# Massively Multilingual Lexical Specialization of Multilingual Transformers

**Tommaso Green**[1], **Simone Paolo Ponzetto**[1], **Goran Glavaš**[2]
[1] Data and Web Science Group, University of Mannheim, Germany
[2] CAIDAS, University of Würzburg, Germany
{tommaso.green, ponzetto}@uni-mannheim.de
goran.glavas@uni-wuerzburg.de

## Abstract

While pretrained language models (PLMs) primarily serve as general-purpose text encoders that can be fine-tuned for a wide variety of downstream tasks, recent work has shown that they can also be rewired to produce high-quality word representations (i.e., static word embeddings) and yield good performance in type-level lexical tasks. While existing work primarily focused on the lexical specialization of single monolingual PLMs, in this work we expose massively multilingual transformers (MMTs, e.g., mBERT or XLM-R) to multilingual lexical knowledge at scale, leveraging BabelNet as the readily available rich source of multilingual and cross-lingual type-level lexical knowledge. Concretely, we use BabelNet's multilingual synsets to create synonym pairs (or synonym-gloss pairs) across 50 languages and then subject the MMTs (mBERT and XLM-R) to a lexical specialization procedure guided by a contrastive objective. We show that such multilingual lexical specialization brings substantial gains in two standard cross-lingual lexical tasks, bilingual lexicon induction and cross-lingual word similarity, as well as in cross-lingual sentence retrieval. Crucially, we observe gains for languages unseen in specialization, indicating that multilingual lexical specialization enables generalization to languages with no lexical constraints. In a series of controlled experiments, we show that the number of specialization constraints plays a much greater role than the set of languages from which they originate.

## 1 Introduction

Massively multilingual transfomers (MMTs) such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), among others, have been the primary vehicle of cross-lingual NLP transfer, offering state-of-the-art performance for many tasks and target languages in various zero-shot and few-shot transfer scenarios (Pires et al., 2019; Wu and Dredze, 2019; Cao et al., 2020; Artetxe et al., 2020; Lauscher et al., 2020a; Zhao et al., 2021; Ruder et al., 2021, *inter alia*). Much less work, however, investigated their capabilities as multilingual type-level word encoders (Vulić et al., 2020b). Recent work, focusing primarily on monolingual PLMs, demonstrated that they can be turned into effective type-level lexical encoders using lexical constraints (Vulić et al., 2021; Liu et al., 2021), i.e., a process commonly referred to as *lexical specialization*. Existing work, however, investigated lexical specialization in monolingual or bilingual settings only, namely specializing PLMs for a single language or a pair of languages using either English monolingual or noisy translated bilingual lexical constraints (Vulić et al., 2021). This is not only computationally inefficient, since one specialization needs to be executed for each language or language pair, but also does not tap into the wealth of multilingual knowledge of MMT's simultaneous pretraining on many (100+) languages, as well as large amounts of manually curated knowledge from massively multilingual knowledge bases like BabelNet (Navigli and Ponzetto, 2010).

In this work, in contrast, we investigate massively multilingual lexical specialization of MMTs, i.e., the potential benefits and limitations of a *single lexical specialization procedure in multiple languages*. To this end, we tap into BabelNet as the readily available massively multilingual lexico-semantic resource. Concretely, we release a dataset of synonym pairs or synonym-gloss pairs that cover 50 languages (representing 14 different language families) obtained from BabelNet's multilingual synsets and leverage them as positive instances in a contrastive specialization training procedure. Our evaluation on two multilingual lexical tasks – bilingual lexicon induction (BLI) and cross-lingual semantic word similarity (XLSIM) – as well as on the task of cross-lingual sentence retrieval demonstrate the effectiveness of the multilingual lexical special-



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 7700–7715

ization when compared against vanilla MMTs.

We complement our evaluation with diagnostic experiments aimed at studying properties of the multilingual lexical constraints that might drive the downstream lexical performance of the specialized models, namely the choice of the languages and the number of constraints. For this, we perform experiments where we control MMT specialization for (i) the linguistic diversity of the language represented in the specialization dataset and (ii) the size of the specialization dataset, i.e., the number of synonym pairs from BabelNet used in contrastive training. The results of our diagnostic experiments suggest that, counterintuitively, the typological diversity of the languages used in specialization (i.e., the specialization languages) has barely any effect in defining the downstream performance. These findings for multilingual specialization for (type-level) lexical tasks contrast the observations for higher-level tasks, requiring the modeling of sentence or sentence-pair semantics, in which both multi-source specialization/fine-tuning on diverse languages (Chen et al., 2019; Ansell et al., 2021) and linguistic proximity between training and evaluation languages (Lin et al., 2019; Lauscher et al., 2020a) have been shown to strongly affect the transfer performance. At the same time, we find that the alignment performance quickly saturates with few constraints: this corroborates the *rewiring hypothesis* of Vulić et al. (2021), here in a massively multilingual setting. To encourage further research on this topic, we release our code and all our datasets of lexical constraints and synonym-gloss pairs at `https://github.com/umanlp/babelbert`.

## 2 Related Work

External lexical knowledge from lexico-semantic resources (e.g., WordNet, ConceptNet) has been extensively leveraged for improving distributional representations of words – a process commonly referred to as semantic specialization. Earlier work on semantic specialization of word embeddings can roughly be divided into (i) *joint specialization* approaches (Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014; Kiela et al., 2015; Liu et al., 2015; Ono et al., 2015, inter alia), which integrate the external lexical constraints in the word embedding (pre)training and (ii) *retrofitting* or *postprocessing* techniques that post-hoc modify the pretrained embeddings, conforming them to external lexical constraints (Faruqui et al., 2015; Wieting et al., 2015; Mrkšić et al., 2017; Vulić et al., 2018; Glavaš and Vulić, 2018; Ponti et al., 2018).

External lexical knowledge has also been used to enrich monolingual PLMs by coupling masked language modeling with an auxiliary pretraining objective using lexical relations from WordNet (Lauscher et al., 2020b; Levine et al., 2020). However, these approaches are all aimed at enriching the Transformer with additional knowledge to be exploited in various downstream tasks, rather than producing better type-level word representations.

Several paradigms for obtaining semantically improved multilingual representation spaces have been proposed for static word embeddings: (i) *align-and-specialize* (Mrkšić et al., 2017) starts from mutually unaligned monolingual embedding spaces and both aligns them and semantically specializes for semantic similarity (as opposed to other types of semantic relatedness) by means of both multilingual (i.e., monolingual constraints for multiple languages) and cross-lingual lexical constraints; (ii) in *cross-lingual specialization transfer* (Vulić et al., 2018; Glavaš and Vulić, 2018) the embedding space of a target language is first projected into the (unspecialized) space of the high-resource language (typically English) using a limited amount of cross-lingual source-target lexical alignments and then specialized with the specialization model trained using abundant monolingual lexical constraints of the resource-rich source language; (iii) *constraint transfer* (Ponti et al., 2019) noisily translates the (abundant) source language constraints into the target languages and uses those to specialize the monolingual embedding spaces of the target languages.

More recently, Vulić et al. (2020b) showed that pretrained transformers encode a wealth of lexical knowledge in their parameters and that it is possible to obtain from them type-level (i.e., static, out-of-context) word representations that outperform representations obtained with word embedding models like fastText (Bojanowski et al., 2017). In a subsequent approach dubbed LexFit (Vulić et al., 2021), they specialize monolingual BERT models for a range of languages using English monolingual constraints from WordNet and Roget's Thesaurus, and automatically translated constraints (Ponti et al., 2019) for languages other than English. Finally, once they obtain static word representations for each language, they induce a bilingual space in a standard fashion, by means of an orthogonal pro-

jection (Smith et al., 2017; Artetxe et al., 2019). The effectiveness of LexFit stems from the fact that language-specific PLMs tend to produce better representations for a language than the MMTs (e.g., mBERT) due to the "curse of multilinguality" (Conneau et al., 2020; Pfeiffer et al., 2022). This, however, limits its applicability to a few languages with existing monolingual PLMs, crucially excluding low-resource ones. The approach is also computationally expensive as it entails (i) a separate specialization process for each language (plus a noisy translation of English constraints to the target language), and (ii) a bilingual alignment for each language pair. Additionally, it does not exploit cross-lingual lexical constraints in specialization, merely in post-hoc alignment. In this work, we propose a single massively multilingual lexical specialization approach (dubbed BabelBERT) to be applied to pretrained MMTs. To this end, we propose to obtain the lexical constraints from BabelNet, a massively multilingual lexico-semantic resource (Navigli and Ponzetto, 2010; Navigli et al., 2021). BabelBERT has several advantages: (i) we leverage multilingual (i.e., monolingual from multiple languages) as well as *cross-lingual* lexical constraints and (ii) perform a single multilingual specialization procedure, instead of specializing separately for each language or language pair; (iii) the multilingual specialization on top of MMTs removes the need for language-specific transformers and additionally allows for specialization effects to propagate to unseen languages, i.e., languages without readily available lexical constraints.

## 3 Multilingual Lexical Specialization

We first describe how we create lexical constraints from BabelNet in §3.1 and then how we leverage them in a contrastive learning procedure in §3.2.

### 3.1 Constraint Mining

In line with most work on lexical specialization for semantic similarity, we exploit the lexico-semantic relation of synonymy to obtain the specialization constraints. The multilingual synsets of BabelNet[1] allow us to simultaneously obtain both monolingual and cross-lingual synonym pairs. Let $L = \{L_1, \ldots, L_n\}$ be the predefined set of languages for which we mine the constraints and $W = \{w_1, \ldots, w_N\}$ a list of seed words. For each word, we first obtain all BabelNet synsets containing that word, discarding synsets that represent named entities and keeping only synsets that have at least two glosses in languages from $L$. We then iterate over all the fetched synsets, creating all possible (monolingual and cross-lingual) synonym pairs ($w_1$, $w_2$), each associated with a language pair ($L_1^{word}$, $L_2^{word}$). We also extract glosses ($g_1$, $g_2$), i.e., sentences that explain the concept of the synset, in languages ($L_1^{gloss}$, $L_2^{gloss}$) different from $L_1^{word}$ and $L_2^{word}$, respectively.[2] To ensure the quality of the words we extract, we only keep a word if it lies in the top-$k$ words in its language frequency list.[3] We additionally discard words that are automatic translations or Wikipedia redirections, remove multi-words and delete duplicates.

### 3.2 Specializing for Semantic Similarity

Our lexico-semantic specialization procedure is based on a Bi-Encoder architecture (often also referred to as Dual-Encoder or Siamese architecture) and a contrastive training objective.

**Type-Level Word Representations.** Following related work (Vulić et al., 2020b; Bommasani et al., 2020; Vulić et al., 2021), we obtain type-level word representations from a PLM independently for each word from the synonym pair: we tokenize each word into its constituent subwords $sw_1 \ldots sw_m$ and feed the sequence $[SPEC_1][sw_1] \ldots [sw_m][SPEC_2]$ into the MMT, with $[SPEC_1]$ and $[SPEC_2]$ denoting the special sequence start and sequence end tokens of the MMT (e.g., $[CLS]$ and $[SEP]$, respectively, for BERT). We get the final representation $\mathbf{e}_w^{type}$ by mean pooling the representations of its subwords (without the special tokens) from the last layer of the Transformer encoder.

**Sense-Level Word Representations.** Type-level representations of polysemous words conflate, by construction, all of its senses into one embedding. To address this and make the representations capture a specific sense, we leverage additional context through sense-level information provided by the knowledge base. For example, among the senses of the word bat found in BabelNet, the list of the most frequent senses contains the synset for the

[1] We use version 5.0, which covers 500 languages, under the non-commercial license (https://babelnet.org/full-license)

[2] $L_1^{word}$ may, however, be the same as $L_2^{gloss}$ and so may $L_2^{word}$ be the same as $L_1^{gloss}$.

[3] We use the *wordfreq* library (Speer, 2022) and fastText vocabularies for the languages that *wordfreq* does not cover.

nocturnal animal and the sports club (e.g., the one used for example in cricket). Besides the set of synonyms, i.e., the different synsets these senses belong to, their different meaning is captured by the glosses. The gloss for the animal sense reads as follows:

> Nocturnal mouselike mammal with forelimbs modified to form membranous wings [...]

whereas for the sport stick we find:

> A cricket bat is a specialised piece of equipment used by batters in the sport of cricket [...]

In light of this, we make the MMT additionally encode word-gloss pairs, in order to obtain sense-level representations: to this end, we append one of the mined glosses $g$ to the word as follows:

$$[SPEC_1][sw_1]\dots[sw_m][SPEC_2]\, g\, [SPEC_2]$$

and feed this as input to the MMT. As with the type-level representations, we get the final representation $\mathbf{e}_w^{sense}$ by mean pooling the representations of *only* the subwords $sw_1 \dots sw_m$ (i.e., gloss is there just for the contextualization of the word) from the last layer of the Transformer encoder.

**Contrastive Objective.** We train in batches $\mathcal{B} = (w_1^{(i)}, w_2^{(i)}, syn_{id}^{(i)})_{i=1}^{N_B}$ of synonym pairs, with $syn_{id}^{(i)}$ denoting the BabelNet synset of the synonym pair $(w_1^{(i)}, w_2^{(i)})$. In sense-level training, each data point additionally contains the glosses $(g_1^{(i)}, g_2^{(i)})$. We train by minimizing a variant of the popular InfoNCE contrastive loss (van den Oord et al., 2018). In a single batch, there might be more than one pair belonging to the same synset, we thus form all possible *ordered* [4] positive pairs in a set $\mathcal{P}$, i.e., pairs of words with the same $syn_{id}$.

$$\mathcal{L}^{\mathcal{B}}_{\text{InfoNCE}} = \frac{-1}{|\mathcal{P}|} \sum_{(w_1^{(i)}, w_2^{(j)}) \in \mathcal{P}} \log\left(sim(\mathbf{e}_{w_1}^{(i)}, \mathbf{e}_{w_2}^{(j)})\right)$$

$$- \log\left(sim(\mathbf{e}_{w_1}^{(i)}, \mathbf{e}_{w_2}^{(j)}) + \sum_{n \in \mathcal{N}^{(i)}} sim(\mathbf{e}_{w_1}^{(i)}, \mathbf{e}^{(n)})\right)$$

where $sim(\mathbf{e}_{w_1}^{(i)}, \mathbf{e}_{w_2}^{(j)}) = \exp(\cos(\mathbf{e}_{w_1}^{(i)}, \mathbf{e}_{w_2}^{(j)}))/\tau$, with $\tau$ as the temperature hyperparameter and $\mathcal{N}^{(i)}$ as the set of in-batch negatives, i.e., words from the other pairs in the batch that come from a BabelNet synset other than $syn_{id}^{(i)}$.

**Adapter-Based Fine-Tuning.** Besides full fine-tuning of the MMT's parameters, we experiment with lexical specialization via adapter-based fine-tuning (Houlsby et al., 2019). Adapters, shown useful in various sequential and transfer learning scenarios (Pfeiffer et al., 2020b; Rücklé et al., 2020; Lauscher et al., 2021; Hung et al., 2022) are parameter-light modules that are inserted into a PLM's layers before specialization (i.e., fine-tuning): during specialization, only adapter parameters are tuned, while the PLM's pretrained parameters are kept fixed. We adopt the architecture of Pfeiffer et al. (2020b), in which one bottleneck adapter is inserted into each Transformer layer.

## 4 Experimental setup

**Multilingual Lexical Constraints.** We focus on the 50 diverse languages from the popular XTREME-R benchmark (Ruder et al., 2021) which we report with language codes for brevity: *af*, *ar*, *az*, *bg*, *bn*, *de*, *el*, *en*, *es*, *et*, *eu*, *fa*, *fi*, *fr*, *gu*, *he*, *hi*, *ht*, *hu*, *id*, *it*, *ja*, *jv*, *ka*, *kk*, *ko*, *lt*, *ml*, *mr*, *ms*, *my*, *nl*, *pa*, *pl*, *pt*, *qu*, *ro*, *ru*, *sw*, *ta*, *te*, *th*, *tl*, *tr*, *uk*, *ur*, *vi*, *wo*, *yo*, *zh*. The sample covers 14 language families (Afro-Asiatic, Austro-Asiatic, Austronesian, Dravidian, Indo- European, Japonic, Kartvelian, Kra-Dai, Niger- Congo, Sino-Tibetan, Turkic, Uralic, Creole, and Quechuan) and additionally contains Basque and Korean as two language isolates.

We collect the constraints from BabelNet with the procedure described in §3.1. As seed words, we select the top-$N$ English most frequent words ($N = 1,000$, filtering for stopwords using NLTK (Bird and Loper, 2004)) and retain only words that belong to the top-$k$ ($k = 15,000$) words in the frequency list of a language. The total training set consists of 761,273 lexical constraints: we provide additional statistics of the dataset in appendix A and a few examples in Table 3.

**Evaluation Tasks.** We evaluate on two standard cross-lingual word-level tasks, bilingual lexicon induction (BLI) and cross-lingual word similarity (XLSIM). We couple this with an evaluation on unsupervised cross-lingual sentence retrieval. For the word-level tasks (XLSIM and BLI): a) we make sure not to include in the training set of lexical constraints from BabelNet any word that appears

[4] This is because the last $sim(\cdot)$ term depends on which word occupies the first position.

in the test sets; b) we take the mean-pooling of the embeddings of the subwords from the best-performing layer (see Table 4) of the MMT as word representations.

**Task 1: Bilingual Lexicon Induction.** BLI tests the quality of a multilingual (bilingual) representation space by means of type-level word alignment between languages. For a given query word from a source language, words from the vocabulary of a target language are ranked based on their similarity with the query. The position of the correct translation of the query in the target language ranking reflects the quality of the type-level word alignment between the languages. We evaluate on two well-established benchmarks: G-BLI (Glavaš et al., 2019) covers 28 language pairs between 8 languages (*de*, *en*, *fi*, *fr*, *hr*, *it*, *ru*, *tr*), whereas PanlexBLI (Vulić et al., 2019) spans 15 diverse languages (*bg*, *ca*, *eo*, *et*, *eu*, *fi*, *he*, *hu*, *id*, *ka*, *ko*, *lt*, *no*, *th*, *tr*) for a total of 210 language pairs. For both datasets, we evaluate on the test portions consisting of 2,000 word pairs per language pair and report the performance in terms of Mean Reciprocal Rank (MRR) as recommended by Glavaš et al. (2019). Following Vulić et al. (2020b), the vocabularies used for retrieval cover the top 100K most frequent words from the respective fastText Wikipedia vectors (Bojanowski et al., 2017).

**Task 2: Cross-Lingual Word Similarity.** XL-SIM measures the correlation between the similarities of cross-lingual word pairs obtained based on their representations in the multilingual (bilingual) representation space and similarity scores assigned by human annotators. We evaluate the performance on 66 language pairs between 12 languages (*zh*, *cy*, *en*, *et*, *fi*, *fr*, *he*, *pl*, *ru*, *es*, *sw*, *yue*) from the MultiSimLex dataset (Vulić et al., 2020a) and use Spearman's $\rho$ between the cosine similarities between words' embeddings and the corresponding human-assigned similarity scores.

**Task 3: Cross-Lingual Sentence Retrieval.** For cross-lingual sentence retrieval, we use the Tatoeba dataset (Artetxe and Schwenk, 2019) which comprises 112 languages, where each language has 1,000 sentences paired with their translations in English. We obtain the sentence embedding by mean-pooling the representations of all of its subword tokens at the output of the best-performing Transformer layer. We straightforwardly compute the similarities between sentences as the cosine of the angle between their embeddings. We compare each query sentence to its nearest neighbour and compute accuracy as our evaluation measure.

**Training Details.** We experiment using two different MMTs: multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-R (Conneau et al., 2020).[5] In all experiments, we set the temperature of the InfoNCE loss to $\tau = 0.07$. To account for the very skewed distribution of constraints across language pairs (cf. Figure 3), following Conneau and Lample (2019), we sample batch constraints from the multinomial distribution $\{q_{i,j}\}$ over language pairs $(L_i, L_j)$ as follows:

$$q_{i,j} = \frac{p_{i,j}^{\alpha}}{\sum_{k,l} p_{k,l}^{\alpha}}, \quad p_{i,j} = \frac{n_{i,j}}{\sum_{k,l} n_{k,l}}$$

where $n_{i,j}$ denotes the number of synonym pairs for a language pair $(L_i, L_j)$ and $\alpha$ is the smoothing factor. We set $\alpha$ to 0.5.

For model selection (both hyperparameter search and checkpoint selection), we proceed as follows. We randomly select two language pairs from G-BLI and two language pairs from PL-BLI and pick the corresponding training sets – consisting of 5,000 pairs – as our validation set. Before training, we run one validation loop to get the MRR score of the unspecialized vanilla MMT for each of the language pairs. During training, we stop every quarter of an epoch to perform validation: we track for each of these four validation language pairs the relative improvement of MRR w.r.t. the vanilla score. We use the average of these four relative improvements as our overall validation metric that guides model selection. We train for 15 epochs using AdamW (Loshchilov and Hutter, 2019) and use PytorchLightning (Falcon and The PyTorch Lightning team, 2019) for our implementation, coupled with Huggingface Transformers (Wolf et al., 2020) and Pytorch Metric Learning (Musgrave et al., 2020) libraries. For the adapter-based models, we use the adapter-transformers library (Pfeiffer et al., 2020a).

**Hyperparameters and training details** For the fully fine-tuned models, we search for the optimal learning rate $lr \in \{2e-5, 5e-6, 1e-6\}$ and the batch size $N_B \in \{32, 64\}$. For adapter-based models, we additionally try $lr = 1e-4$ and set the adapter reduction ratio to 16 (i.e., we set the bottleneck size of the adapters to 48). Every experiment

[5] We use `bert-base-multilingual-uncased` and `xlm-roberta-base` from HuggingFace (Wolf et al., 2020).

| MMT | Model | BLI | | XLSIM | Ttb |
|---|---|---|---|---|---|
| | | G-BLI | PL-BLI | | |
| mBERT | vanilla | 14.5 | 10.7 | 10.3 | 34.1 |
| | Babel-Ad | 20.0 | 12.3 | 25.4 | 43.2 |
| | Babel-FT | **20.9** | **12.4** | **25.8** | **43.7** |
| | Babel-Gl | 19.5 | 11.8 | 24.1 | 41.7 |
| XLM-R | vanilla | 8.5 | 5.4 | 5.9 | 37.6 |
| | Babel-Ad | 17.8 | 8.7 | 32.0 | 55.6 |
| | Babel-FT | **20.8** | **10.5** | 34.2 | 55.7 |
| | Babel-Gl | 19.7 | 10.0 | **34.4** | **58.6** |

Table 1: Results of multilingual lexical specialization for two MMTs – mBERT and XLM-R on three tasks (four datasets): BLI – on G-BLI and PanLex-BLI (PL-BLI), XLSIM, and cross-lingual sentence retrieval on the Tatoeba dataset (Ttb). Performance reported in terms of MRR ($\times 100$) for BLI, Spearman's $\rho$ for XLSIM and accuracy ($\times 100$) for Ttb. For each dataset, we report averages across all language pairs (28 for G-BLI, 210 for PL-BLI, 66 for XLSIM, and 112 for Ttb). The highest scores per column and MMT are in **bold**.

is done on a single NVIDIA V100 or A100 GPU on a computing cluster at our disposal. Taking into account failed experiments, grid searches and successful runs we report 331 days of compute (including CPU time for preprocessing and retrieval) as logged by the Weights & Biases logger (Biewald, 2020). We provide the full list of hyperparameter values in the Appendix (Table 4).

## 5 Results and Discussion

We present our main results in Table 1, where we compare the multilingual lexical specialization procedure with type-level representations using full fine-tuning (Babel-FT) and adapter-based tuning (Babel-Ad) and full-fine tuning of the MMT using sense-level representations (Babel-Gl) against the baseline models provided by the unspecialized vanilla MMTs. We show the results for the word representations that come from the layer for which the best average performance is obtained on the given dataset: we provide the information on optimal layers for lexical representations in Table 4 in the Appendix. For each model, we compute each language-pair score as the average over 3 runs with different random seeds.

Overall, the results indicate that multilingual lexical fine-tuning improves the performance of both MMTs (mBERT and XLM-R) on all tasks. All three specialization variants (Babel-Ad/FT/Gl) yield similar performance on all three tasks for mBERT. The same is, however, not the case for XLM-R, where full fine-tuning (Babel-FT and Babel-Gl) leads to substantially better performance across the board compared to adapter-based training (Babel-Ad). Training with sense-level information in the form of synset glosses (Babel-Gl) seems particularly beneficial for cross-lingual sentence retrieval on Tatoeba (Ttb) – we assume that this is because, much like the rest of the sentences in Tatoeba, the glosses in specialization training provide sentential context to word representations. Interestingly, despite the fact that vanilla mBERT produces substantially higher quality word representations than vanilla XLM-R (e.g., 14.5 vs. 8.5 on G-BLI or 10.3 vs. 5.9 on XLSIM), our multilingual lexical specialization tilts the results in favour of XLM-R, with comparable BLI results between the two and much better XLM-R performance on XLSIM (9-point gap) and Ttb (15-point gap). This suggests that XLM-R actually contains richer multilingual lexical knowledge than mBERT, which is, however, buried deeper in the model (this is corroborated by the fact that for XLM-R we generally get better lexical representations from lower Transformer layers, see the Appendix): once uncovered by means of our multilingual lexical specialization, this richer lexical information of XLM-R surfaces.

Figure 1 shows the five language pairs for each evaluation task/dataset for which we observe the largest performance improvement with Babel-FT mBERT. In PL-BLI, we find two languages, Norwegian and Catalan, for which no constraint was seen during training: this indicates that the benefits of massively multilingual lexical specialization propagate to unseen languages. What we find particularly encouraging is the fact that Tatoeba pairs with the largest gains include some low-resource languages and dialects (e.g., Wu and Yu Chinese, Tamil, and Interlingue), some of them not even present in MMTs in pretraining.

### 5.1 Additional Analysis

We perform additional ablations to try to isolate the factors that lead to performance gains from massively multilingual lexical specialization. To this end, we carry out experiments in which we vary (i) the typological diversity of the sample of languages from which we take BabelNet constraints and (ii) the size of the training set of lexical constraints

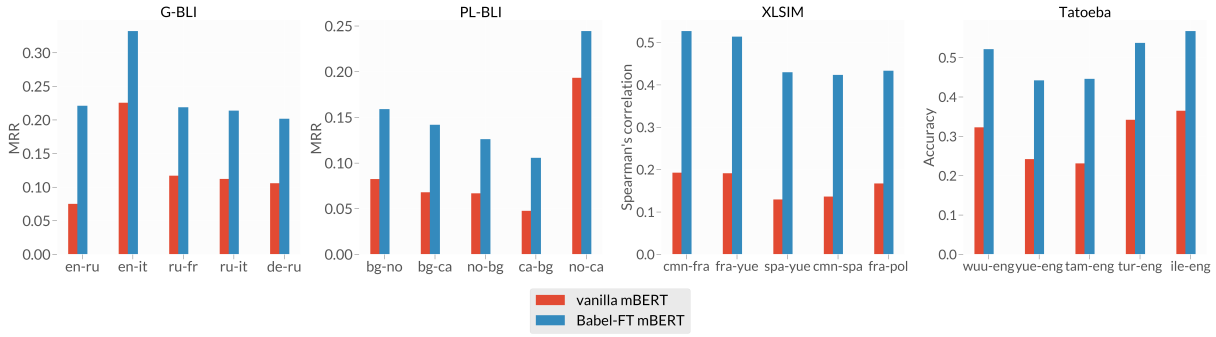


Figure 1: The five language pairs for each evaluation task/dataset for which we observe the largest performance improvement with Babel-FT mBERT. Each language pair score is obtained using the best layer for the specific dataset-model combination as reported in Table 4.

| Language sample | $d_{\text{typ}}$ | $sim_{train-test}$ | PL-BLI Avg |
|---|---|---|---|
| {af, de, el, en, es, fr, nl, pt, ro, uk} | 0.1830 | 0.7369 | 0.1154 |
| {af, en, hi, it, ms, nl, pt, ro, ru, uk} | 0.2406 | 0.7275 | 0.1176 |
| {af, de, en, fr, it, pl, pt, ta, uk, ur} | 0.2661 | 0.7113 | 0.1164 |
| {bn, de, en, es, hi, it, jv, ro, ru, uk} | 0.2906 | 0.7114 | 0.1159 |
| {en, it, mr, ms, pl, pt, ru, ta, uk, zh} | 0.3180 | 0.6953 | 0.1166 |
| {az, de, en, fa, fr, it, pt, sw, ta, uk} | 0.3576 | 0.6764 | 0.1172 |
| {af, de, el, fr, jv, ml, mr, qu, ta, uk} | 0.3746 | 0.6520 | 0.1176 |
| {bn, el, fr, hi, jv, mr, ms, qu, ro, tl} | 0.4022 | 0.6530 | 0.1175 |
| {bn, gu, it, my, pt, ro, sw, ta, vi, zh} | 0.4231 | 0.6345 | 0.1162 |
| {ar, de, el, en, ja, jv, kk, my, sw, ur} | 0.4600 | 0.6207 | 0.1171 |

Table 2: Selection of 10 samples of 10 languages each in increasing $d_{\text{typ}}$ order, together with the similarity between training and test languages and average MRR on PL-BLI.

used for specialization.

**Role of Linguistic Diversity.** We investigate how changing typological diversity of the sample of languages from which we draw the specialization constraints affects the generalization to unseen languages. That is, we test whether a selection of languages with different degrees of linguistic diversity and no overlap with the test languages impacts the multilingual lexical specialization of the MMTs. To quantify linguistic diversity, we borrow the typological diversity index $d_{\text{typ}}$ from Ponti et al. (2020). For a sample of languages $S$, we compute the index based on the URIEL vectors (Littell et al., 2017) of languages in the sample.[6] We obtain $d_{\text{typ}}$ of the sample $S$ by computing an entropy value for each of the features across all languages in $S$: such value is 0 if all languages have identical values for that feature. We then average the entropy scores across all features.

We choose PL-BLI as our benchmark for this analysis as it proved to be the most challenging lexical task. We first create 10 samples, each containing 10 languages, as follows: we first sample 1 million different language samples of 10 languages, making sure none of them contains any of the PL-BLI test languages. We then divide them into 10 bins according to $d_{\text{typ}}$ and randomly pick one sample from each bin at random. For each sample, we mine the synonym pairs from BabelNet from scratch, considering only those languages, and making sure that each language pair has exactly 100 constraints – resulting in a total of 5,500 instances for each sample (counting both cross-lingual and monolingual constraints). In an effort to limit test language leakage in the training procedure, differently from the main experiments, we only validate on two language pairs from G-BLI for model selection. We conduct experiments with Babel-FT mBERT as our best-performing model on PL-BLI. We fine-tune for 10 epochs with the same hyperparameters used in the main experiments but without the upsampling as all language pairs have the same number of constraints. We report the results, together with the language samples and their $d_{\text{typ}}$ score in Table 2. Surprisingly, we observe a poor correlation between $d_{\text{typ}}$ of the language sample and the corresponding PL-BLI performance.

[6]We use the `syntax_knn` vectors which contain 103 manually-coded features indicating various syntactic properties of languages.

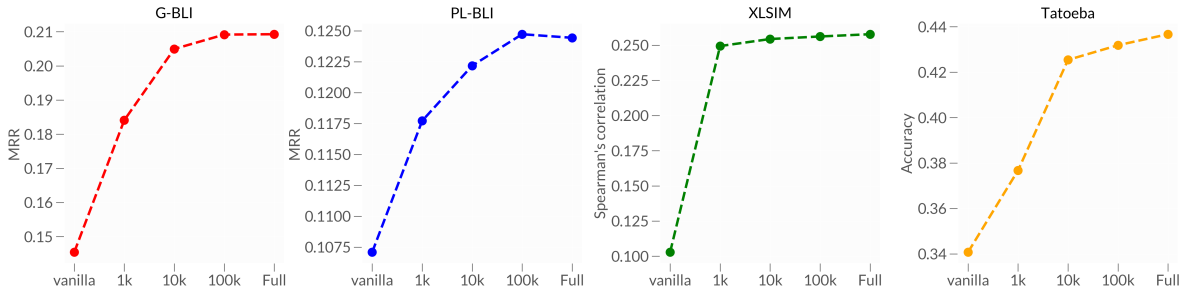


Figure 2: Performance on the four datasets plotted against the size of the specialization dataset: vanilla refers to the unspecialized MMT and Full to the full training set size (761,273).

We additionally quantify the degree of similarity between the languages of each sample and the languages included in PL-BLI: $sim_{train-test}$ (also shown in Table 2) is the average of pairwise similarities of URIEL vectors between the languages of the two sets. We again observe a poor correlation between $sim_{train-test}$ and PL-BLI performance, indicating that the proximity of training and testing languages does not play a role. While counterintuitive, this is actually favourable: it indicates that we can, with similar specialization success, leverage constraints from a wide range of languages, which allows us to mine them for high-resource languages for which structured lexical knowledge is more abundant.

**The Role of Constraint Size.** We perform additional experiments to investigate the effect of training set size (i.e., the total number of constraints). For this, we create three additional training sets of sizes 1K, 10K, and 100K instances, with the same relative distribution of constraints across language pairs as in the full training. We then subject mBERT to Babel-FT specialization for 10 epochs (same hyperparameters and training procedure as in the main evaluation). For each training set, we execute 3 different runs and report averages in Figure 2. On all four benchmarks, we observe the same overall behaviour: the performance saturates already with 10K constraints. Tatoeba, the only sentence-level task in our setup, seems to be the task that benefits the most from a larger training set. In contrast, XLSIM performance saturates already with 1K lexical constraints, which seems to be in line with findings of Vulić et al. (2021).

Vulić et al. (2021) proposed and empirically validated the *rewiring hypothesis* – that lexical specialization primarily exposes knowledge that is already present in the pre-trained weights, rather than injecting new knowledge. We believe our results confirm this claim: if specialization mainly resurfaces lexical knowledge hidden in the weights, this puts a cap on the downstream performance. In our case, we find that such knowledge seems to be independent of both the typological diversity of the training languages and the typological similarity of training and test languages: in this sense, the MMT appears to be learning a language-agnostic lexical alignment function that affects its entire representation space. Moreover, learning of this function only seems to require about 10K samples, with performance quickly saturating with more constraints.

## 6 Conclusion

In this work, we presented a multilingual lexical specialization approach that leverages the massively multilingual lexical knowledge available in BabelNet. Differently from prior approaches, which perform monolingual or bilingual specialization procedures, we subject MMTs to a single training regime with lexical constraints from 50 languages and report substantial gains over the unspecialized baseline (i.e., MMT not subject to lexical specialization). We perform both type-level lexical specialization, i.e. with words fed in isolation to the transformer, and sense-level lexical specialization, by accompanying each word with a gloss. We perform a series of additional experiments to study the driving factors of lexical specialization: in one experiment, we keep the training set size fixed while diversifying the languages in the training set. We observe that this does not seem to have a significant impact on the overall performance. In a subsequent experiment, we again use constraints encompassing 50 languages, but limit the number of constraints per language pair: we find that more constraints help the model perform better, however, few samples are necessary to reach peak performance. Our results support the rewiring hypothesis

of Vulić et al. (2021) that lexical specialization resurfaces the existing lexical knowledge stored in MMTs, rather than injecting it. Such extraction seems to be independent of the training languages and quickly saturates with few constraints.

**Limitations.** While we try to perform multilingual lexical specialization on a set of typologically diverse languages, we are still restricting our analysis to a small fraction of all the languages of the world. In addition to this, our analysis investigates only two MMTs – albeit arguably the two most widely used. Due to hardware limitations, we experimented with XLM-R Base: the results we report may be substantially different for XLM-R Large (or other larger MMTs like mT5), which possibly encodes more lexical knowledge.

**Ethical considerations.** We leverage lexical constraints from BabelNet, a resource constructed semi-automatically. BabelNet may contain lexical associations reflecting negative social biases (e.g., sexism or racism). Biased constraints, when used as training data in our specialization, may strengthen societal biases present in MMTs.

**Acknowledgments** Tommaso Green and Simone Ponzetto have been supported by the JOIN-T 2 project of the Deutsche Forschungsgemeinschaft (DFG). Goran Glavaš has been supported by the EUINACTION grant funded by NORFACE Governance (462-19-010) through Deutsche Forschungsgemeinschaft (DFG; GL950/2-1). We additionally acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. We thank our colleague Sotaro Takeshita for insightful discussions during the development of this project and Ines Rehbein for her comments on a draft of this paper.

## References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

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.

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

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *Machine Learning and Knowledge Discovery in Databases*, pages 132–148, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. 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 and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45, Melbourne, Australia. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Chia-Chien Hung, Anne Lauscher, Simone Ponzetto, and Goran Glavaš. 2022. DS-TOD: Efficient domain specialization for task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904, Dublin, Ireland. Association for Computational Linguistics.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048, Lisbon, Portugal. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020a. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020b. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1501–1511, Beijing, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. Pytorch metric learning.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of babelnet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989, Denver, Colorado. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. 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.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293, Brussels, Belgium. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Cross-lingual semantic specialization via lexical relation induction. 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 2206–2217, Hong Kong, China. Association for Computational Linguistics.

Andreas Rücklé, Jonas Pfeiffer, and Iryna Gurevych. 2020. MultiCQA: Zero-shot transfer of self-supervised text matching models on a massive scale. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2471–2486, Online. Association for Computational Linguistics.

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.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word

vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020a. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. 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 516–527, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? 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 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021. LexFit: Lexical fine-tuning of pretrained language models. 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 5269–5283, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

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

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 1219–1228, New York, NY, USA. Association for Computing Machinery.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550, Baltimore, Maryland. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. 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 5751–5767, Online. Association for Computational Linguistics.

## A Training set statistics

We report the number of constraints for each language pair in Figure 3.

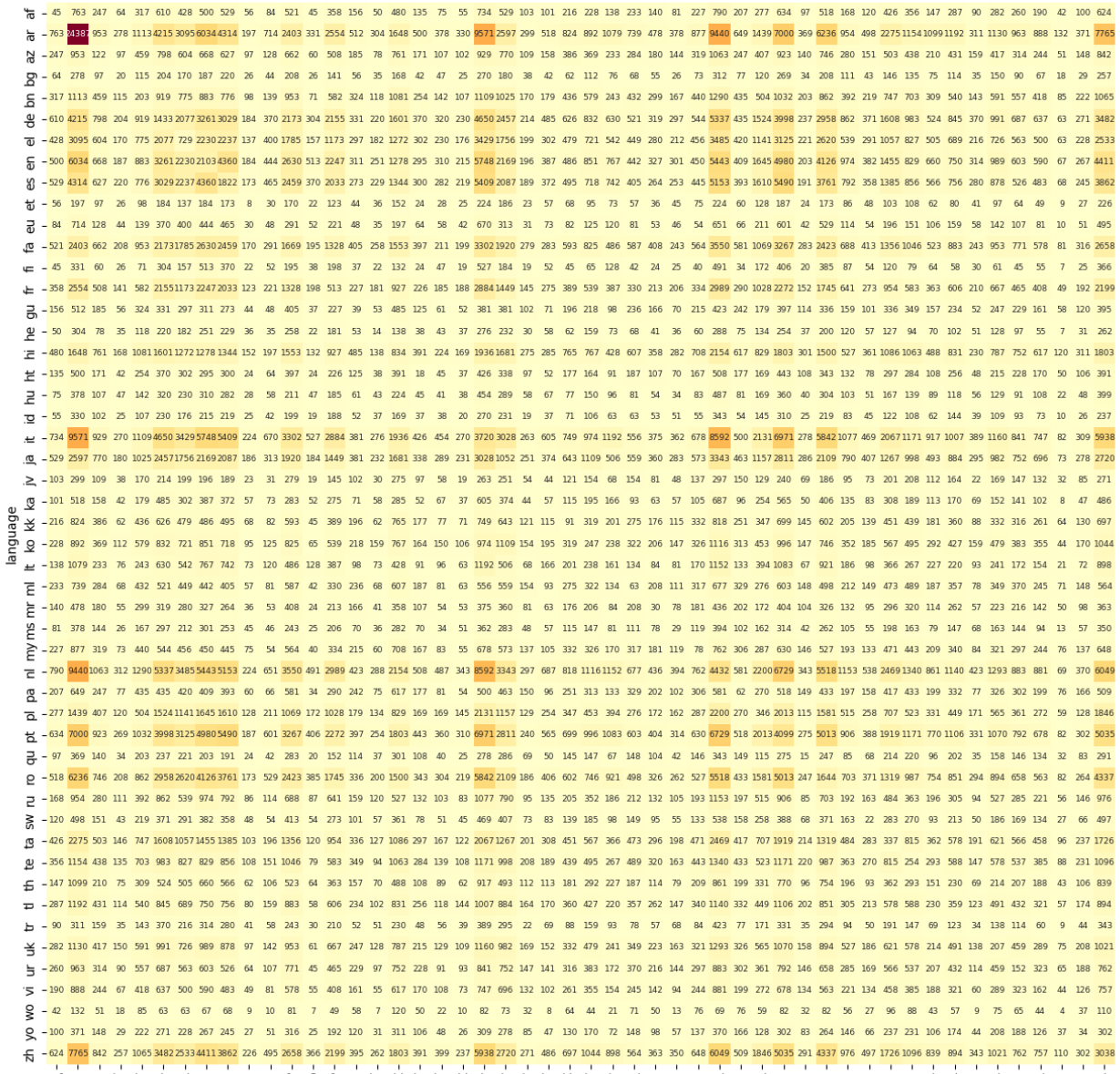


Figure 3: The numbers of (monolingual and cross-lingual) synonymy constraints in the training set for the selection of 50 languages from the XTREME-R benchmark.

| $w_1$ | $w_2$ | $g_1$ | $g_2$ | $syn_{id}$ |
|---|---|---|---|---|
| production [en] | produit [fr] | Нещо, произведено от човешка или механична дейност или при естествен процес. [bg] | The amount of an artifact that has been created by someone or some process. [en] | bn:00064584n |
| passato [it] | gestern [de] | Минало е период от време, поредица от събития, които вече са се случили. [bg] | Iragana edo lehena gauden unearen auretik gertatu den oro dugu. [eu] | bn:00060927n |
| lajan [ht] | centen [nl] | Το χρήμα είναι οποιοδήποτε στοιχείο ή επαληθεύσιμη εγγραφή που είναι γενικά αποδεκτό ως πληρωμή για αγαθά [...] [el] | Montant non spécifié d'une devise. [fr] | bn:00055644n |
| lungsod [tl] | oraș [ro] | A large, busy city, especially as the main city in an area or country or as distinguished from surrounding rural areas. [en] | Una metropoli è una città di grandi dimensioni con più di 1 milione di abitanti con un'area comunale [...] [it] | bn:00019319n |

Table 3: Examples of synonym pairs together with their glosses and synset identifiers.

| MMT | Model | $lr$ | $N_B$ | G-BLI Layer | PL-BLI Layer | XLSIM Layer | Ttb Layer |
|---|---|---|---|---|---|---|---|
| mBERT | vanilla | - | - | 6 | 6 | 6 | 8 |
| | Babel-Ad | $5e-6$ | 32 | 12 | 12 | 12 | 11 |
| | Babel-FT | $1e-6$ | 64 | 12 | 12 | 12 | 11 |
| | Babel-FT-Gl | $1e-6$ | 64 | 12 | 12 | 12 | 11 |
| XLM-R | vanilla | - | - | 1 | 1 | 1 | 7 |
| | Babel-Ad | $1e-4$ | 64 | 8 | 8 | 12 | 10 |
| | Babel-FT | $1e-6$ | 32 | 11 | 12 | 12 | 11 |
| | Babel-FT-Gl | $5e-6$ | 32 | 9 | 8 | 12 | 10 |

Table 4: The best-found hyperparameters of the models together with the index of the best-performing layer for each dataset.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Limitation section*

☑ A2. Did you discuss any potential risks of your work?
*Yes, ethical considerations*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract and introduction (section 1)*

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

### B ☑ Did you use or create scientific artifacts?

*Yes, we release a dataset (3.1) and use publicly released datasets/models (sec. 4)*

☑ B1. Did you cite the creators of artifacts you used?
*Section 4.*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*We use Babelnet under a non-commercial license (footnote 1)*

☑ 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)?
*We use BabelNet as in the license and release a dataset with the intention of fostering research on lexical specialization*

☑ 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?
*We remove named entities from the resulting dataset (section 3)*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Appendix A*

☒ 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.
*Appendix A, Sec.4*

### C ☑ Did you run computational experiments?

*Sec 4*

☑ 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 4*

*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 4, Appendix B*

☑ 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?
*Sec 5*

☒ 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.)?
*Left blank.*

**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.*