xSIM++: An Improved Proxy to Bitext Mining Performance for Low-Resource Languages

Mingda Chen*, **Kevin Heffernan***, **Onur Çelebi, Alex Mourachko, Holger Schwenk**
{mingdachen,kevinheffernan,celebio,alexmourachko,schwenk}@meta.com
Meta AI Research

Abstract

We introduce a new proxy score for evaluating bitext mining based on similarity in a multilingual embedding space: xsim++. In comparison to xsim, this improved proxy leverages rule-based approaches to extend English sentences in any evaluation set with synthetic, hard-to-distinguish examples which more closely mirror the scenarios we encounter during large-scale mining. We validate this proxy by running a significant number of bitext mining experiments for a set of low-resource languages, and subsequently train NMT systems on the mined data. In comparison to xsim, we show that xsim++ is better correlated with the downstream BLEU scores of translation systems trained on mined bitexts, providing a reliable proxy of bitext mining performance without needing to run expensive bitext mining pipelines. xsim++ also reports performance for different error types, offering more fine-grained feedback for model development.

1 Introduction

When training neural machine translation (NMT) systems, it has been shown in prior works that generally, the quality of such systems increases with the availability of high-quality training data (Koehn and Knowles, 2017). However, for many low-resource languages there are few public corpora available, posing many challenges. In order to address this sparsity, one approach is to supplement existing datasets with automatically created parallel corpora, and a technique which has shown to be successful for such issues is the task of bitext mining (Schwenk et al., 2021b).

In bitext mining, the aim is to find pairs of sentences with the same sentence meaning across collections of monolingual corpora. In this work, we adopt a *global mining* approach (Schwenk et al., 2021a), which has shown recent success in providing high-quality data for low-resourced languages (NLLB Team et al., 2022).

In order to evaluate any bitext mining method, a natural approach is to train a NMT system on the automatically created alignments. However, this is extremely costly. As an alternative, the BUCC task (Zweigenbaum et al., 2018) offers a method for evaluating bitext mining algorithms by embedding known alignments within monolingual corpora, and then reporting on the number of correctly aligned pairs. However, this task currently only covers 5 high-resourced languages (English, French, Russian, German and Chinese), and so is not applicable to the low-resource domain. In order to address this, another approach to evaluate bitext mining is to align existing multilingual parallel test sets. Two such test sets are Tatoeba[1] and FLORES200.[2] However, as shown by Heffernan et al. (2022), the Tatoeba corpus is not very reliable given that for some sentence pairs there are only a few hundred sentences. Therefore, we opt to use FLORES200, which is also n-way parallel.

One existing method for evaluating bitext mining on parallel test sets is xsim.[3] This method reports the error rate of misaligned sentences, and follows a margin-based global mining approach (Artetxe and Schwenk, 2019a). However, although using xsim on test sets such as FLORES200 has been shown to be useful as a proxy metric for bitext mining (NLLB Team et al., 2022), it has the following limitations:

1. Using FLORES200 alone has proven to not be difficult enough as for many language pairs, existing approaches quickly saturate at 0% error (NLLB Team et al., 2022).

*Equal contribution

[1] https://github.com/facebookresearch/LASER/tree/main/data/tatoeba/v1

[2] https://github.com/facebookresearch/flores/tree/main/flores200

[3] https://github.com/facebookresearch/LASER/tree/main/tasks/xsim

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 101–109

Transformation Category	Original Sentence	Transformed Sentence
Causality Alternation	Apart from the fever and a sore throat, I feel well and in good shape to carry out my work by telecommuting.	Apart from the fever and a sore throat, I feel well and in bad shape to carry out my work by telecommuting
Entity Replacement	Charles was the first member of the British Royal Family to be awarded a degree.	M. Smith was the first member of The University to be awarded a degree.
Number Replacement	Nadal bagged 88% net points in the match winning 76 points in the first serve.	Nadal bagged 98% net points in the match winning 71 points in the sixth serve.

Table 1: Examples of the transformations applied to the English sentences from FLORES200 dev set. The red texts indicate the places of alternations.

2. As the dev and devtest sets are quite small (997/1012 respectively), this is arguably not a good approximation for performance when mining against billions of possible candidate sentences.

3. We have observed that there is not a significant overlap in the semantics between candidate sentences, meaning that it is not possible to test difficult scenarios that arise in bitext mining when choosing between multiple (similar) candidate pairs.

In order to address these limitations, in this work we introduce `xsim++`. This is an improved proxy for bitext mining performance which expands the dev and devtest sets of FLORES200 to include both more data points, and also difficult to distinguish cases which provide far greater challenges to the models. Our contributions can be summarised as follows:

1. We create a more semantically challenging and expanded English test set for FLORES200.

2. We validate this new test set by independently performing 110 bitext mining runs, training 110 NMT systems on the output mined bitexts, and then determining both the correlation and statistical significance between `xsim++` and the resulting BLEU scores.

3. We open-source the expanded FLORES200 dev and devtest sets, and also the `xsim++` code to evaluate them[4].

2 Methodology

2.1 Background: `xsim`

Given two lists of sentences in different languages, `xsim` seeks to align each sentence in the source language to a corresponding sentence in the target language based on a margin-based[5] similarity (Artetxe and Schwenk, 2019a). In doing so, `xsim` leverages the mining approach described in Artetxe and Schwenk (2019b) to first encode sentences into embedding vectors, assign pairwise scores between sentences in the lists, and then take the sentence in the target language that achieves the maximum score as the final prediction. `xsim` relies on human-annotated parallel corpora and measures the performance of bitext mining using the fraction of misaligned source sentences, i.e., error rates.

2.2 `xsim++`

As the effectiveness of `xsim` is limited by the availability of parallel corpora, we choose to create `xsim++` by automatically expanding the English sentences, and evaluate the sentence encoders on into-English language directions, following prior work on low-resource bitext mining (Heffernan et al., 2022). Aside from the expanded candidate set, `xsim++` follows the same procedure as `xsim`.

`xsim++` seeks to capture more subtle improvements in bitext mining by adding challenging negative examples. The examples are human-written sentences transformed by various operations. These operations intend to perturb semantics through minimal alternations in the surface text. In particular, we use the following categories of transformations: causality alternation, entity replacement, and number replacement. We focus on these three transformation types only as they easily allow us to create negative examples. Examples of the transformed sentences are shown in Table 1. For these transformations, we adapt the implementation in Dhole et al. (2021)[6] and describe the details

[4] https://github.com/facebookresearch/LASER

[5] In this work we report all results using the *absolute* margin

[6] Although this library has additional transformation methods available, many would create positive examples in this use case (e.g. paraphrases).

	Total #	# per orig.
Original	997	-
Causality	1868	1.87
Entity	37745	37.86
Number	3476	3.49

Table 2: Total numbers of original sentences and transformed sentences in different transformation categories. We also report the averaged numbers of transformations per original sentence for each category.

of these transformations below.

Causality Alternation. To alter causality in a sentence, we (1) replace adjectives with their antonyms; (2) negate the meaning of sentences by adding or removing negation function words (e.g. "did not" and "was not") to the sentences; or (3) leverage the negation strengthening approach (Tan et al., 2021), which changes the causal relationships through more assertive function words (e.g. replacing "may" with "will"). For example, as shown in Table 1 we replace "good" with the antonym "bad".

Entity Replacement. We collect candidate entities from large amounts of monolingual data. Then we replace entities in sentences with the ones randomly sampled from the candidate set. For both stages, we use the named entity recognizer from NLTK (Bird et al., 2009).

Number Replacement. We use spaCy (Honnibal and Montani, 2017) to detect dates, ordinals, cardinals, times, numbers, and percentages and then randomly replace their values.

Given the strategies above, for each sentence we create multiple transformations (i.e. negative examples) of that source sentence. For example, consider Table 1. In the "Entity Replacement" example we create a transformation by replacing two named entities. We can then continue this process by replacing these with other named entities until we have reached the desired number of total transformations[7]. Note that since the opportunity to change each category is dependent on the frequency of that category in the evaluation sets, some transformations occurred more than others (e.g. entities were more frequent than numbers). We summarize the data statistics for xsim++ on the FLORES200 dev set in Table 2. Results for the devtest set are in appendix A.

[7]We set a maximum threshold of 100 transformations per category per sentence.

3 Experiment

In order to establish xsim++ as a proxy for bitext mining performance, we measure the correlation between both xsim and xsim++ error rates, and the BLEU scores resulting from NMT systems trained on mined bitexts. More specifically, for each language we choose a sentence encoder model, followed by bitext mining using each respective encoder, and then train and evaluate bilingual NMT systems on the resulting mined bitexts. We use the FLORES200 development sets when computing the BLEU scores.

In order to validate xsim++ against varied embedding spaces, we encode (and mine) using two different multilingual encoder methods: LASER (Artetxe and Schwenk, 2019b) and LaBSE (Feng et al., 2022). For LASER, we trained our own custom encoders (details below). For LaBSE, we used a publicly available model[8] as the code and data for training LaBSE are not publicly available.

We randomly choose 10 low-resource languages to perform both encoder training (if applicable) and bitext mining. The languages are: Faroese (fao), Kabuverdianu (kea), Tok Pisin (tpi), Kikuyu (kik), Friulian (fur), Igbo (ibo), Luxembourgish (ltz), Swahili (swh), Zulu (zul), Bemba (bem).

Encoder Training. We trained LASER encoders using the teacher-student approach described in Heffernan et al. (2022). We choose a LASER model (Artetxe and Schwenk, 2019b) as our teacher, and then trained specialised students for each language. In order to train each student, we used both publicly available code[9] and bitexts (e.g. OPUS[10])

Bitext Mining. For each chosen encoder model, we perform bitext mining against approximately 3.7 billion sentences of English. For low-resource languages, the sizes of monolingual data range from 140k to 124 million. Details are in the appendix. We make use of monolingual data available from both Commoncrawl and Paracrawl[11], and operationalize the mining using the stopes library (An-

[8]https://huggingface.co/sentence-transformers/LaBSE

[9]https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/laser_distillation

[10]https://opus.nlpl.eu

[11]https://paracrawl.eu

drews et al., 2022).[12] For LASER, we use 1.06 as the margin threshold following Heffernan et al. (2022) and for LaBSE, we use 1.16.[13] Following mining, for each language we concatenate publicly available bitexts and the mined bitext as training data for NMT bilingual models using `fairseq`,[14] translating from each foreign text into English. For all NMT systems, we keep the hyperparameters fixed (details in Appendix).

Evaluation. Model selection involves two use cases: comparisons within a model and across different models. For the former comparison, given our custom encoders, we choose to compare 10 checkpoints from each model.[15] For cross model comparisons, we compare each chosen encoder checkpoint against another existing system. In this case, the LaBSE encoder. To quantitatively measure these two cases, we report pairwise ranking accuracy (Kocmi et al., 2021) for `xsim` and `xsim++`. Formally, the accuracy is computed as follows

$$\frac{|\mathrm{s}(\mathrm{proxy}\Delta) = \mathrm{s}(\mathrm{mining}\Delta) \text{ for all system pairs}|}{|\text{all system pairs}|}$$

where proxyΔ is the difference of the `xsim` or `xsim++` scores, miningΔ is the difference of the BLEU scores, $\mathrm{s}(\cdot)$ is the sign function, and $|\cdot|$ returns the cardinal number of the input.

In this work, we have 550 system pairs with 55 pairs per language direction (i.e. $\binom{11}{2}$ pairs given 10 custom LASER encoder checkpoints + LaBSE). We always compare systems within a language direction as the scores for system pairs across different directions are not comparable.[16]

3.1 Results

As shown in Table 3, `xsim++` significantly outperforms `xsim` on the pairwise ranking accuracy. Additionally, when comparing the computational cost to mining, `xsim++` costs over 99.9% less GPU hours and saves approximately 3 metric tons of carbon emissions, but still manages to achieve a competitive accuracy. We observe similar trends for the within a model and across models use cases and report their separate accuracies in the appendix.

Metric	Accuracy	GPU hours
`xsim`	35.48	0.43
`xsim++`	72.00*	0.52
Mining BLEU (Oracle)	100	19569

Table 3: Pairwise ranking accuracy along with the total number of GPU hours. For all experiments, we used NVIDIA A100 GPUs. An $*$ indicates that the result passes the significance test proposed by Koehn (2004) with p-value < 0.05 when compared to `xsim`.

	Accuracy
`xsim++`	72.00
Causality	63.09
Entity	65.45
Number	60.73
Misaligned	40.73
Causality + Entity	68.55
Causality + Entity + Misaligned	70.55
Causality + Misaligned	68.00
Causality + Number	66.73
Causality + Number + Misaligned	71.45
Entity + Misaligned	70.55
Number + Entity	67.45
Number + Entity + Misaligned	71.09
Number + Misaligned	64.36

Table 4: Pairwise ranking accuracy when using combinations of error categories. Causality=Causality Alternation, Entity=Entity Replacement, Number=Number Replacement.

To better understand the contributions of each transformation category (cf. subsection 2.1) in measuring the final mining performance, we report accuracies for different combinations of categories in Table 4. In cases where an incorrect bitext alignment do does not map to any of the augmented sentences of the true alignment, we denote these as "misaligned". We find that entity replacement helps most in improving the accuracy and combing all the transformations gives the best performance.

4 Related Work

As `xsim++` uses rule-based data augmentation, it is related to work in other areas that also employ similar data augmentation methods, such as part-of-speech tagging (Şahin and Steedman, 2018), contrastive learning (Tang et al., 2022), text classification (Kobayashi, 2018; Wei and Zou, 2019), dialogue generation (Niu and Bansal, 2018) and summarization (Chen and Yang, 2021).

[12] https://github.com/facebookresearch/stopes

[13] We did grid search on threshold values from 1.11 to 1.25 on three languages (swh, ltz, and zul), decided the optimal one based on the BLEU scores, and used the threshold for the rest of languages.

[14] https://github.com/facebookresearch/fairseq

[15] Evenly spaced between epochs 1 and 30.

[16] There are factors varied across language directions that are unrelated to the quality of sentence encoders but could affect mining performance, such as amounts of monolingual data available for mining.

5 Conclusion and Future Work

We proposed a proxy score xsim++ for bitext mining performance using three kinds of data augmentation techniques: causality alternation, entity replacement, and number replacement. To validate its effectiveness, we conducted large-scale bitext mining experiments for 10 low-resource languages, and reported pairwise ranking accuracies. We found that xsim++ significantly improves over xsim, doubling the accuracies. Analysis reveals that entity replacement helps most in the improvement. In future work, we plan to extend xsim++ to non-English language pairs.

6 Limitations

We highlight three limitations of our work. The first is that xsim++ is automatically constructed. There could be noisy sentences leading to errors that are irrelevant to the quality of encoders. The second is that xsim++ applies transformations solely to English sentences. Generalizing it to non-English language pairs requires additional research. Finally, we have experimented with the two most popular multilingual encoders: LASER and LaBSE. There are other available approaches which would be interesting to also validate xsim++ against.

References

Pierre Andrews, Guillaume Wenzek, Kevin Heffernan, Onur Çelebi, Anna Sun, Ammar Kamran, Yingzhe Guo, Alexandre Mourachko, Holger Schwenk, and Angela Fan. 2022. stopes - modular machine translation pipelines. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation.

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.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *Findings of EMNLP*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496, Brussels, Belgium. Association for Computational Linguistics.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

Fiona Anting Tan, Devamanyu Hazarika, See-Kiong Ng, Soujanya Poria, and Roger Zimmermann. 2021. Causal augmentation for causal sentence classification. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 1–20, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zilu Tang, Muhammed Yusuf Kocyigit, and Derry Tanti Wijaya. 2022. AugCSE: Contrastive sentence embedding with diverse augmentations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 375–398, Online only. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th workshop on building and using comparable corpora*, pages 39–42.

A Data Statistics for xsim++ with FLORES200 devtest set

	Total #	# per orig.
Original	1012	-
Causality	1916	1.89
Entity	38855	38.39
Number	3262	3.22

Table 5: Total numbers of original sentences and transformed sentences in different transformation categories. We also report the averaged numbers of transformations per original sentence for each category.

We report the data statistics for xsim++ with FLORES200 devtest set in Table 5.

B Sizes of Monolingual data for Low-Resource Languages

Language	Size
kik	147,902
kea	226,507
fur	737,178
fao	1,179,475
tpi	1,661,743
bem	2,302,805
ibo	8,124,418
zul	20,477,331
swh	55,399,821
ltz	123,944,670

Table 6: Number of monolingual sentences for each language.

We report the sizes of monolingual data for each language in Table 6.

C Hyperparameters for NMT systems

encoder layers	6
encoder attention heads	8
encoder embed dim	512
encoder FFNN embed dim	4096
decoder layers	6
decoder attention heads	8
decoder embed dim	512
decoder FFNN embed dim	4096
optimiser	Adam
adam betas	(0.9, 0.98)
learning rate	0.001
dropout	0.3
spm vocab size	7000

Table 7: Hyperparameters for NMT systems.

We report hyperparameters for NMT evaluations in Table 7.

D Within and Across Model Accuracies

Metric	Within	Across
xsim	31.33	54.04
xsim++	69.77*	82.00*

Table 8: Pairwise ranking accuracy for comparisons within a model and across different models. An $*$ indicates that the result passes the significance test proposed by Koehn (2004) with p-value < 0.05 when compared to xsim.

We report accuracies for within a model (i.e., LASER) and across different models (i.e., the 10 LASER checkpoints vs LaBSE) in Table 8.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☐ 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?
Left blank.

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

B ☑ Did you use or create scientific artifacts?

Section 2

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

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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.?
Section 2

☑ 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.
Section 2

C ☑ Did you run computational experiments?

Section 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?
Section 3

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?
Section 3 and Appendix

☑ 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?
Section 3 and Appendix

☑ 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.)?
Section 3

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

Left blank.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.