2000 in total).

We scored the samples using the procedure described by Fomicheva et al. (2022). Annotators were asked to rate the translation quality for each sentence on a scale 0–100. Sentences were initially annotated by three annotators. If the range of the most diverging scores exceeded 30 points, an additional annotator was asked to replace the most diverging one until convergence was achieved. The annotators' raw scores were converted to z-scores[3]; the final score is the average of all scores after convergence. More information about annotators, and annotation procedure is presented in Appendix A.

3 Analyses and Results

3.1 C-XNLI and DA Scores

To demonstrate that our extension has similar properties to its parent XNLI, we perform the following analyses. We tokenize C-XNLI's sentences with MOSES tokenizer and obtain the average number

[1] https://github.com/facebookresearch/fairseq
[2] https://github.com/Helsinki-NLP
[3] The normalization according to each individual annotator's overall mean and standard deviation.

	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	hr
XX-En BLEU	35.2	38.7	39.3	42.1	45.8	41.2	27.3	27.1	21.3	22.6	29.9	24.4	23.6	24.6	41.8
En-XX BLEU	15.8	34.2	38.8	42.4	48.5	49.3	37.5	24.9	24.6	21.4	21.9	24.1	39.9	23.2	42.1

Table 2: BLEU scores calculated on XNLI test set reported by Conneau et al. (2018), extended with Croatian using MOSES tokenizer. XX-En stands for any language to English, whereas En-XX stands for English to any language translation.

of tokens in premises (19.0) which is nearly double the number in hypotheses (9.3) – a ratio that is consistent with other XNLI languages (see Appendix C).

Another analysis Conneau et al. (2018) provide is the BLEU score of their MT systems translating to and from the target language. We have extended their results to include those for the Croatian language (Table 2). Our translations from English to Croatian (EN-XX in the table) have the fourth-best BLEU score. These findings are not too surprising since the MT we use is more recent. The distribution of DA scores for Croatian and German is shown in Figure 1. We can observe that Croatian, although is a lower-resourced language, it has a slightly higher translation quality, as the mean of Croatian DA scores is almost identical to a German one.

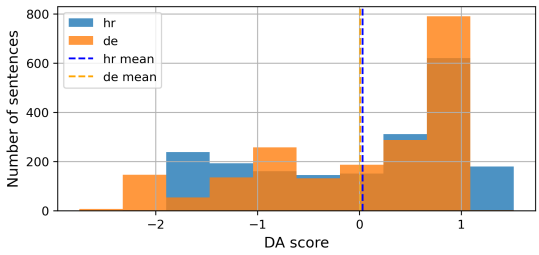

Figure 1: Distributions of DA scores across languages.

The correlations between the LASER and SBERT cosine scores and DA scores for both languages are shown in Table 3, with $p < 0.05$. The correlations for German are higher, and the LASER cosines tend to correlate less.

	hr	de
SBERT	0.57	0.61
LASER	0.45	0.54

Table 3: Spearman correlation calculated between cosine score and DA annotations.

In Figure 2 we can see that the Croatian model is more likely to make a mistake on premises compared to the German model.

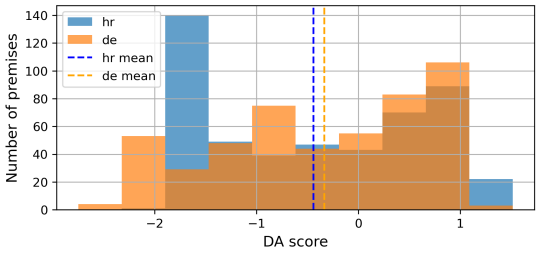

Figure 2: Premise distributions of hr/de DA scores.

3.2 Overlaps

The analysis presented here extends Artetxe et al.'s (2020) work where authors demonstrate that the overlap between hypotheses and premises is an overlooked bias in the XNLI dataset, caused by access to premise during hypothesis generation in English, and no access to it during translation into other languages. They decrease the bias by back-translating data and improve their results. To demonstrate the existence of that bias, we take a more direct approach and define a metric that represents overlap – the proportion of copied text from premise to hypothesis. It is the number of character N-grams which occur in both hypothesis and premise, divided by the number of possible character N-grams in the hypothesis. In Table 4 we presented those overlaps using bi-grams, $N = 2$. We can observe that in the training set, the overlap is 5% to 20% higher compared to development and test sets. In order to investigate that even further, we asked our professional translator to translate 1% of our C-XNLI dataset: 100 sentences which consist of 25 premises and 75 of their hypotheses. We made sure that the premise was given alongside each hypothesis so that it provides context to it in order to measure the influence on the overlap since, in the translation effort, premises and hypotheses were given separately. Our representative sample contained similar genre distribution, overlap distribution, and similar development vs. test overlap ratio. Our results show that when using $N = 2$, biased sample has 8% increase in overlap, whereas for $N = \{3, 4, 5\}$, it increased by $\sim 17\%$.

split	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	hr
dev	0.43	0.50	0.56	0.47	0.57	0.55	0.54	0.40	0.45	0.52	0.41	0.48	0.38	0.50	0.21	0.54
test	0.44	0.51	0.56	0.48	0.58	0.56	0.55	0.40	0.46	0.52	0.41	0.49	0.39	0.49	0.21	0.53
train	0.54	0.60	0.62	0.56	0.62	0.62	0.61	0.52	0.55	0.54	0.52	0.54	0.36	0.56	0.48	0.56

Table 4: Average overlap between hypotheses and premises for each language and split.

Paradigm	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	hr	Avg	Avg^{+hr}	C_{tr}	C_{de}	C_{te}
Fine-tune multilingual model on English training set (ZERO-SHOT)																					
en	72.0	77.7	76.4	75.8	84.5	78.8	78.0	69.7	75.5	65.4	71.6	72.6	65.5	74.3	72.9	78.0	74.0	74.3	0.84	0.73	0.74
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																					
Conneau et al. (2020)	77.3	81.3	80.3	80.4	85.4	82.2	81.4	76.1	79.7	73.1	77.9	78.6	73.0	79.7	80.2	-	79.1	-	0.82	0.65	0.67
all	77.3	80.8	80.5	79.8	84.8	81.7	80.7	75.5	79.0	72.6	77.4	77.8	72.0	79.0	78.8	80.6	78.5	78.6	0.87	0.75	0.76
all_plus_hr	77.2	81.1	80.3	79.9	84.8	81.9	80.9	75.4	78.6	72.2	77.2	77.7	71.1	79.3	78.8	81.0	78.4	78.6	0.85	0.73	0.75
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																					
	74.0	79.8	79.3	77.0	84.5	80.8	79.0	72.9	77.8	69.7	67.1	75.2	66.9	78.4	77.6	80.2	76.0	76.3	0.83	0.76	0.77
Translate everything to English and use English-only model (TRANSLATE-TEST)																					
en	72.8	77.7	76.6	76.4	84.5	78.9	77.6	67.7	73.4	63.4	68.6	72.3	63.0	70.7	72.8	79.5	73.1	73.5	0.79	0.70	0.71

Table 5: We present the accuracy of baseline XLM-R Base models on each XNLI language, with the addition of Croatian, together with an average accuracy for all languages without Croatian (Avg) and with Croatian (Avg^{+hr}). Our XLM-R models are averaged over three different seeds. We also calculate the Spearman's correlation between accuracies of each model's setup and train set overlaps (C_{tr}), development set overlaps (C_{de}), and test set overlaps (C_{te}). For overlaps we used $N = 2$.

3.3 XLM-R Setups

We tested cross-lingual transfer using zero-shot and translate-based setups. For each, we employ pre-trained XLM-R Base model (Conneau et al., 2020), implemented in Transformers library (Wolf et al., 2020). In the zero-shot approach, we fine-tune our model on English samples. In the translate-train approach, we fine-tune on translations of a training set, whereas in translate-train-all, we fine-tune it on concatenated training translations. Evaluations are done in all languages. In the translate-test approach, we use the same model from our zero-shot approach and evaluate it on English translations of other languages. We experimented with various hyperparameter configurations and found appropriate ranges. Hyperparameter optimization is done for each setup, and details are presented in the Appendix B.

Results of baseline setups are shown in Table 5. To demonstrate the comparability of our training setup, we compare XLM-R's reported accuracy with ours, which is only 0.6 points lower in the train-translate-all setup. The performance of the Croatian model is consistently among the TOP5 models. The reason for that might be in the high BLEU score shown in Table 2. Focusing on the best overall model – translate-train-all, we notice that adding Croatian did not drastically change the average performance and decreased it only for distant languages like Urdu and Swahili. Whereas for other languages, it increased or did not change significantly.

Finally, Table 5 also shows how the performance of models on the test set of each language correlates with the bi-gram overlaps in the train, development, and test sets of that particular language. There is a consistent high correlation between the overlap in all sets and models' performance ($p < 0.05$). However, a lower correlation is seen in the development and test sets. This observation could be attributed to the fact that increasing the overlap of a particular language makes it more similar to the English set, in terms of overlap, thus improving the performance. However, as we showed in Subsection 3.2, the overlap in the development and test sets is artificially lower due to biased translation. Alternatively, high training overlaps might indicate that the model is learning to detect the occurrence of overlapping cues.

4 Conclusion

In this work, we extended XNLI to include the Croatian language. The development and test sets were translated by a professional translator. We have successfully demonstrated that the quality of the development and test sets is comparable to that of the other languages. To validate the machine-translated training set, we compare our

Croatian translations with those available for a high-resourced language – German. The comparison is based on 2000 manually scored sentences from German and Croatian train sets using a variant of DA scores normalized by z-score. Our results show that the Croatian MT model performs slightly better because it's more up-to-date, even though it's a lower-resourced language. We also found that the Croatian translation model performs poorly on longer sentences – premises.

Finally, we present an overlap metric to measure the textual overlap between the premise and hypothesis. We find that the training set has larger overlaps than the development and test sets. These overlaps resulted in a high correlation between the models' scores, indicating that a model uses cues from the data that also correlate with overlaps.

We provide our datasets under the same license[4] as the XNLI dataset, and also make the accompanying code available on GitHub[5]. We hope that by sharing our datasets, researchers will have the opportunity to gain further insights and expand their knowledge in the field of cross-lingual transfer.

Limitations

In each contribution of this work, we can isolate several potential limitations. In creating C-XNLI, the MT model for the formation of the train set was chosen based on the results from a single dataset. Additionally, an assumption that the model is plagued with typical issues that affect MT models was investigated on a small dataset. Although we are skeptical of the MT model's performance and perform QE scoring of the small dataset by a group of annotators and analysis to ascertain its performance, we are only comparing Croatian machine-translation results to results from a single language (German), assuming that results would hold for other high-resource languages. Also, for some MT evaluations, we use a single metric (BLEU) known to have many problems but only generally considered to correlate with human judgment.

Our hyperparameter optimizations are of limited scope. All hyperparameters are fixed, except the learning rate with four possible values we search over. Furthermore, we only used three seeds. We could not perfectly reproduce the results outlined in the paper of our baseline XLM-R Base model, partly due to a lack of elucidation in the original paper and partly due to limited hyperparameter optimizations.

Finally, we do not elucidate further and experiment with the discovered correlation between the models' performance and the overlap in datasets, and we leave it for future work.

[4]CC BY-NC 4.0

[5]https://github.com/lobadic/C-XNLI

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. 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.

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.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources*

and Evaluation Conference, pages 4963–4974, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. 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 datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Direct Assessment

A.1 Annotation

In order to increase the quality of our annotations, we firstly provided a set of 50 training samples, and only later provided the other samples to the annotators. The annotators are instructed to score each sentence on a scale 0–100 according to the perceived translation quality (Fomicheva et al., 2022). Specifically, the 0–10 range for incorrect translations; 11–29 for translations with a few correct keywords, but wrongly conveyed meaning; 30–50 for the ones containing major mistakes; 51–69 for translations that convey the meaning of the source, but contain grammatical errors; 70–90 for translations that preserve semantics of the source sentence; and 91–100 for correct translations.

Also, our Croatian annotators are Croatian native students majoring in Linguistics or pursuing a Translation degree. German annotators have the language competence of C1 or above. They were paid per hour. On the contrary, our professional translator was paid according to the regular translation rate in Croatia for a large corpus on a card basis (1800 characters including white spaces).

A.2 Scores Dataset Creation

When resolving final DA scores, if we ended up in a scenario where the outlier was on either side (e.g. $[0, 20, 40]$), we randomly chose one. Furthermore, the process described by Fomicheva et al. (2022) is biased towards the first three annotations, meaning that if two of them are outliers, we'll keep discarding annotations until a third outlier comes. In our process, it happened in $\sim 1\%$ of cases.

B Hyperparameters

Here we outline hyperparameters used for hyperparameter search of our XLM-R Base model on different XNLI training setups. Every model was trained on 3 epochs, and the best one (out of 3 epochs) was chosen based on evaluation results on the dev set. For all of our experiments we used 2 NVIDIA 3090 GPUs.

Name	Value(s)
Max epochs	3
Optimizer	AdamW
Batch size	32
Warmup proportion	6%
Weight decay	0.01
Learning rate	$[8e^{-6}, 1e^{-5}, 3e^{-5}, 5e^{-5}]$
Learning rate scheduler	linear with warmup
Max seq length	128

Table 6: Considered Hyperparameters.

C XNLI Additional Analyses

	ar	**bg**	**de**	**el**	**en**	**es**	**fr**	**hi**	**ru**	**sw**	**th**	**tr**	**ur**	**vi**	**zh**	**hr**
Premise	20.7	20.9	21.1	21.0	21.7	22.1	24.1	23.2	19.6	18.7	22.1	16.8	24.1	27.6	21.8	19.0
Hypothesis	10.2	10.4	10.8	10.6	10.7	10.9	12.4	11.9	9.7	9.0	10.4	8.4	12.3	13.5	10.8	9.3

Table 7: Average token lengths per language for hypotheses and premises reported by Conneau et al. (2018), extended with Croatian using MOSES tokenizer.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5 (first after the conclusion)

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
0,1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2,3

☑ B1. Did you cite the creators of artifacts you used?
1,2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2,3

C ☑ Did you run computational experiments?

3

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
2

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
2, appendix A.1

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
2