On the Role of Parallel Data in Cross-lingual Transfer Learning

Machel Reid[*]
Google DeepMind
machelreid@google.com

Mikel Artetxe[†]
Reka AI
mikel@rekai.ai

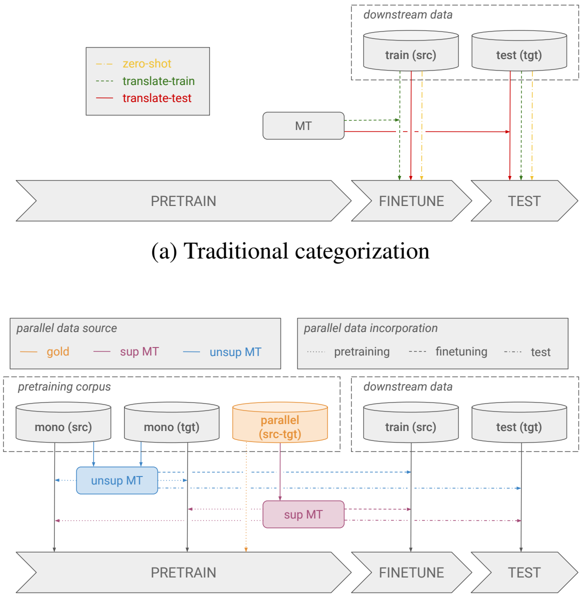
Abstract

While prior work has established that the use of parallel data is conducive for cross-lingual learning, it is unclear if the improvements come from the data itself, or it is the modeling of parallel interactions that matters. Exploring this, we examine the usage of unsupervised machine translation to generate synthetic parallel data, and compare it to supervised machine translation and gold parallel data. We find that even model generated parallel data can be useful for downstream tasks, in both a general setting (continued pretraining) as well as the task-specific setting (translate-train), although our best results are still obtained using real parallel data. Our findings suggest that existing multilingual models do not exploit the full potential of monolingual data, and prompt the community to reconsider the traditional categorization of cross-lingual learning approaches.

1 Introduction

Multilingual models have been shown to generalize across languages in a zero-shot fashion (Pires et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Kale et al., 2021). These models are usually pretrained on concatenated monolingual corpora in multiple languages using some form of language modeling or denoising objective. The models are then finetuned using labeled downstream data in the source language (typically English), which makes them capable of generalizing to the target language thanks to the aligned representations learned at pretraining.

While this paradigm does not require any data in the target language other than the monolingual pretraining corpus, prior work has reported improved results by incorporating **parallel data** into the pipeline, either at pretraining or finetuning time.



(a) Traditional categorization

(b) All possibilities to use different types of pretraining data

Figure 1: **Cross-lingual transfer settings.** Monolingual and parallel data can be used at different stages of the pipeline, either directly or indirectly through MT (b), but the traditional categorization falls short at capturing them (a).

During pretraining, parallel data has been incorporated through an auxiliary objective, such as Translation Language Modeling (TLM) in XLM (Conneau and Lample, 2019) or bitext denoising in PARADISE (Reid and Artetxe, 2022). Regarding finetuning, it is common to use Machine Translation (MT)—which is trained on parallel data under the hood—to translate the downstream training data into the target language(s) (Conneau et al., 2020), which can be seen as a form of data augmentation.

Nevertheless, it is still unclear **why** parallel data is beneficial for cross-lingual transfer learning. Is the **data itself** that matters, given the additional information that it contains? Or is it the explicit **modeling** of parallel interactions that is important? To answer this question, we systematically compare the use of parallel data from different sources:

[*] Work done while at the University of Tokyo
[†] Work done while at Meta AI

Findings of the Association for Computational Linguistics: ACL 2023, pages 5999–6006
July 9-14, 2023 ©2023 Association for Computational Linguistics

ground truth parallel data, or synthetic parallel data generated by either supervised MT, unsupervised MT, or word-by-word translation. Most notably, our unsupervised MT variant relies on the exact same monolingual corpus used to pretrain the model, so any potential improvement can only come from the modeling side.

Our experiments on Natural Language Inference (NLI), Question Answering (QA) and Named Entity Recognition (NER) show that the explicit modeling of parallel interactions is indeed beneficial, demonstrating that existing pretraining and finetuning methods do not exploit the full potential of monolingual data. However, our best results are obtained using real parallel data—either directly or indirectly through supervised MT—showing that there is also some inherent value on it.

In the light of these results, we argue that the traditional categorization of cross-lingual transfer approaches into *zero-shot*, *translate-train* and *translate-test* (Figure 1a) falls short at capturing the required detail for a fair comparison across different approaches. Given this, we encourage further research on understanding what the contribution of monolingual and parallel data is, and how to best leverage them (directly or indirectly through MT, and at different parts of the pipeline), which requires thinking beyond the boundaries of the existing categorization (Figure 1b).

2 Experimental setup

2.1 Tasks

We run experiments on 3 tasks: NLI on XNLI (Conneau et al., 2018), extractive QA on XQuAD (Artetxe et al., 2020), and NER on WikiANN (Pan et al., 2017). In all cases, we use the original training set in English, and evaluate transfer performance in other languages. Due to compute constraints, we restrict evaluation to the following set of languages: English (en), Arabic (ar), German (de), Hindi (hi), French (fr), Swahili (sw), Russian (ru), Thai (th) and Vietnamese (vi).

Our finetuning incorporation experiments in §3.2 involve machine translating the training data into the target languages. For XNLI, we just translate the premise and hypothesis and leave the label unchanged. For XQuAD and WikiANN, which have token-level labels (as opposed to sequence-level), we translate the input text and project the answer spans by using the `awesome` (Dou and Neubig, 2021) word aligner , taking the aligned spans as the target labels.

2.2 Model

We use XLM-R base (Conneau et al., 2020) for all of our experiments, which was trained through Masked Language Modeling (MLM) on CC-100 (a monolingual corpus covering 100 languages). For finetuning, we experiment with learning rates of 1e-5, 5e-5, and 1e-4 using the Adam optimizer. We train for up to 10 epochs and choose the checkpoint with the best validation performance averaged across the languages in consideration.

2.3 Parallel data sources

We compare the following sources of parallel data in our experiments:

Gold. Ground-truth parallel data generated by humans. We use the same parallel data as Reid and Artetxe (2022), which combines data from IWSLT, WMT, and other parallel corpora.

Supervised MT. Synthetic parallel data generated through a conventional MT system. The MT system is supervised, so this approach is also leveraging ground-truth parallel data indirectly. We use the 420M M2M-100 model (Fan et al., 2020).

Unsupervised MT. Synthetic parallel data generated through an unsupervised MT system (Artetxe et al., 2018; Conneau and Lample, 2019). The MT system is trained on a subset of the monolingual data used for pretraining, so this approach does not use any additional data neither directly nor indirectly, other than the synthetically generated one. More concretely, we use XLM-R base to initialize our unsupervised MT model, and finetune it in 16 directions (en↔{ar,de,hi,fr,sw,ru,th,vi} using the iterative denoising autoencoding and backtranslation approach proposed by Conneau and Lample (2019).[1] We train for a total of 750k iterations using a batch size of 128k tokens. We use 200MB of text from CC100 for each language, amounting to a total of 1.8GB of training data.

Dictionary. Synthetic parallel data generated through random word replacement with a dictionary. We use the same dictionaries as Reid and Artetxe (2022), which combine dictionaries from MUSE (Lample et al., 2018) and those extracted using word aligners (Östling and Tiedemann, 2016). Following Reid and Artetxe (2022), we replace

[1] `https://github.com/facebookresearch/XLM`

	XNLI (acc)							XQuAD (F1)							WikiANN (F1)								
	en	ar	de	hi	fr	sw	avg	en	ar	hi	ru	th	vi	avg	en	ar	fr	hi	ru	th	vi	sw	avg
1) XLM-R	83.9	71.9	75.2	69.1	77.4	62.2	73.3	86.5	68.6	76.7	80.1	74.2	79.1	77.5	81.3	53.0	80.5	73.0	69.1	1.3	79.4	70.5	63.5
2) *+ unsup MT*	83.4	72.4	77.1	72.2	78.2	67.8	75.2	**86.7**	70.2	80.7	81.5	75.8	79.6	79.0	81.3	54.1	82.1	74.9	71.1	3.8	80.7	71.7	64.9
3) *+ sup MT*	83.2	74.4	77.5	**72.7**	78.3	70.1	76.0	86.6	**73.5**	81.1	**83.0**	77.4	**81.9**	**80.7**	81.6	57.0	82.3	75.4	71.6	**5.8**	**81.6**	73.4	66.1
4) *+ gold*	**84.0**	**75.2**	**77.7**	72.4	**78.6**	**70.4**	**76.4**	86.3	72.3	**82.3**	82.7	**78.2**	**81.9**	80.6	**82.4**	**57.3**	**82.4**	**75.6**	**71.8**	4.6	81.5	**73.7**	**66.2**

Table 1: **Pretraining incorporation results.** We compare the original XLM-R model (1) with three variants where we continue pretraining it on either synthetic (2, 3) or real (4) parallel data. All models are finetuned on English downstream data and zero-shot transferred to the target language.

	XNLI (acc)							XQuAD (F1)							WikiANN (F1)								
	en	ar	de	hi	fr	sw	avg	en	ar	hi	ru	th	vi	avg	en	ar	fr	hi	ru	th	vi	sw	avg
1) XLM-R	83.9	71.9	75.2	69.1	77.4	62.2	73.3	**86.5**	68.6	76.7	80.1	74.2	79.1	77.5	81.3	53.0	80.5	73.0	69.1	1.3	79.4	70.5	63.5
2) *+ dict*	83.7	72.6	77.6	70.7	78.9	65.6	74.9	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3) *+ unsup MT*	84.0	73.2	77.1	71.6	78.6	67.9	75.4	86.0	70.4	80.3	81.0	76.3	79.8	78.9	80.6	56.0	82.7	75.7	71.8	3.7	80.9	72.3	65.5
4) *+ sup MT*	**84.2**	**74.6**	**78.2**	**73.1**	**79.4**	**70.6**	**76.7**	86.3	**73.2**	**81.6**	**83.4**	**77.2**	**81.4**	**80.5**	**82.2**	**57.4**	**83.1**	**76.4**	**72.4**	**5.2**	**82.1**	**73.4**	**66.6**

Table 2: **Finetuning incorporation results.** We compare finetuning XLM-R on the original English data (1), and machine translated data through either word-by-word replacement (2), unsupervised MT (3) or supervised MT (4).

words that are included in our dictionary with a probability of 0.4.

3 Experiments and results

3.1 Pretraining incorporation

In these experiments, we incorporate parallel data into the pretraining process. We take XLM-R as our starting point, which was trained on monolingual data through MLM, and continue pretraining it on both MLM and TLM for 70k steps, using a batch size of 64k tokens. We use a learning rate of 5e-5 with a linear warmup and cosine decay schedule. We use the MLM objective 70% of the time, and the TLM objective 30% of the time. The latter applies the same masking objective over concatenated parallel sentences, and we compare different sources of parallel data as detailed in §2.3. For parallel data generated through MT, we translate a random subset of CC100 (keeping consistent with the data used in pretraining). The model is then finetuned on the downstream tasks using the original training data in English, and zero-shot transferred to the target languages.

We report our results in Table 1. We find that all variants incorporating parallel data outperform the original XLM-R model,[2] and the improvements are consistent across all target languages. However, different from Reid and Artetxe (2022), we do not find any clear improvements on English. Regarding the source of parallel data, we find that supervised MT performs at par with gold data, even for less-resourced languages for which MT tends to suffer. Unsupervised MT lags behind them, but consistently outperforms the baseline.

These results suggests that the mere facilitation of parallel interaction is helpful even when not using any new data, but incorporating ground-truth parallel data brings further improvements. However, the way in which parallel data is incorporated—either directly or through MT—does not have any clear impact, as evidenced by the similar performance of supervised MT and gold.

3.2 Finetuning incorporation

In these experiments, we incorporate parallel data into the finetuning process. We translate the downstream training data in English into the rest of languages, and finetune XLM-R in the combined data in all languages. This is commonly referred to as *translate-train-all* in the literature.

We report our results in Table 2. Similar to the finetuning incorporation, we find that incorporating parallel data outperforms the baseline in all tasks and target languages for all data sources that we explore. Supervised MT obtains the best results, followed by unsupervised MT and word-by-word translation with dictionaries. Similar to the pretraining incorporation results, this suggests

[2] The skeptical reader might attribute this improvement to the additional training steps we perform, irrespective of the use of parallel data. However, we find strong evidence that the improvements are brought by the use of parallel data given that (i) XLM-R was trained until convergence using a huge amount of compute, and our continued training represents an insignificant fraction on top (96 GPU days, compared to 13k GPU days, or a relative 0.7% further), and (ii) we get improvements in all target languages but not in English, suggesting that the additional steps improve the cross-lingual capabilities of the model but not its general quality.

that synthetic parallel data can bring improvements even when generated exclusively through monolingual data, but using real parallel data brings further improvements. Finally, we find that even simplistic ways to incorporate parallel signals can bring improvements, as evidenced by the dictionary replacement results.

3.3 Discussion

While prior work has reported strong results from incorporating parallel data for cross-lingual transfer learning, our results show that this improvement can partly—but not exclusively—be attributed to the explicit use of a parallel training signal, which can also be achieved through unsupervised MT without the need for any real parallel data. In fact, we find that the facilitation of parallel interactions is more important than the use of real parallel data in all tasks but XQuAD, where the latter has a larger impact. Despite the popularity of multilingual pretrained models, which predominantly rely on monolingual data both for pretraining and finetuning, this calls into question the extent to which existing approaches are able to exploit the full potential of such monolingual data. In addition, it is striking that we obtain similar results for both pretraining and finetuning incorporation, as well as supervised MT and gold standard parallel data. While further evidence is necessary to draw a more definitive conclusion, this suggests that parallel data brings similar improvements regardless of when (pretraining vs. finetuning) and how (directly vs. indirectly through MT) it is incorporated.

4 Reconsidering the categorization of cross-lingual learning approaches

As illustrated in Figure 1a, approaches to cross-lingual learning have traditionally been classified into 3 categories: *zero-shot* (finetune a multilingual model on English and zero-shot transfer into the target language), *translate-train* (translate the English training data into the target languages through MT and finetune a multilingual model), and *translate-test* (translate the test set into English and run inference using a monolingual model). This distinction is primarily based on which stage of the pipeline MT is incorporated into. While relevant from a practical perspective, we believe that, if taken in a rigid manner, such a framework can hinder addressing the more fundamental question of what the contribution of each data source is, and how to best leverage each of them.

More concretely, as shown in Figure 1b, there are different data **types** that one can use (monolingual source corpora, monolingual target corpora and parallel corpora, in addition to downstream data), which can be incorporated at different **stages** of the pipeline (pretraining, finetuning, testing) and via different **procedures** (directly or indirectly through MT). We argue that research in cross-lingual learning should aim to understand how the variants in each dimension as well the interactions between them impact downstream performance, which can require thinking beyond the boundaries of the 3 conventional categories. For instance, our variant using unsupervised MT to translate the downstream training data would fall within the definition of *translate-train*. However, this approach is more comparable to *zero-shot* in that it only uses monolingual data, and it would be unfair to compare it to conventional *translate-train* systems that rely on parallel data to train the MT system.

5 Related work

Prior work has explored the extent to which monolingual pretraining relies on knowledge transfer from unlabeled corpora by using synthetic data (Chiang and Lee, 2020; Krishna et al., 2021) or downstream data (Krishna et al., 2022) instead, and similar ideas have also been explored in computer vision (Kataoka et al., 2020; Asano et al., 2020). However, to the best of our knowledge, we are first to examine if cross-lingual learning also relies on knowledge transfer from parallel data. Our use of synthetic parallel corpora is also connected with back-translation, which is widely used in MT (Sennrich et al., 2016). However, conventional MT systems are trained on parallel data, and back-translation is usually motivated as a way to leverage additional (monolingual) data. In contrast, our unsupervised MT variant does not use any additional data compared to regular pretraining.

6 Conclusions

In this work, we show that even model-generated parallel data can be useful for cross-lingual learning—greatly expanding the possibilities for multilingual models to improve their performance by taking advantage of their own machine translation capabilities. Given this, we advocate for investigating the optimal way to leverage monolingual and/or parallel data for cross-lingual learning, which might require thinking beyond the boundaries of the conventional *zero-shot*, *translate-train* and *translate-test* categories.

7 Limitations

In this work, we only consider the pre-train then fine-tune paradigm which assumes that model weights are tuned for adaptation to specific tasks. Future work, once more capable multilingual LLMs are released, may also consider the few shot, and in-context learning-based setups to accommodate for more recent approaches towards adaptation in NLP. Future work may also consider setups more relevant to different, more diverse tasks (e.g. including webtext).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. 2020. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*.

Cheng-Han Chiang and Hung-yi Lee. 2020. Pre-training a language model without human language.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. 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.

Mihir Kale, Aditya Siddhant, Noah Constant, Melvin Johnson, Rami Al-Rfou, and Linting Xue. 2021. nmt5 – is parallel data still relevant for pre-training massively multilingual language models? *arXiv preprint arXiv:2106.02171*.

Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. 2020. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision*.

Kundan Krishna, Jeffrey Bigham, and Zachary C. Lipton. 2021. Does pretraining for summarization require knowledge transfer? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3178–3189, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kundan Krishna, Saurabh Garg, Jeffrey P. Bigham, and Zachary C. Lipton. 2022. Downstream datasets make surprisingly good pretraining corpora.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. 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.

Machel Reid and Mikel Artetxe. 2022. PARADISE: Exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810, Seattle, United States. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
Our work is primarily an analysis and does not entail any clear risk

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

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

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ 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.?
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.
Not applicable. Left blank.

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?
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?
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?
Not applicable. Left blank.

☑ 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?
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.?
Not applicable. Left blank.

□ 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)?
Not applicable. Left blank.

□ 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?
Not applicable. Left blank.