# Translation Errors Significantly Impact Low-Resource Languages in Cross-Lingual Learning

**Anonymous ACL submission**

## Abstract

Popular benchmarks (e.g., XNLI) used to evaluate cross-lingual language understanding consist of parallel versions of English evaluation sets in multiple target languages created with the help of professional translators. When creating such parallel data, it is critical to ensure high-quality translations for all target languages for an accurate characterization of cross-lingual transfer. In this work, we find that translation inconsistencies *do exist* and interestingly they *disproportionally impact low-resource languages* in XNLI. To identify such inconsistencies, we propose measuring the gap in performance between zero-shot evaluations on the human-translated and machine-translated target text across multiple target languages; relatively large gaps are indicative of translation errors. We also corroborate that translation errors exist for two target languages, namely Hindi and Urdu, by doing a manual reannotation of human-translated test instances in these two languages and finding poor agreement with the original English labels these instances were supposed to inherit.

## 1 Introduction

Multilingual benchmarks, such as XNLI, XTREME, play a vital role in assessing the cross-lingual generalization of multilingual pretrained models ((Conneau et al., 2018b), (Hu et al., 2020)). A common strategy adopted in zero-shot multilingual benchmark creation is to translate development and test sets from English into various target languages with the help of professional human translators. However, such a translation process is susceptible to human errors and could lead to incorrect estimates of cross-lingual transfer to target languages. We find translation errors do emerge and they disproportionately affect translations in certain low-resource languages such as Hindi and Urdu.[1]

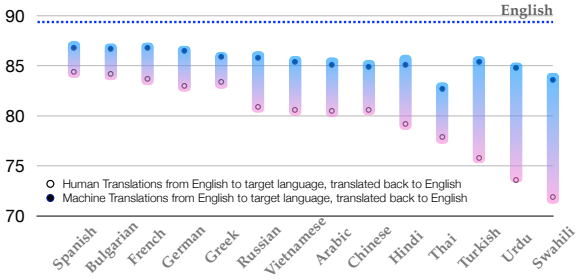


Figure 1: XNLI performance gap by evaluating on translations of human-annotated data in target languages versus paraphrases of the original English data via back-translations pivoted on each target language.

Consider the well-known Cross-Lingual Natural Language Inference (XNLI) benchmark (Conneau et al., 2018a) that contains human translations of English premise-hypothesis pairs (with the labels reproduced from English) into 14 typologically-diverse target languages. Prior work raised concerns about whether the semantic relationships between premise and hypothesis are preserved in such human translations, but did not probe into this issue further (Artetxe et al., 2020a, 2023). We find that there are indeed errors introduced in the human translations leading to label inconsistencies and that this issue disproportionately affects low-resource languages.

To visualize the impact of low-quality translations on low-resource languages, Figure 1 compares zero-shot XNLI performance on all 14 target languages using the XLMR model (Conneau et al., 2020) finetuned on English NLI with the following two input types: 1. Human translations of the original English NLI instances to the target language from XNLI, translated back to English. 2. Machine

[1] In the context of multilingual models, we refer to a language as low (or high)-resource based on the proportion of its data used in model pretraining. XLMR (Conneau et al., 2020) is pretrained on the CC-100 corpus that includes roughly 50GB each of data from *high-resource* languages such as French, Greek and Bulgarian, and only 20.2GB, 5.7GB and 1.6GB of data in *low-resource* languages such as Hindi, Urdu and Swahili, respectively.

translations of the original English NLI instances to the target language, translated back to English. We see a clear differential trend with larger gaps between the (scores over the) two input types for low-resource languages such as Swahili, Urdu and Turkish (appearing on the right) and smaller gaps for high-resource languages such as Spanish, German and French (appearing on the left). We also observe that the *cross-lingual transfer gap* when comparing the performance of human-translations for each target language with that of English (the latter shown as a dotted line) is largely overestimated for low-resource languages.

To summarize, our main contributions are:

(1) We highlight the problem of translation errors in XNLI disproportionately affecting low-resource languages, and propose a practical way of identifying low-quality human translations by comparing their performance with machine translations derived from the original English sentences.

(2) We find the translation errors persist under various train/test settings, including training data derived from machine-translations and paraphrases via backtranslations.

(3) For two low-resource languages Hindi and Urdu, we manually annotate a subset of NLI data and find large discrepancies in the newly annotated labels when compared to the labels projected from the original English sentences.

## 2 Experimental Setup

### 2.1 Tasks and Models

Our main focus is on the popular XNLI (Conneau et al., 2018b) benchmark, which is a three-way classification task to check whether a premise entails, contradicts or is neutral to a hypothesis. Parallel to English NLI ((Bowman et al., 2015), (Williams et al., 2018)), XNLI consists of development sets (2490 instances) and test sets (5010 instances) in 14 typologically-diverse languages[2] Translation-based gap analysis on two other multilingual tasks (MLQA and PAWSX) is included in Appendix A.

We use XLM-Roberta (XLMR) (Conneau et al., 2020) as the pretrained multilingual model in all our experiments. (Appendix B reports scores using mBERT (Devlin et al., 2019) for XNLI that follow the same trends.)

### 2.2 Training and Test Variants

(Artetxe et al., 2020a) showed that using machine-translated data to finetune the pretrained model helps it generalize better to both machine and human-translated test data. Motivated by this finding, we construct the following training variants:

(1) ORIG: Original English training data.

(2) Backtranslated-train (B-TRAIN): English paraphrases of the original English data via backtranslations, with Spanish as a pivot.

B-TRAIN is a training variant introduced in (Artetxe et al., 2020a) that we adopt in our work.

We also evaluate on the following four variants of test data:

(1) Zero-shot (ZS): Human-translated dev/test sets in the target languages.

(2) Translate-test (TT): Machine translations of target language dev/test sets to English.

(3) Translate-from-English (TE): Machine translations of original English to the target languages.

(4) Backtranslation-via-target (BT): Machine translations of original English to the target language and back to English.

We use two translation systems to create the above variants: 1) A state-of-the-art open-source multilingual translation model from the No Language Left Behind (NLLB) project (NLLB Team et al., 2022), and 2) Google's Cloud Translate API.[3] Due to the prohibitive cost of the latter for the creation of training data, we use NLLB to create all our training variants (unless specified otherwise).[4] Test variants were created using both translation systems. More implementation details and translation-related details are provided in Appendix C and Appendix D. Some of the types of translation errors in the human-translated dev/test sets in ZS and TT are illustrated in Appendix E.

[2]Languages include French (fr), Spanish (es), German (de), Greek (el), Bulgarian (bg), Russian (ru), Turkish (tr), Arabic (ar), Vietnamese (vi), Thai (th), Chinese (zh), Hindi (hi), Swahili (sw) and Urdu (ur).

[3]https://cloud.google.com/translate

[4]We found NLLB to be poor in quality when translating from English to Chinese. We used the M2M translation system (Fan et al., 2020) for English-to-Chinese that was far superior.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZS | 89.3 | 83.5 | 84.8 | 83.4 | 82.4 | 83.7 | 80.5 | 79.4 | 79.2 | 79.9 | 78.3 | 79.4 | 77.2 | 72.7 | 74.0 | 79.9 |
| TT-n | - | 82.1 | 83.1 | 80.7 | 82.3 | 82.6 | 79.3 | 75.9 | 78.0 | 78.7 | 73.8 | 77.6 | 77.7 | 70.5 | 71.3 | 78.1 |
| TT-g | - | 83.7 | 84.4 | 83.0 | 83.4 | 84.2 | 80.9 | 75.8 | 80.5 | 80.6 | 77.9 | 80.6 | 79.2 | 71.9 | 73.6 | 79.9 |
| BT-n | - | 84.5 | 84.9 | 83.5 | 82.9 | 82.7 | 82.3 | 81.1 | 81.4 | 82.4 | 76.4 | 79.6 | 82.9 | 79.4 | 80.8 | 81.8 |
| TE-n | - | 84.4 | 85.5 | 83.9 | 83.6 | 83.9 | 83.4 | 81.7 | 81.5 | 81.9 | 78.7 | 81.0 | 82.1 | 77.0 | 80.3 | 82.1 |
| TE-g | - | 85.3 | 85.9 | 85.9 | 84.8 | 86.1 | 84.9 | 83.8 | 82.7 | 84.0 | 82.0 | 84.3 | 82.1 | 77.3 | 81.8 | 83.6 |
| BT-g | - | **86.6** | **86.8** | **86.5** | **85.9** | **86.7** | **85.8** | **85.4** | **85.1** | **85.4** | **82.7** | **84.9** | **85.1** | **83.6** | **84.8** | **85.4** |
| Δ-g | | 2.9 | 2 | 3.1 | 2.5 | 2.5 | 4.9 | 6 | 4.6 | 4.8 | 4.4 | 4.3 | 5.9 | 10.9 | 10.8 | 4.9 |

Table 1: Results of ORIG (model trained on original English data) evaluated on different test set variants described in Section 2.2. -n refers to using NLLB as the translator, -g refers to using Google-translate as the translator. Highest scores in each column are shown in bold and next highest is underlined.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZS | 89.2 | 84.5 | 85.9 | 84.6 | 84.3 | 85.5 | 82.9 | 81.0 | 81.8 | 82.6 | 79.8 | 80.9 | 79.6 | 74.7 | 75.6 | 81.7 |
| TT-n | - | 84.0 | 85.7 | 82.4 | 84.4 | 84.4 | 81.8 | 78.9 | 81.0 | 80.9 | 77.4 | 80.5 | 80.5 | 73.6 | 74.4 | 80.7 |
| TT-g | - | 84.8 | 86.5 | 84.1 | 85.1 | 85.9 | 82.7 | 78.9 | 83.1 | 82.7 | 80.4 | 82.6 | 81.4 | 74.9 | 76.9 | 82.1 |
| BT-n | - | 85.9 | 86.8 | 85.1 | 84.8 | 84.6 | 84.3 | 82.8 | 83.5 | 84.2 | 79.3 | 81.4 | 84.8 | 81.9 | 82.5 | 83.7 |
| TE-n | - | 85.8 | 86.8 | 85.2 | 84.9 | 85.2 | 84.6 | 83.0 | 83.5 | 83.6 | 80.6 | 82.0 | 83.4 | 79.1 | 81.4 | 83.5 |
| TE-g | - | 86.6 | 87.0 | 86.9 | 85.5 | 86.4 | 86.4 | 84.3 | 84.6 | 84.9 | 83.3 | 84.6 | 83.5 | 78.9 | 82.9 | 84.7 |
| BT-g | - | **88.0** | **87.7** | **87.6** | **86.7** | **87.5** | **87.1** | **85.9** | **86.4** | **86.2** | **84.2** | **85.9** | **85.9** | **85.4** | **86.1** | **86.5** |
| Δ-g | | 3.2 | 1.2 | 2.5 | 1.6 | 1.6 | 4.2 | 4.9 | 3.3 | 3.5 | 3.8 | 3.3 | 4.5 | 10.5 | 9.2 | 4.3 |

Table 2: Results of B-TRAIN on different test set variants described in Section 2.2.

## 3 Cross-lingual Transfer Gap in XNLI

### 3.1 Using Original English NLI Train Set

Table 1 presents XNLI F1 scores for all four test variants using ORIG training data. Test translations are generated using both NLLB (-n) and Google Translate (-g). Δ-g in Table 1 refers to the performance gap when using human vs. machine translations. It is the difference between the F1 for BT-g (machine-translated target language text) and the best F1 among ZS and TT-g (human-translated target language text). It is striking that Δ-g values for low-resource languages like Urdu and Swahili are as high as 10.8 and 10.9, respectively, and as low as 2.9 and 2 for high-resource languages like French and Spanish, respectively.

### 3.2 Using Translated Train Sets

Table 2 shows test accuracies using an XLMR model finetuned on B-TRAIN. Across all target languages and all test set variants, we see consistent improvements in performance compared to ORIG in Table 1. This is consistent with the observation in (Artetxe et al., 2020a) that finetuning on backtranslation-driven paraphrases helps generalize better to both human and machine translated test sets. Interestingly, even with the overall improvements using B-TRAIN, the large performance gap between ZS and TE (and TT and BT) for low-resource languages like Urdu and Swahili persists.

**Overestimated Cross-lingual Gap.** Based on (Hu et al., 2020), we compute cross-lingual transfer gap as the difference between English F1 and the average of F1 scores across all other languages. From Table 2, the previously reported cross-lingual gap was 7 using ZS, which reduces to 2.7 using BT-g. The largest gaps for an individual language were previously 14.5 and 13.6 for Swahili and Urdu and have now reduced to 3.8 and 3.1 with BT-g, respectively. This suggests a quick recipe for a quality check of human translations. For target languages supported by machine-translation systems, the performance gap between either ZS and TE or between TT and BT could be a quick way to check whether the human translations might have issues during the data collection phase (thus yielding large gap values).

## 4 Human Evaluation

For two low-resource languages Hindi and Urdu, we reannotate a subset of the human-translations

with NLI labels and check how well they match the labels inherited from the original English text. We pick random, non-overlapping sets of 200 instances each in English, Hindi and Urdu and get them relabelled by native speakers. (Appendix F provides more annotation details.) The new labels matched the original labels 90.5%, 66.5% and 60% of the time for English, Hindi and Urdu, respectively. This clearly highlights the large drop in label agreement for Hindi and Urdu compared to English, with relative reductions of 24% and 30.5% for Hindi and Urdu, respectively. In (Conneau et al., 2018a), the same experiment was conducted using English and French and the original labels were recovered 85% and 83% of the time, respectively. The authors concluded there was no loss of information in the translations. However, we find there to be a significant loss of information in translations for languages such as Hindi and Urdu.

To verify if machine translations (TE), rather than XNLI's human translations (ORIG), align better with the labels from the original English, we relabel 200 instances translated from English to Hindi and Urdu (via Google Translate). The annotators recovered the ground-truth labels 80% and 71% of the time for Hindi and Urdu, respectively, highlighting that label inconsistencies in Hindi/Urdu human translations (ORIG) are significantly worse than with machine translations (TE).

## 5 Attention-based Analysis

We assess how the attention distributions learned for XNLI over the English test instances correlate with the attention distributions learned for human-annotated Hindi/Urdu/Swahili test instances and Google-translated (English to) Hindi/Urdu/Swahili test instances. For each correctly predicted English instance, we consider both human-translated (HT) and machine-translated (MT) target language translations and compute word alignments between English and these translations using awesome-align (Dou and Neubig, 2021a). Aligned words whose attention score is greater than the mean attention score for the sequence are counted and normalized by the total number of such words in a sequence. Finally, we compute an average over all these overlap fractions across instances in the dataset. These mean overlap scores shown in Table 3 are computed separately using the human translations (HT) and machine translations (MT). For all three languages, we find the overlap fraction to be higher for the Google-translated sentences compared to the human-translated sentences. This suggests that MT aligns better with the original English text compared to HT.

| text/lang | ur | hi | sw |
|---|---|---|---|
| HT | 0.75 | 0.78 | 0.79 |
| MT | **0.86** | **0.84** | **0.84** |

Table 3: Aggregate attention scores over aligned words in Human Translated (HT) and Machine Translated (MT) XNLI test instances with parallel English data.

## 6 Related Work

There is growing interest in building multilingual benchmarks for the evaluation of cross-lingual transfer. E.g., XTREME (Conneau et al., 2019) covering a wide range of languages and tasks including XNLI (Conneau et al., 2018a), XQuAD (Artetxe et al., 2020b), PAWS-X (Yang et al., 2019) and MLQA (Lewis et al., 2019). Recently, many extensions of XTREME: IndicXTREME (Doddapaneni et al., 2022) focusing on 18 Indian languages, XTREME-R (Ruder et al., 2021) and XTREME-UP (Ruder et al., 2023) have also been released. Translation artifacts have only been studied in select prior works. Mohammad et al. (2016) study how translations can alter sentiment labels in Arabic text. In very recent work, Artetxe et al. (2023) advocate for the use of English-only finetuning using machine-translation systems. However, this relies on high-quality human translations in the target languages which we highlight needs to be carefully examined especially for low-resource languages.

## 7 Conclusions

This work studies the problem of translation irregularities in evaluation sets of multilingual benchmarks like XNLI that are created by translating English into multiple target languages. We find that the translation sets of low-resource languages like Urdu, Swahili exhibit most number of inconsistencies while translations of high-resource languages like French, German are more immune to this problem. We suggest an effective way to check the quality of human translations by comparing performance with machine translations, and show how the cross-lingual transfer estimates can significantly vary with improved translations.

## 8 Limitations

For tasks that have output labels directly corresponding to the input text (e.g., sequence labeling tasks like POS-tagging, question answering, etc.), it would be trickier to use our technique since translations could change the word order and subsequently affect the output labels as well.

We highlight the problem of the cross-lingual transfer gap for low-resource languages being mischaracterized due to poor performance on these languages stemming from poor-quality translations and not necessarily because the model has difficulty with the given target languages. We do not offer a solution to deal with translation errors. Rather, we ask for additional checks when collecting translations for low-resource languages.

We identify that the existing translation datasets for low-resource languages in XNLI have inconsistencies. While we did not create manually-corrected versions of these translation sets, we will be releasing the machine-translated text from English to these target languages upon publication.

## Ethics Statement

We would like to emphasize our commitment to upholding ethical practices throughout this work. We aimed to ensure that human annotators received a fair compensation for their annotation efforts and was commensurate with the time and effort invested in their work. For translations using Google Translate, we used the paid Cloud API service in accordance with the terms and conditions of usage.

## References


Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. *arXiv preprint arXiv:2305.14240*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. 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, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

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*, pages 632–642, Lisbon, Portugal. 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, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018a. Xnli: Evaluating cross-lingual sentence representations.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Indicxtreme: A multi-task benchmark for evaluating indic languages.

Zi-Yi Dou and Graham Neubig. 2021a. Word alignment by fine-tuning embeddings on parallel corpora.

Zi-Yi Dou and Graham Neubig. 2021b. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

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 generalization. *CoRR*, abs/2003.11080.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: evaluating cross-lingual extractive question answering. *CoRR*, abs/1910.07475.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *J. Artif. Int. Res.*, 55(1):95–130.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification.

| **F1/EM** (# sents) | **en** (4918) | **hi** (4918 ) | **en** (5495 ) | **vi** (5495) |
|---|---|---|---|---|
| ZS | 83.2/69.8 | 70.6/52.9 | 83.4/70.6 | 74.0/52.7 |
| TT-n | - | 78.4/64.5 | - | 74.9/61.3 |
| BT-n | - | **78.4/64.7** | - | **76.7/63.2** |

Table 4: Results on TT-n and BT-n MLQA test sets. BT-n Hi indicates backtranslated data pivoted through Hindi, TT-n Hi indicates test set in Hi translated to En. (Note that for MLQA only questions are translated.)

## A Performance Gap Analysis for MLQA, PAWS-X

Multilingual (Extractive) Question Answering ((Lewis et al., 2019), MLQA) consists of questions in English translated to six different languages including Arabic(ar), German(de), Spanish(es), Hindi(hi), Vietnamese(vi) and Chinese(zh) amounting to 5K instances in each target language. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification (Yang et al., 2019) consists of dev/test paraphrases in English translated to six different languages: French(fr), Spanish(es), German(de), Chinese(zh), Japanese(ja), and Korean(ko) with the help of human translators.

**MLQA.** For MLQA, we translate questions in the two low-resource languages, Hindi and Vietnamese, to English using NLLB (TT). We also create a BT version of the original English questions (2.2) using Hindi and Vietnamese as pivots.

Table 4 shows TT and BT scores for Hindi are nearly identical and there is a small improvement using BT for Vietnamese compared to TT. This indicates that the professional annotators did not

| Instructions |
| --- |
| Given premise and hypothesis, label each pair as "entailment", "contradiction" or "neutral" as follows:<br>1. if hypothesis is entailed by the premise, it's an "entailment" ,<br>2. if the hypothesis contradicts the premise (hypothesis cannot be True given the premise), it's a "contradiction",<br>3. if the hypothesis is independent of the premise (hypothesis may or may not be True given the premise), it's a "neutral" relationship. |

Table 5: Task description shared with the annotators for the NLI task

introduce semantic inconsistencies during translation for MLQA. In general, classification tasks like XNLI appear to be more susceptible to translation inconsistencies since the annotators are not made aware of the ground-truth labels during translation and are only asked to independently translate the premise/hypothesis pairs.

**PAWS-X.** Table 6 shows the results of the different settings ZS, TE, TT, and BT for the six languages. The model used for inference is xlm-roberta-large trained on the English train set. TE is better than ZS mainly for Korean (by 4.9% in test set) and Chinese (4.9% in dev set) and is nearly equal for other languages. BT is better than TT again for Korean and Chinese and nearly equal for other languages. This indicates the presence of human translation inconsistency for the two languages.

## B Comparing the Performance of mBert and XLMR

As can be seen in Table 7, XLMR outperforms mBert by a huge margin on every language. Thus, we used XLMR for evaluating all our experiments.

## C Details of Model Training

The models mBert and XLMR were trained using the same setting as mentioned in the XTREME repository.[5]

**XNLI.** mBert is trained for 2 epochs with a learning rate of 2e-5, with a batch size of 8 and gradient accumulation of 4(i.e an effective batch size of 32). XLMR is trained for 2 epochs with a learning rate of 5e-6, batch size of 5 and gradient accumulation steps of 6 (i.e effective batch size of 30). The final model is selected from the best checkpoint, which is based on the model's performance on the English dev set. For training the different variants of the model (ORIG, T-TRAIN, B-TRAIN, BT-enes, MT-hi-g, MT-hi-n) we use the same hyperparameter setting as mentioned above.
We use xlm-roberta-large for all our experiments. Model training was done on a single Nvidia Geforce GTX 1080 Ti GPU, which has a RAM of 12GB. It took us around one day to train a single model for 2 epochs. For data translation using NLLB(3.3B parameter model), we made use of the NVIDIA A100-SXM4-80GB gpu for faster processing. Translating the test sets took couple of hours(1-1.5).

**MLQA.** To evaluate the performance on MLQA dataset, we trained XLMR on the SQUAD dataset (Rajpurkar et al. (2016)). The model is trained for 3 epochs with a learning rate of 3e-5, batch size of 1 and gradient accumulation of 32 (i.e an effective batch size of 32).

**PAWS-X.** We trained xlm-roberta-large model on the English train set. The model is trained for 5 epochs with a learning rate of 2e-5, batch size of 2 and gradient accumulation of 16 (i.e an effective batch size of 32).

## D Details of Train and Test Translations

To train the model on back-translated (using Spanish as the pivot) and machine-translated(translated to Hindi and Spanish) data, we made use of the open-source 3.3B parameter NLLB model hosted on Hugging-Face[6]. We found that the English to Chinese translation using NLLB is of lower quality, so we tried the open source 1.2B parameter M2M ((Fan et al., 2020)) model [7]and it performed better compared to the NLLB translator.

## E Examples of Translation Errors

Table 8 highlights a few examples of premise-hypothesis pairs in Hindi and Urdu that are no longer semantically consistent with the original labels (copied from English) after translation. These examples would be marked as errors in predictions, when in fact the predictions are reasonable given the semantic deviations in the human-translated Hindi/Urdu sentences from the original English sentences.

[5]https://github.com/google-research/xtreme

[6]https://huggingface.co/facebook/nllb-200-3.3B

[7]https://huggingface.co/facebook/m2m100_1.2B

| dev/test | en | de | es | fr | ja | ko | zh | avg |
|---|---|---|---|---|---|---|---|---|
| sents | (2000/2000) | (2000/2000) | (2000/2000) | (2000/2000) | (2000/2000) | (2000/2000) | (2000/2000) | - |
| ZS | 95/95.9 | 89/90.9 | 90.4/90.4 | 91.4/91.6 | 82.9/80.5 | 83.6/80.8 | 83.9/84.2 | 86.9/86.4 |
| TT-n | - | 88.9/89.9 | 89.8/91 | 90.4/91.6 | 83/79 | 82.2/80.4 | 81.6/80.9 | 86.0/85.5 |
| TE-n | - | 91.2/92.3 | 92.1/92.3 | 90.9/91.2 | 83.7/83.4 | 86.8/85.7 | 88.8/88.6 | 88.9/88.9 |
| BT-n | - | 90.6/91.5 | 91.6/92.2 | 90.8/90.8 | 81.9/80.6 | 84/84.4 | 89/88.2 | 88.0/88.0 |

Table 6: Results on ZS, TE, TT, and BT PAWS-X.

| dev | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XLMR | 89.9 | 84.2 | 85.0 | 84.3 | 81.8 | 83.2 | 79.7 | 79.9 | 79.2 | 81.6 | 78.0 | 80.0 | 78.3 | 72.1 | 74.6 | 80.8 |
| mBert | 83.0 | 74.9 | 74.8 | 72.2 | 67.8 | 68.2 | 68.4 | 63.4 | 65.4 | 69.8 | 54.8 | 70.6 | 61.5 | 52.4 | 53.3 | 66.7 |

Table 7: Zero shot performance of ORIG mBert and XLMR models on the XNLI target dev sets.

## F Details of Human Annotations

Each task (set of random 200 sentences) is annotated independently by two annotators. The task description shared with the annotators is included in Table 5. The sentences in agreement between the two annotators are reviewed and approved for the dataset by the final annotator. If there is a mismatch, it is sent to the two annotators for review and possible corrections. If the mismatch persists, a third annotator performs a fresh annotation. The final annotator reviews the 3 answers and submits the final answer for the dataset. We also computed the Cohen's Kappa score between the two annotators and found them to be: 0.64 for English sentences, 0.43 for Hindi sentences, and 0.37 for Urdu sentences.

## G Tools and Libraries

We made use of awesome-align ((Dou and Neubig, 2021b)) to align words between English and any target language. The model used by awesome-align was bert-base-multilingual-cased. We used the Pytorch framework[8] and Hugging-face library[9] for all our model training and inferencing tasks. To integrate Labse ((Feng et al., 2020)), we made use of the Sentence-transformers library[10]. To convert the transliterated sentences to the original scripts, we made use of both google-translate and Indictrans(Bhat et al. (2015)) (for Indian languages). We made use of the google-cloud-translate api to use the google-translate services.

[8]https://pytorch.org/
[9]https://huggingface.co/
[10]https://www.sbert.net/

## H More Trained Models

We trained a few more models in different settings to check their impact on the cross-lingual performance despite presence of semantic irregularities. The additional models we trained include:

1. T-TRAIN is the model trained on English train set machine translated to Spanish. (See Table 9.)

2. BT-enes, i.e train the model on backtranslated english (using Spanish as a pivot) + the original English.

3. MT-hi-g, i.e train the model on machine-translated train set where the train set is translated to Hindi using google-translate. Here we used only 1/3rd of training data to train the model(to incur low costs of translation).

4. MT-hi-n, this is the same as above, except that the translation is performed using NLLB translator.

Using T-TRAIN is more effective in improving test performance across all target languages compared to using ORIG
Tables 10, 11, 12 shows the results of the trained models across different test settings (test sets translated using NLLB). The figures highlight the potential semantic gap that exists between BT and TT (also ZS and TE) across all the models which increases more towards the low resource languages.
In Table 13 and 14, we compare the zero shot and translate-test results of all the trained models across different languages. B-TRAIN and BT-enes performs the best across majority of the languages.

| Premise | Hypothesis | En-Premise | En-Hypothesis | Label | Pred | Comment |
|---|---|---|---|---|---|---|
| Aise hi choti si baatein bhane mera karm par ek bada antar bana diya | Mei kuch hasil karne ki koshish kar raha tha. | Little things like that made a big difference in what I was trying to do. | I was trying to accomplish something. | E | N | Incorrect translation of premise changes the relationship between the label and the premise-hypothesis pair. |
| Mei tumhe ek ghante mei wapas phone karta hoo, ve kehte hai. | Usne kaha ki ve bol rahe the. | I'll call you back in about an hour, he says. | He said they were done speaking. | C | E | Hypothesis is incorrectly translated leading to a change in meaning (i.e "they were done speaking" is translated to "they were speaking"). |
| Wo qaed nahin rehna chahte they | Unhe kuch mawaqe par pakda ja sakta tha lekin wo is se bachna chahte they | They didn't want to stay captive. | They had been captured at some point but wanted to escape. | N | C | Tense is incorrect in the translation of the hypothesis. The premise implies that they have already been captured while the incorrect translation implies that they did not want to get caught, hence predicting a contradiction. |
| Ye tha, ye ek khoobsoorat din tha | Aj ek aramdah din tha | That was, that was a pretty scary day. | It was a relaxing day. | C | N | Tense is incorrectly altered to present and "pretty scary" is translated to simply "khoobsoorat"(pretty), thus inverting the overall sentiment. |

Table 8: Semantically incorrect examples of premise-hypothesis pairs in Hindi (first two) and Urdu (latter two). E, N and C implies entailment, neutral and contradiction labels.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZS | 88.9 | 84.8 | 85.7 | 84.8 | 84.4 | 85.0 | 82.9 | 80.9 | 81.2 | 81.9 | 78.9 | 80.7 | 79.6 | 74.9 | 75.9 | 81.7 |
| TT-n | - | 83.2 | 84.5 | 82.4 | 83.9 | 84.1 | 81.3 | 78.4 | 80.6 | 80.7 | 76.6 | 79.7 | 80.1 | 73.1 | 74.2 | 80.2 |
| TT-g | - | 84.3 | 85.9 | 84.2 | 84.8 | 85.2 | 82.8 | 77.8 | 82.5 | 81.9 | 79.9 | 82.2 | 81.1 | 74.3 | 76.0 | 81.6 |
| BT-n | - | 85.2 | 86.2 | 84.6 | 84.8 | 84.2 | 83.9 | 82.3 | 83.3 | 83.9 | 79.2 | 81.6 | 84.4 | 81.4 | 81.9 | 83.4 |
| TE-n | - | 85.3 | 86.3 | 85.1 | 84.4 | 84.9 | 84.7 | 82.5 | 83.1 | 83.9 | 79.9 | 81.8 | 83.0 | 79.0 | 81.4 | 83.2 |
| TE-g | - | 86.2 | 86.6 | 86.5 | 85.1 | 86.8 | 86.0 | 83.9 | 84.1 | 85.0 | 82.7 | 84.5 | 83.4 | 79.4 | 82.8 | 84.5 |
| BT-g | - | **87.0** | **87.3** | **87.3** | **86.7** | **87.0** | **86.7** | **85.7** | **86.0** | **86.1** | **83.8** | **85.5** | **85.8** | **84.6** | **85.5** | **86.1** |
| Δ-g | | 2.2 | 1.4 | 2.5 | 1.9 | 1.8 | 3.8 | 4.8 | 3.5 | 4.2 | 4.1 | 3.3 | 4.7 | 9.7 | 9.5 | 4.1 |

Table 9: Results of T-TRAIN on different test set variants described in Section 2.2.

Table 15, 16 compares the zero-shot and translate-test results of the MT-hi models, it can be seen that both the models perform equally across the languages, also because of training on less amount of data, their zero-shot performance is very slightly inferior to the ORIG model.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZS | 89.8 | 85.1 | 86.2 | 84.6 | 84.1 | 85.2 | 82.4 | 81.3 | 81.2 | 81.9 | 79.3 | 80.9 | 78.6 | 74.9 | 76.0 | 82.1 |
| TT-n | - | 84.2 | 85.2 | 82.6 | 84.8 | 84.8 | 81.9 | 78.8 | 81.7 | 81.1 | 78.2 | 80.3 | 80.7 | 73.8 | 75.1 | 80.9 |
| BT-n | - | **85.9** | **86.6** | **85.0** | **85.0** | **85.2** | **84.2** | **83.2** | **83.6** | **84.8** | **79.4** | **81.9** | **85.2** | **82.1** | **82.8** | **83.9** |
| TE-n | - | **85.9** | **87.0** | **85.2** | **84.5** | **85.3** | **84.6** | **83.1** | **83.6** | **84.2** | **80.1** | **82.7** | **82.9** | **78.7** | **80.8** | **83.5** |

Table 10: Results of BT-enes (model trained on back-translated(en→es→ en) + original English train set) on different test set data settings 2.2, **-n** refers to using NLLB translator.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZS | 87.4 | 82.9 | 84.2 | 82.7 | 83.4 | 83.4 | 81.1 | 80.8 | 79.9 | 80.4 | 78.1 | 79.9 | 78.8 | 74.1 | 75.3 | 80.8 |
| TT-n | - | 81.7 | 82.6 | 80.1 | 82.2 | 82.3 | 80.3 | 76.2 | 79.4 | 79.3 | 75.8 | 77.9 | 78.5 | 72.2 | 72.5 | 78.6 |
| BT-n | - | **83.9** | **84.4** | **83.4** | **82.7** | 81.8 | **82.3** | **80.1** | **81.5** | **82.2** | **77.5** | **80.0** | **83.3** | **79.9** | **81.0** | **81.7** |
| TE-n | - | **83.7** | **84.9** | **83.6** | 83.0 | **83.5** | **82.8** | **81.5** | **82.0** | **82.3** | **79.4** | **81.1** | **82.7** | **78.2** | **81.4** | **82.1** |

Table 11: Results of MT-hi-g (model trained on data translated to Hindi (en→hi) using google-translate) on different test set data settings 2.2.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ZS | 87.2 | 83.4 | 83.6 | 82.9 | 82.7 | 83.4 | 81.8 | 79.9 | 79.9 | 80.1 | 78.7 | 80.6 | 78.4 | 73.6 | 74.9 | 80.7 |
| TT-n | - | 82.2 | 83.6 | 80.6 | 82.6 | 82.6 | 80.38 | 76.4 | 79.6 | 79.5 | 76.9 | 78.8 | 79.4 | 72.73 | 73.2 | 79.2 |
| BT-n | - | **83.7** | **84.7** | **83.4** | **83.0** | **82.7** | **82.3** | **80.6** | **81.9** | **82.9** | **78.2** | **80.7** | **83.4** | **80.2** | **81.6** | **82.1** |
| TE-n | - | **83.8** | **84.8** | **83.5** | **82.9** | **83.7** | **82.6** | **81.2** | **82.1** | **81.9** | **79.2** | **81.3** | **82.6** | **78.1** | **80.9** | **82.0** |

Table 12: Results of MT-hi-n (model trained on data translated to Hindi (en→hi) using NLLB-translate) using different data settings 2.2.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ORIG | 89.3 | 83.5 | 84.8 | 83.4 | 82.4 | 83.7 | 80.5 | 79.4 | 79.2 | 79.9 | 78.3 | 79.4 | 77.2 | 72.7 | 74.0 | 80.5 |
| B-train | 89.2 | 84.5 | 85.9 | 84.6 | 84.3 | **85.6** | **82.9** | 81.0 | **81.8** | **82.6** | **79.8** | **80.9** | **79.6** | 74.7 | 75.6 | **82.2** |
| BT-enes | **89.8** | **85.1** | **86.2** | 84.6 | 84.1 | 85.2 | 82.4 | **81.3** | 81.2 | 81.9 | 79.3 | **80.9** | 78.6 | **74.9** | **76.1** | 82.1 |
| T-TRAIN | 88.9 | 84.8 | 85.7 | **84.8** | **84.4** | 85.0 | 82.2 | 80.9 | 81.2 | 81.9 | 78.9 | 80.7 | **79.6** | **74.9** | 75.9 | 81.9 |

Table 13: Comparing zero-shot test set results of different trained models (translations performed using NLLB).

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ORIG | - | 82.1 | 83.1 | 80.7 | 82.3 | 82.6 | 79.3 | 75.9 | 78.0 | 78.7 | 73.8 | 77.6 | 77.7 | 70.5 | 71.3 | 78.1 |
| B-TRAIN | - | 84.0 | **85.7** | 82.4 | 84.4 | 84.4 | 81.8 | **78.9** | 81.0 | 80.9 | 77.4 | **80.5** | 80.5 | 73.6 | 74.4 | 80.7 |
| BT-enes | - | **84.2** | 85.2 | **82.6** | **84.8** | **84.8** | **81.9** | 78.8 | **81.7** | **81.1** | **78.2** | 80.3 | **80.7** | **73.8** | **75.1** | **80.9** |
| T-TRAIN | - | 83.2 | 84.5 | 82.4 | 83.9 | 84.1 | 81.3 | 78.4 | 80.6 | 80.7 | 76.6 | 79.7 | 80.1 | 73.1 | 74.2 | 80.2 |

Table 14: Comparing translate-test (using NLLB translator) test set results of different trained models.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MT-hi-g | **87.4** | 82.9 | **84.2** | 82.7 | **83.4** | **83.4** | 81.1 | **80.8** | **79.9** | **80.4** | 78.1 | 79.9 | **78.8** | **74.1** | **75.3** | **80.8** |
| MT-hi-n | 87.2 | **83.4** | 83.6 | **82.9** | 82.7 | **83.4** | **81.8** | 79.9 | **79.9** | 80.1 | **78.7** | **81.2** | 78.4 | 73.6 | 74.9 | 80.7 |

Table 15: Comparing zero-shot test set results of models trained on machine-translated Hindi (1/3rd of training data), hi-g implies using google translator and hi-n implies using NLLB translator.

| test | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MT-hi-g | - | 81.7 | 82.6 | 80.1 | 82.2 | 82.3 | 80.3 | 76.2 | 79.4 | 79.3 | 77.9 | 76.5 | 78.5 | 72.2 | 72.5 | 78.7 |
| MT-hi-n | - | **82.2** | **83.6** | **80.6** | **82.6** | **82.6** | **80.4** | **76.4** | **79.6** | **79.5** | **76.9** | **78.8** | **79.4** | **72.7** | **73.2** | **79.2** |

Table 16: Comparing translate-test (using NLLB translator) test set results of models trained on machine-translated Hindi(1/3rd of training data), hig implies using google translator and hin implies using NLLB translator.