Probing Cross-Lingual Lexical Knowledge from Multilingual Sentence Encoders

Ivan Vulić[1] **Goran Glavaš**[2] **Fangyu Liu**[1]
Nigel Collier[1] **Edoardo Maria Ponti**[3,1] **Anna Korhonen**[1]
[1]Language Technology Lab, TAL, University of Cambridge
[2]CAIDAS, University of Würzburg
[3]EdinburghNLP, University of Edinburgh
{iv250, alk23}@cam.ac.uk

Abstract

Pretrained multilingual language models (LMs) can be successfully transformed into multilingual sentence encoders (SEs; e.g., LaBSE, xMPNET) via additional fine-tuning or model distillation with parallel data. However, it remains unclear how to best leverage them to represent sub-sentence *lexical* items (i.e., words and phrases) in cross-lingual lexical tasks. In this work, we *probe* SEs for the amount of *cross-lingual lexical knowledge* stored in their parameters, and compare them against the original multilingual LMs. We also devise a simple yet efficient method for *exposing* the cross-lingual lexical knowledge by means of additional fine-tuning through inexpensive contrastive learning that requires only a small amount of word translation pairs. Using bilingual lexical induction (BLI), cross-lingual lexical semantic similarity, and cross-lingual entity linking as lexical probing tasks, we report substantial gains on standard benchmarks (e.g., +10 Precision@1 points in BLI). The results indicate that the SEs such as LaBSE can be 'rewired' into effective cross-lingual lexical encoders via the contrastive learning procedure, and that it is possible to expose more cross-lingual lexical knowledge compared to using them as off-the-shelf SEs. This way, we also provide an effective tool for harnessing 'covert' multilingual lexical knowledge hidden in multilingual sentence encoders.

1 Introduction

Transfer learning with pretrained Language Models (LMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) offers unmatched performance in many NLP tasks (Wang et al., 2019; Raffel et al., 2020). However, despite the wealth of semantic knowledge stored in the pretrained LMs (Rogers et al., 2020; Vulić et al., 2020b), they do not produce coherent and effective sentence representations when used off-the-shelf (Liu et al., 2021c). To this effect, further specialization for sentence-level semantics – not unlike the standard task fine-tuning – is needed (Reimers and Gurevych, 2019; Li et al., 2020; Yan et al., 2021, *inter alia*). LMs get *transformed* into sentence encoders (SEs) via dual-encoder frameworks that leverage contrastive learning objectives (van den Oord et al., 2018; Musgrave et al., 2020), in supervised (i.e., leveraging labeled external data such as NLI or sentence similarity annotations) (Reimers and Gurevych, 2019; Vulić et al., 2021b; Liu et al., 2021a) or, more recently, fully unsupervised fine-tuning (Liu et al., 2021c; Gao et al., 2021) setups.

Following the procedures from monolingual setups, another line of research has been transforming multilingual LMs into *multilingual SEs* (Feng et al., 2022; Reimers and Gurevych, 2020), which enable effective sentence matching and ranking in multiple languages as well as cross-lingually (Litschko et al., 2022). The transformation is typically done by coupling **1)** LM objectives on monolingual data available in multiple languages with **2)** cross-lingual objectives such as Translation Language Modeling (TLM) (Conneau and Lample, 2019) and/or cross-lingual contrastive ranking (Yang et al., 2020). Such multilingual SEs consume a large number of parallel sentences for the latter objectives. Consequently, they outperform multilingual off-the-shelf LMs in cross-lingual sentence similarity and ranking applications (Liu et al., 2021d; Litschko et al., 2022). However, as we show in this work, such multilingual SEs may still lag behind traditional *static* cross-lingual word embeddings (CLWEs) when encoding sub-sentence *lexical items* (e.g., words or phrases) (Liu et al., 2021c) for cross-lingual lexical tasks (e.g., BLI).

In this work, we *probe* multilingual SEs for *cross-lingual lexical knowledge*, relying on standard semantic similarity tasks in cross-lingual setups as our '*lexical probes*' (Vulić et al., 2020b). We demonstrate that, due to their fine-tuning on multilingual and parallel data, they indeed store a

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2089–2105
May 2-6, 2023 ©2023 Association for Computational Linguistics

wealth of such knowledge, much more than what 'meets the eye' when they are used 'off the shelf'. However, this lexical knowledge needs to be *exposed* from the original multilingual SEs, (again) through additional fine-tuning. In other words, we show that multilingual SEs can be 'rewired' into effective cross-lingual *lexical encoders*, as illustrated in Figure 1. This rewiring is again done via a quick and inexpensive contrastive learning procedure: with merely 1k-5k word translation pairs, we successfully convert multilingual SEs into state-of-the-art bilingual lexical encoders for any language pair (present in a specific dataset).[1]

We probe the original LMs and SEs as well as demonstrate the usefulness of the proposed contrastive procedure for 'exposing' cross-lingual lexical knowledge on three standard lexical cross-lingual tasks using standard evaluation data and protocols: BLI, cross-lingual lexical semantic similarity (XLSIM), and cross-lingual entity linking (XL-EL). We show that the 'exposure' procedure is highly effective for both vanilla multilingual LMs (mBERT and XLM-R) and multilingual SEs (LaBSE and xMPNET): e.g., we observe ≈+10 Precision@1 points gains on standard BLI benchmarks (Glavaš et al., 2019). Multilingual SEs offer substantially better cross-lingual lexical performance than vanilla LMs, both before and after being subjected to contrastive cross-lingual lexical fine-tuning (see Figure 1). This indicates that it is possible to expose more cross-lingual lexical knowledge from multilingual SEs than from their vanilla LM counterparts, likely owing to their additional exposure to parallel data.

Finally, inspired by Li et al. (2022), we validate that word vectors produced by cross-lingual lexical encoders (i.e., after the contrastive cross-lingual lexical 'exposure') can be effectively interpolated with static CLWEs (Artetxe et al., 2018) and offer even stronger performance in cross-lingual lexical tasks. Encouragingly, our cross-lingual lexical specialization of multilingual SEs (as well as the further interpolation with static CLWEs), yields particularly massive performance gains for pairs of low-resource languages, as demonstrated on the low-resource BLI benchmark (Vulić et al., 2019).

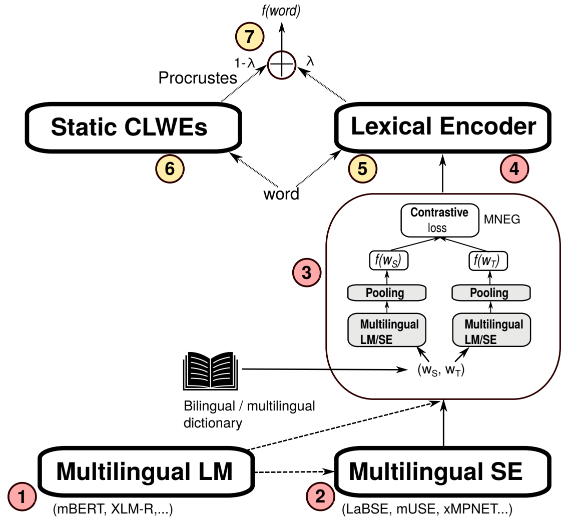


Figure 1: Illustration of the pipeline of exposing cross-lingual lexical knowledge from multilingual language models (LMs) and sentence encoders (SEs) (§2). Multilingual LMs (①) can be transformed into multilingual SEs (②) as done in previous work (Reimers and Gurevych, 2020; Feng et al., 2022). A contrastive cross-lingual lexical fine-tuning procedure (③) (requiring an external bilingual dictionary) can be applied on both ① and ②, yielding a fine-tuned cross-lingual lexical encoder (④). At inference, a word/phrase is encoded by the lexical encoder (⑤). In addition, its encoding can be interpolated with the corresponding static (cross-lingual) word embedding (⑥), producing the final embedding of the word/phrase (⑦). Before the interpolation, static CLWEs must be mapped into the vector space of the lexical encoder (④): to this end, we learn the standard orthogonal (Procrustes) projection matrix.

2 From Multilingual Sentence Encoders to Cross-Lingual Lexical Encoders

Motivation. The motivation for this work largely stems from the research on *probing and analyzing* pretrained LMs for various types of knowledge they (implicitly) store in their parameters (Ethayarajh, 2019; Jawahar et al., 2019; Rogers et al., 2020). In this paper, we focus on a *particular knowledge type*: cross-lingual lexical knowledge, and its extraction from multilingual LMs and SEs. The work combines two research threads, being inspired by the work on probing monolingual PLMs for lexical knowledge (Vulić et al., 2020b), as well as on interpreting representations in multilingual PLMs (Bjerva et al., 2019; Libovický et al., 2020; Beinborn and Choenni, 2020; Deshpande et al., 2022; Chai et al., 2022, *inter alia*).

Previous work also tried to prompt multilingual LMs for word translations via masked natural lan-

[1] Note that this is a typical requirement of standard mapping-based approaches for learning static cross-lingual word embeddings which excel in the BLI task (Mikolov et al., 2013; Conneau et al., 2018; Glavaš and Vulić, 2020).

guage templates (Gonen et al., 2020) and extract type-level word embeddings from LMs (i) directly without context (Vulić et al., 2020a, 2021a) or (ii) by averaging contextual embeddings over a large auxiliary corpus in the target language (Bommasani et al., 2020; Litschko et al., 2022). This existing body of work **1)** demonstrated that even sophisticated templates and extraction strategies cannot outperform cross-lingual word embedding spaces (e.g., induced from monolingual fastText vectors) in cross-lingual lexical tasks such as BLI (Vulić et al., 2020b) and **2)** did not attempt to expose cross-lingual lexical knowledge from multilingual SEs and compare it against the (same type of) knowledge extracted from vanilla multilingual LMs.

Multilingual Sentence Encoders. Off-the-shelf LMs contextualize (sub)word representations, but are unable to encode the precise meaning of input text out of the box. SEs – LMs fine-tuned via sentence-level objectives – in contrast, directly produce a precise semantic encoding of input text. A large body of work focuses on inducing multilingual encoders that capture sentence meaning across languages (Artetxe and Schwenk, 2019; Feng et al., 2022; Yang et al., 2020, *inter alia*).

The most popular approach obtains multilingual SEs (Reimers and Gurevych, 2020) by distilling the knowledge from the monolingual English SE teacher (trained on English semantic similarity and NLI data) into multilingual LM student (e.g., mBERT), using parallel sentences to guide the distillation process. SEs, being specialized for sentence similarity, encode sentence meaning more accurately and are useful in various (unsupervised) text similarity and ranking tasks, monolingually and across languages (Artetxe and Schwenk, 2019).

While SEs' primary purpose is sentence encoding, they can, in principle, be applied to sub-sentential text: words and phrases. In this work, we show that multilingual SEs can be turned into effective cross-lingual lexical encoders. We achieve this through additional cross-lingual lexical fine-tuning (Vulić et al., 2021a), requiring as supervision only a small set of word translation pairs.

2.1 Cross-Lingual Lexical Fine-Tuning

For a given language pair L_s-L_t, our contrastive cross-lingual lexical fine-tuning of multilingual encoders (LMs and SEs alike) requires a dictionary spanning N (typically $N \leq 5{,}000$) word translation pairs, $\mathcal{D} = \{(w_{i,s}, v_{i,t})\}_{i=1}^{N}$.[2] We consider the translation pairs from $\mathcal{D}$ to be *positive examples* for the contrastive fine-tuning procedure. For each of the N source language words in the dictionary ($w_{i,s}$), we precompute a set of K hard negative samples: these are the L_t words that are the closest to $w_{i,s}$ in the representation space of the multilingual encoder, but not its direct translation $v_{i,t}$. For one-to-many and many-to-many seed dictionaries $\mathcal{D}$, the set K does not contain any L_t word paired with $w_{i,s}$. Let $f_\theta(\cdot)$, be the encoding function of the multilingual LM/SE, with θ as parameters, and let $S(\cdot,\cdot)$ be a function of similarity between two vectors. For a source word $w_{i,s}$, we select as hard negatives words v_t from L_t that have the highest $S(f_{\theta_0}(w_{i,s}),f_{\theta_0}(v_t))$ score, with θ_0 as the original encoder's parameters, before fine-tuning.

We encode all training words – those from the seed dictionary $\mathcal{D}$ and (at most) $N \cdot K$ precomputed hard L_t negatives – independently and *in isolation*. Concretely, for an input w with M subword tokens $[sw_1] \ldots [sw_M]$, we feed the sequence $[SPEC1][sw_1] \ldots [sw_M][SPEC2]$ into the multilingual encoder (with $[SPEC1]$ and $[SPEC2]$ as encoder's sequence start and end tokens, resp.), and take the average of the transformed representation (from the last Transformer layer) of the w's subword tokens as the w's encoding $f_\theta(w)$. Put simply, we process sub-sentential text input in the same way that multilingual SEs handle sentence-level input. We experimented with other encoding strategies from prior work, e.g., taking the representation of the sequence start token $[SPEC1]$ (Liu et al., 2021b; Li et al., 2022); in preliminary experiments, however, we obtained the best results by averaging subword representations.[3]

Following common practice in contrastive learning (Henderson et al., 2019; Vulić et al., 2021a), we define S as the scaled cosine similarity: $S(f_\theta(w_i),f_\theta(w_j)) = C \cdot cos(f_\theta(w_i),f_\theta(w_j))$, with C as the scaling constant. We then train in batches of B translation pairs, with the variant of the widely used multiple negatives ranking loss (MNEG) (Cer et al., 2018; Henderson et al., 2019, 2020) as the fine-tuning objective:

[2] Note that such bilingual dictionaries are one of the most widespread and cheapest-to-obtain resources in multilingual NLP (Ruder et al., 2019; Wang et al., 2022).

[3] Note that $[SPEC1]$ and $[SPEC2]$ are placeholders for the encoder's special tokens: e.g., in case of multilingual BERT [SPEC1] is the [CLS] token, while [SPEC2] is the [SEP] token.

$$
\begin{aligned}
\mathcal{L} = & -\sum_{i=1}^{B} S(f_\theta(w_i), f_\theta(v_i)) \qquad \text{(positives)} \\
& + \sum_{i=1}^{B} \log \sum_{j=1, j\neq i}^{B} e^{S(f_\theta(w_i), f_\theta(v_j))} \quad \text{(in-batch negatives)} \\
& + \sum_{i=1}^{B} \log \sum_{k=1}^{K} e^{S(f_\theta(w_i), f_\theta(v_{k,i}))} \qquad \text{(hard negatives)}
\end{aligned}
$$

where $v_{k,i}$ denotes the k-th hard negative from the language L_t for the L_s word w_i. MNEG combines the K hard negatives per each positive example with B-1 in-batch negatives (i.e., for a source language word $w_{i,s}$, each target language word $v_{j,t}$, $j \neq i$ from B is used as an in-batch negative of $w_{i,s}$). MNEG aims to reshape the representation space of the encoder by simultaneously (a) maximising the similarity for positive pairs – i.e., bringing closer together ('attracting') the words from the positive pairs and (b) minimising the similarity for (both in-batch and hard) negative pairs – i.e., pushing ('repelling') the words from negative pairs further away from each other).[4]

2.2 Interpolation with Static CLWEs

Li et al. (2022) recently showed that further performance benefits in the BLI task might be achieved by combining the type-level output of the encoding function *f* with static CLWEs, but they experimented only with multilingual LMs, and limited their analyses to the BLI task.

Static CLWEs and multilingual encoder-based representations of the same set of words can be seen as two different views of the same data point. Following Li et al. (2022), we learn an additional linear orthogonal mapping from the static cross-lingual WE space – e.g., a CLWE space induced from monolingual fastText embeddings (Bojanowski et al., 2017) using VECMAP (Artetxe et al., 2018) – into the cross-lingual space spanned by the multilingual encoder. The mapping transforms ℓ_2-normed d_1-dimensional static CLWEs into d_2-dimensional cross-lingual WEs obtained through the multilingual encoder (fine-tuned f_θ or original f_{θ_0}).

Learning the linear map $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$, when $d_1 < d_2$,[5] is formulated as a Generalized Procrustes problem (Schönemann, 1966; Viklands, 2006). It operates on all (i.e., both L_s and L_t) words from the seed translation dictionary $\mathcal{D}$. To learn the mapping $\boldsymbol{W}$, for pairs from $\mathcal{D}$ we decouple L_s words $w_{i,s}$ from their L_t translations $v_{i,t}$ to create vector pairs $(clwe(w_{i,s}), f_\theta(w_{i,s}))$ and $(clwe(v_{i,t}), f_\theta(v_{i,t}))$ – with $clwe(w)$ as the static CLWE of w (e.g., its VECMAP embedding), and $f_\theta(w)$ its encoder-based representation – based on which we learn of the orthogonal mapping $\boldsymbol{W}$ (the so-called Procrustes method gives a closed-form solution). Unless noted otherwise, a final representation of an input word w is then computed as:

$$
(1-\lambda)\frac{clwe(w)\boldsymbol{W}}{\|clwe(w)\boldsymbol{W}\|_2} + \lambda \frac{f_\theta(w)}{\|f_\theta(w)\|_2}, \qquad (1)
$$

where λ is a tunable interpolation hyper-parameter, $clwe(w)$ denotes the static CLWE of w, and $f_\theta(w)$ the representation of w obtained with the (contrastively fine-tuned or original) multilingual LM/SE. This simple procedure yields an 'interpolated' shared cross-lingual WE space.

3 Experimental Setup

Multilingual Sentence Encoders. We probe two widely used multilingual SEs: **1)** Language-agnostic BERT Sentence Embedding (**LABSE**) (Feng et al., 2022) which adapts pretrained multilingual BERT (**MBERT**) (Devlin et al., 2019) into a multilingual SE; **2)** Multilingual **XMPNET** is a distillation-based adaptation (Reimers and Gurevych, 2020) of **XLM-R** (Conneau et al., 2020) as the student model into a multilingual SE, based on the monolingual English MPNet encoder (Song et al., 2020) as the teacher model. LABSE is the current state-of-the-art multilingual SE and supports 109 languages, while XMPNET is the best-performing multilingual SE in the Sentence-BERT repository (Reimers and Gurevych, 2019): For further technical details regarding the models in our comparison, we refer to the original papers.

Along with LABSE and XMPNET as SEs, we experiment with the original multilingual LMs – mBERT and XLM-R – using the the same training and evaluation protocols (see Figure 1 and §2), aiming to quantify: (i) the extent to which cross-lingual lexical knowledge can be exposed from LMs that have *not* been specialized for sentence-level semantics, as well as (ii) the increase in quality of lexical knowledge brought about with sentence-level specialisation (i.e., when multilingual LMs get transformed into multilingual SEs).

[4] In practice, we rely on the implementation of the MNEG loss from the SBERT repository www.sbert.net (Reimers and Gurevych, 2019); the default value $C = 20$ is used.

[5] The assumption $d_1 < d_2$ typically holds as fastText WEs are 300-dimensional while the dimensionality of standard multilingual LMs and SEs is $d_2 = 768$ or $d_2 = 1,024$.

Evaluation Tasks. We evaluate on the standard and diverse cross-lingual lexical semantic tasks treated as 'cross-lingual lexical probes'. In other words, we fine-tune the models to steer them towards becoming better lexical encoders and then we check how well they fare across a set of representative (intrinsic) lexical tasks which could be seen as such 'lexical probes'.

Task 1: Bilingual Lexicon Induction (BLI), a standard task to assess the "semantic quality" of static cross-lingual word embeddings (CLWEs) (Ruder et al., 2019), allows us to **1)** directly assess the extent to which cross-lingual word translation knowledge can be exposed from multilingual LMs and SEs and **2)** immediately test the ability to transform multilingual sentence encoders into bilingual lexical encoders. We run a series of BLI evaluations on two standard BLI benchmarks. **1)** GT-BLI (Glavaš et al., 2019), constructed semi-automatically from Google Translate, comprises 28 language pairs with a good balance of typologically similar and distant languages (Croatian: HR, English: EN, Finnish: FI, French: FR, German: DE, Italian: IT, Russian: RU, Turkish: TR). **2)** PanLex-BLI (Vulić et al., 2019) focuses on BLI evaluation for lower-resource languages, deriving training and test data from PanLex (Kamholz et al., 2014). We evaluate on 10 pairs comprising the following five typologically and etymologically diverse languages: Bulgarian (BG), Catalan (CA), Estonian (ET), Hebrew (HE), and Georgian (KA).

Standard BLI setups and data are adopted: 5k training word pairs are used as seed dictionary $\mathcal{D}$, and another 2k pairs as test data. Note that $\mathcal{D}$ is used to (i) contrastively fine-tune multilingual encoders (§2.1), (ii) learn the (baseline) static VecMap CLWE space, as well as to (iii) learn the projection between the static CLWE space and the representation spaces of multilingual encoders required to obtain the interpolated representations (§2.2). The evaluation metric is standard Precision@1 (P@1).[6] For PanLex-BLI, we also run experiments using smaller $\mathcal{D}$, spanning 1k pairs.

Task 2: Cross-Lingual Lexical Semantic Similarity (XLSIM) tests the extent to which lexical representations can capture the (human perception of) fine-grained semantic similarity of words across languages. We use the comprehensive XLSIM benchmark Multi-SimLex (Vulić et al., 2020a), which comprises cross-lingual datasets of 2k-4k scored word and phrase pairs over 66 language pairs. We evaluate on a subset of language pairs shared with the GT-BLI dataset: EN, FI, RU, FR.

The evaluation metric is the standard Spearman's rank correlation between the average of gold human-elicited XLSIM scores for word pairs and the cosine similarity between their respective word representations. To avoid any test data leakage, we remove all XLSIM test pairs from the bilingual dictionary $\mathcal{D}$ prior to fine-tuning and CLWE mapping.

Task 3: Cross-Lingual Entity Linking (XEL) is a standard task in knowledge base (KB) construction (Zhou et al., 2022), where the goal is to link an entity mention in any language to a corresponding entity in an English KB or in a language-agnostic KB.[7] We evaluate on the cross-lingual biomedical entity linking (XL-BEL) benchmark of Liu et al. (2021d): it requires the model to link an entity mention to entries in UMLS (Bodenreider, 2004), a language-agnostic medical knowledge base. We largely follow the XL-BEL experimental setup of Liu et al. (2021d) and probe the encoders first *without* any additional task-specific fine-tuning on UMLS data, and then *with* subsequent UMLS fine-tuning (i) only on the EN UMLS data, (ii) on all the UMLS data in 10 languages of the XL-BEL dataset.[8] Due to a large number of experiments, we again focus on the subset of languages in XL-BEL shared with GT-BLI: EN, DE, FI, RU, TR.

Static CLWEs and Word Vocabularies. As monolingual static WEs, we select CommonCrawl fastText vectors (Bojanowski et al., 2017) of the top 200k most frequent word types in the training data, following prior work on learning static CLWEs (Conneau et al., 2018; Artetxe et al., 2018; Heyman et al., 2019).[9] Static CLWEs are then induced via the standard and popular supervised mapping-based VecMap method (Artetxe et al., 2018), leveraging the seed dictionary $\mathcal{D}$. These CLWEs are used for interpolation with encoder-based WEs (see §2.2) but also as the baseline approaches for BLI and XLSIM tasks. We compute the type-level WEs from multilingual LMs and SEs for the same

[6] We observed very similar performance trends for P@5 and Mean Reciprocal Rank (MRR) as BLI measures.

[7] Following prior work (Liu et al., 2021b; Zhou et al., 2022), XEL in this work also refers only to entity mention *disambiguation*; it does not cover the mention detection subtask.

[8] See (Liu et al., 2021d) for additional details.

[9] CommonCrawl-based fastText WEs typically outperform other popular choice for monolingual WEs: Wikipedia-based fastText (Glavaš et al., 2019; Li et al., 2022). We note that the main trends in our results also extend to the Wiki-based WEs.

200K most frequent words of each language.

Technical Details and Hyperparameters. The implementation is based on the SBERT framework (Reimers and Gurevych, 2019), using the suggested settings: AdamW (Loshchilov and Hutter, 2018); learning rate of $2e$-5; weight decay rate of 0.01. We run contrastive fine-tuning for 5 epochs with all the models, with the batch size of $B = 128$ positives for MNEG. The number of hard negatives per each positive is set to $K = 10$ (see §2.1).[10] Since standard BLI and XLSIM datasets lack a validation portion (Ruder et al., 2019), we follow prior work (Glavaš et al., 2019) and tune hyperparameters on a *single* language pair from each dataset, and use those values in all other runs. The randomly selected language pairs are EN-TR for GT-BLI and CA–ET for PanLex-BLI.

All reported scores are the averages over 5 runs with 5 fixed random seeds.

Model Configurations. They are labelled as follows: **ENC-{noCL,+CL}** (λ), where (i) ENC denotes the input multilingual Transformer, which can be a multilingual LM (MBERT, XLM-R), or a multilingual SE (LABSE, XMPNET), (ii) 'noCL' refers to using the input model 'off-the-shelf' without any contrastive lexical fine-tuning, while '+CL' variants apply the contrastive fine-tuning, and (iii) λ is the factor that defines the interpolation with the static CLWE space, obtained with VECMAP (see Figure 1 and §2.2). Note that $\lambda = 1.0$ implies no interpolation with static CLWE space, i.e., WEs come purely from the multilingual LM/SE.

Important Disclaimer. We note that the main purpose of the chosen evaluation tasks and experimental protocols is not necessarily achieving state-of-the-art performance, but rather probing different model variants in different cross-lingual lexical tasks, and offering fair and insightful comparisons.

4 Results and Discussion

Bilingual Lexicon Induction (BLI). Table 1 displays our main BLI results, aggregated over all 28 language pairs of GT-BLI. Two trends hold across the board. First, multilingual SEs, LABSE and XMPNET, substantially outperform their multilingual LM counterparts, MBERT and XLM-R. The gains are visible in all four experimental configurations (with/without contrastive cross-lingual lexical specialisation × with/without interpolation with the VECMAP CLWE space). This confirms our intuition that multilingual SEs, having been (additionally) trained on parallel data (Feng et al., 2022; Reimers and Gurevych, 2020), should better reflect the cross-lingual alignments at the lexical level than off-the-shelf multilingual LMs, which have not been exposed to any cross-lingual signal in pretraining. The poor cross-lingual lexical alignment in the representation spaces of MBERT and XLM-R also reflects in the fact that with those encoders, we only surpass the baseline VECMAP performance by a small margin (+1.1 for XLM-R, +1.6 for MBERT) after subjecting them to contrastive lexical fine-tuning *and* interpolating their word encodings with VECMAP WEs.

The behavior of SEs, on the other hand, is much more favorable. LABSE, for example, surpasses baseline VECMAP performance with interpolation alone, even without the contrastive lexical fine-tuning. When contrastively fine-tuned (and then interpolated with VECMAP) both LABSE and XMPNET surpass the baseline VECMAP performance by a much wider margin (+6.4 and +5.2, respectively). This implies that contrastive fine-tuning exposes more of the high quality cross-lingual lexical knowledge from multilingual SEs.

Both (i) contrastive cross-lingual lexical learning (+CL) and (ii) interpolation with VECMAP consistently improve the performance for all four encoders: we reach peak scores by combining contrastive fine-tuning and interpolation with static CLWEs (+CL (λ)). Contrastive fine-tuning crucially contributes to the overall performance: compared to interpolation alone (noCL (λ)), +CL (λ) brings an average gain of over 6 BLI points.

Table 2 shows the BLI results on 10 low(er)-resource language pairs from PanLex-BLI. While overall relative trends are similar to those observed for high(er)-resource languages from GT-BLI (Table 1), the gains stemming from cross-lingual contrastive lexical fine-tuning are substantially larger in this case. The best-performing configuration – contrastive fine-tuning and interpolation (+CL (0.4)) applied on LaBSE – surpasses VECMAP by 11 BLI points on average (compared to 6 points on GT-BLI), with gains for some language pairs (e.g., HE-KA, ET-HE) approaching the impressive margin of 20 BLI points. This finding indicates that cross-lingual lexical knowledge stored in multilingual SEs is even more crucial when dealing

[10] We also tested K=$\{20, 30, 50\}$. They slow down fine-tuning while yielding small-to-negligible performance gains.

Multilingual LMs		mBERT				XLM-R			
Config →	VecMap	noCL (1.0)	noCL (λ)	+CL (1.0)	+CL (λ)	noCL (1.0)	noCL (λ)	+CL (1.0)	+CL (λ)
[BLI] λ=0.3	42.7	9.0	39.2	22.3	44.3	6.4	33.7	21.2	43.8
[XLSIM] λ=0.5	45.8	5.7	35.4	38.4	48.1	1.7	23.5	46.1	51.8

Multilingual SEs		LaBSE				xMPNET			
Config →	VecMap	noCL (1.0)	noCL (λ)	+CL (1.0)	+CL (λ)	noCL (1.0)	noCL (λ)	+CL (1.0)	+CL (λ)
[BLI] λ=0.3	42.7	21.4	45.7	30.8	49.1	17.0	41.7	28.6	47.9
[XLSIM] λ=0.5	45.8	50.4	54.9	48.8	54.1	51.3	56.6	49.6	54.5

Table 1: (a) P@1 scores (×100%) averaged across all 28 language pairs in the GT-BLI dataset ([BLI] rows); (b) Spearman's ρ correlation scores (×100) averaged across a subset of 6 language pairs from Multi-SimLex ([XLSIM rows]). See §3 for the description of different model configurations/variants. $|\mathcal{D}| = 5k$. The number in the parentheses denotes the value for λ (see §3), which differs between the two tasks (0.3 for BLI and 0.5 for XLSIM). The λ value of 1.0 effectively means 'no interpolation' with static VecMap CLWEs. Individual results per each language pair in both tasks and with other λs are in Appendix B and Appendix C.

		mBERT		XLM-R		LaBSE		xMPNET	
Pair ↓ / Config →	VecMap	+CL (1.0)	+CL (0.4)	+CL (1.0)	+CL (0.4)	+CL (1.0)	+CL (0.4)	+CL (1.0)	+CL (0.4)
BG–CA	34.4	9.6	31.9	13.2	33.3	17.9	**38.0**	15.9	35.7
BG–ET	30.0	17.1	32.6	21.3	34.1	29.9	**42.7**	26.1	38.9
BG–HE	26.1	9.9	21.1	10.5	26.3	23.7	**37.2**	10.9	27.2
BG–KA	26.8	16.0	29.8	15.9	30.5	27.2	**37.4**	18.7	32.4
CA–ET	26.3	26.8	32.9	23.5	34.1	28.8	**38.6**	29.0	38.9
CA–HE	23.3	2.3	12.5	4.9	18.5	12.7	**28.9**	8.5	22.7
CA–KA	20.7	1.5	10.3	4.7	20.0	9.6	**26.1**	6.4	21.8
ET–HE	18.6	15.0	21.9	17.7	26.0	31.0	**37.8**	18.5	27.0
ET–KA	16.5	7.2	18.2	12.7	24.3	19.3	**30.3**	12.5	25.8
HE–KA	12.7	15.6	23.8	13.3	23.1	25.3	**30.2**	15.1	24.4
Average	23.5	12.1	23.5	13.8	27.0	22.5	**34.7**	16.2	29.5

Table 2: P@1 scores over a representative subset of 10 language pairs from the PanLex-BLI dataset of Vulić et al. (2019). See §3 for the description of different model configurations/variants. $|\mathcal{D}| = 5k$. Highest scores per row are in **bold**. Respective average scores for the *noCL (1.0)* config (i.e., without contrastive learning and without interpolation with static VecMap CLWEs) are: 4.2 (mBERT), 3.1 (XLM-R), 17.0 (LaBSE), 8.3 (xMPNET).

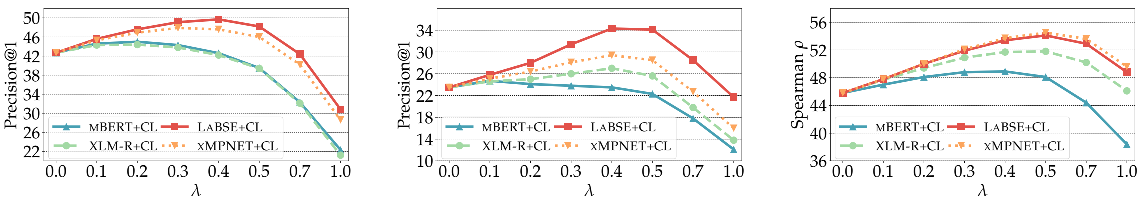
		LaBSE		
Pair ↓	VecMap	noCL (1.0)	+CL (1.0)	+CL (0.5)
BG–CA	15.2	14.0	18.0	**28.9**
BG–ET	12.5	20.3	25.5	**35.1**
BG–HE	5.6	18.3	20.8	**24.7**
BG–KA	9.1	16.0	21.6	**29.7**
CA–ET	9.8	24.8	25.6	**31.2**
CA–HE	5.0	10.6	12.2	**15.6**
CA–KA	5.5	5.7	8.3	**14.8**
ET–HE	3.1	**27.7**	25.1	25.4
ET–KA	4.6	13.2	16.0	**21.1**
HE–KA	3.2	19.0	22.0	**25.4**
Average	7.4	17.0	19.5	**25.2**

Table 3: P@1 scores over 10 language pairs from the PanLex-BLI dataset of Vulić et al. (2019) when $|\mathcal{D}| = 1k$, with different model variants based on LaBSE (see §3). Highest scores per row are in **bold**.

with lower-resource languages.

In Table 3 we compare the results of LaBSE (as the best-performing multilingual SE) against VecMap on PanLex-BLI in a scenario with less external bilingual supervision: $|\mathcal{D}| = 1k$. Interestingly, in this setup LaBSE already substantially outperforms VecMap out of the box (noCL (1.0)); contrastive lexical fine-tuning (+CL (1.0)) and interpolation with VecMap embeddings (+CL (0.5)) again bring further substantial gains, and we again observe a strong synergistic effect of the two components: +CL (1.0) yields gains over noCL (1.0) for 9/10 language pairs, and +CL (0.5) results in further boosts for all 10 pairs. Further, the contrastively fine-tuned LaBSE seems to be much more resilient to training data scarcity than VecMap: reduction of the training dictionary size from 5k to 1k reduces the performance of LaBSE +CL (λ) by 27% (from 34.7 to 25.2 P@1 points) compared to a massive performance drop of almost 70% for VecMap (from 23.5 to mere 7.4 P@1). In sum, the BLI results already indicate the wealth of lexical knowledge 'hidden' in multilingual SEs, which must be 'exposed to surface'.

Cross-Lingual Lexical Semantic Similarity (XLSIM). The average XLSIM results are summarized in Table 1. They again corroborate one of the main findings from BLI experiments: multilingual SEs store more cross-lingual lexical knowledge than



(a) GT-BLI (b) PanLex-BLI (c) Multi-SimLex (XLSIM)

Figure 2: Average scores across different interpolation values λ for the BLI task on **(a)** GT-BLI and **(b)** PanLex-BLI, and **(c)** the XLSIM task on Multi-SimLex. $|\mathcal{D}| = 5k$. Additional results are in Appendix B and C.

Config ↓ / Language L_t →	**DE**		**FI**		**RU**		**TR**	
	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5
XLM-R +noCL	0.0	0.1	0.1	0.2	0.1	0.2	0.4	0.5
XLM-R +noCL+UMLS$_{\text{EN}}$	27.6	32.0	12.2	14.7	21.8	25.9	29.3	35.9
XLM-R +noCL+UMLS$_{all}$	31.8	37.3	18.6	22.2	35.4	41.2	42.8	48.9
XLM-R +CL	14.1	17.1	5.0	6.5	8.7	11.2	21.6	27.1
XLM-R +CL+UMLS$_{\text{EN}}$	25.2	29.0	12.1	14.1	19.8	25.0	31.1	36.1
XLM-R +CL+UMLS$_{all}$	32.1	36.7	19.1	23.8	34.9	42.4	43.4	49.0
XMPNET +noCL	19.5	25.9	12.2	14.8	19.2	24.3	28.9	36.3
XMPNET +noCL+UMLS$_{\text{EN}}$	25.1	29.2	17.8	21.5	21.9	26.9	30.0	36.5
XMPNET +noCL+UMLS$_{all}$	**33.4**	37.8	**23.6**	**27.7**	**39.8**	45.4	**44.6**	**51.4**
XMPNET +CL	20.8	26.5	9.1	12.5	12.8	17.1	30.4	36.5
XMPNET +CL+UMLS$_{\text{EN}}$	25.1	28.7	11.4	14.0	21.8	27.2	31.0	37.5
XMPNET +CL+UMLS$_{all}$	32.0	**38.7**	22.9	27.5	39.2	**45.7**	44.3	51.0

Table 4: A summary of results in the XEL task on the biomedical XL-BEL benchmark of Liu et al. (2021d). We show the results of the better-performing LM (XLM-R), and the more lightweight multilingual SE (XMPNET).

multilingual LMs. This is validated by substantial gains of SEs over corresponding LMs across all configurations in Table 1. Interestingly, due to their contrastive learning objectives on sentence-level parallel data (Feng et al., 2022), LABSE and XMPNET provide very strong XLSIM results when used off-the-shelf (noCL (1.0)), outperforming the CLWE VECMAP embeddings. Contrastive lexical fine-tuning with 5k word translation pairs (+CL (1.0)) in this case does not bring any gains. However, the opposite is true for multilingual LMs: contrastive cross-lingual lexical fine-tuning on only 5k word translation pairs brings large benefits in the XLSIM task (e.g., compare noCL (1.0) and +CL (1.0) configurations for MBERT and XLM-R), and turns them into more effective lexical encoders. This result corroborates a similar finding from prior work in monolingual setups (Vulić et al., 2021a). Finally, interpolation with static CLWEs benefits the final XLSIM performance of all four underlying multilingual encoders: interpolated vectors ($\lambda = 0.5$) yield highest scores across the board, substantially surpassing gains both VECMAP and WEs from fine-tuned encoders ($\lambda = 1.0$).

Interpolation with Static CLWEs. A more detailed analysis over different λ values for BLI and XLSIM, summarized in Figure 2, reveals that interpolation can bring large performance gains, especially for λ in the $[0.3, 0.5]$ interval. The optimal λ value is, however, task- and even dataset-dependent. For instance, for low-resource BLI on PanLex-BLI more knowledge comes from the multilingual encoders as VECMAP CLWEs are of lower quality for such languages: in consequence, the optimal λ value 'moves away' from the static CLWEs towards encodings obtained by fine-tuned multilingual SE. We also note that larger benefits from interpolation are observed when VECMAP CLWEs are combined with contrastively fine-tuned multilingual SEs than with LMs: cf., the large gains in Figure 2b and in Table 3 for the LABSE +CL model variant.

Cross-Lingual Entity Linking (XEL). Experiments on XL-BEL (Liu et al., 2021d), summarized in Table 4, demonstrate that additional contrastive tuning with word or phrase pairs can greatly boost performance of multilingual LMs: even fine-tuning with 5k word translation pairs without any domain-specific knowledge yields strong benefits for XLM-R. As expected, using a much larger and domain-specific external database UMLS yields much higher scores and is more crucial for performance. In fact, contrastively fine-tuning on UMLS generally improves XEL performance with all four underlying models. Again, we observe that SE-based (XMPNET) configurations outperform the respective LM-based (XLM-R) configurations across the board. This finding again indicates that multilingual SEs store more cross-lingual lexical knowledge than multilingual LMs: this difference is particularly salient when the models are used off-the-shelf without any additional contrastive fine-tuning, and xMPNET retains the edge over XLM-R even after task-specific fine-tuning with the UMLS-based domain-specific knowledge.

What is more, for FI, RU, and TR, the multilingual xMPNET-based variants match or surpass the performance of respective XEL models trained on top of monolingual LMs (e.g., for FI, a model based on the Finnish BERT) reported by Liu et al. (2021d). This further validates our hypothesis that multilingual SEs store rich multilingual lexical knowledge, which is then also exposed in domain-specific (multilingual) UMLS fine-tuning, yielding performance gains. Contrastive fine-tuning on UMLS synonyms (+CL+UMLS variants) expectedly outpeforms fine-tuning on (5k) general-domain word translations (+CL), indicating that in specialized domains, if available, in-domain cross-lingual lexical signal should be exploited.

4.1 Summary and Discussion

Since the exposure of knowledge is done through very knowledge-light tuning which also improves representations of lexical items not covered in the dictionaries used for the adaptive fine-tuning, this suggests that the knowledge is stored in the parameters of the large models (both LMs and SEs), but it is more easily 're-purposed' through fine-tuning for SEs. One might posit that the SEs have an 'unfair' advantage over LMs as the primary purpose of the SEs, even before is encoding text items (i.e., sentences) for semantic similarity and search. In this paper, we verify the extent of that advantage for lexical-level encodings/embeddings (as representations of lexical knowledge): while exposing (i.e., re-purposing) works for both model types, they do not reach the same performance peaks and benefits for lexical tasks where such lexical encodings are a paramount. We leave fine-tuning with LM-style objectives for other types of tasks beyond lexical similarity and search for future work.

5 Conclusion and Future Work

We investigated strategies to probe and expose cross-lingual lexical knowledge from pretrained models, including multilingual language models (LMs) and multilingual sentence encoders (SEs). Based on an extensive probing experiments on a suite of cross-lingual lexical tasks, we verified that multilingual SEs (e.g., LaBSE, xMPNET) are superior to multilingual LMs (mBERT, XLM-R) in terms of stored cross-lingual lexical knowledge. We empirically validated that the SEs store more lexical knowledge than 'what meets the eye' when they are used off-the-shelf, but this knowledge must be exposed from them. To this end, we proposed new methods to further fine-tune their representations based on contrastive learning to 'rewire' the models' parameters and transform them from LMs and SEs into more effective cross-lingual lexical encoders. These lexical encoders yield gains for all underlying models, and are especially significant for resource-poor languages and in low-data learning regimes. While this work focused on two widely used state-of-the-art multilingual SEs, the contrastive framework is versatile and model-independent and can be applied on top of other multilingual SEs in future work. We will also investigate other more sophisticated contrastive learning strategies, look into ensembling of knowledge extracted from different SEs, and expand our probing experiments to more tasks and languages.

Acknowledgements

This work has been partially supported by the ERC PoC Grant MultiConvAI (no. 957356) and a Huawei research donation to the University of Cambridge. Ivan Vulić is supported by a personal Royal Society University Research Fellowship (no 221137; 2022-2027). Fangyu Liu is supported by Grace & Thomas C.H. Chan Cambridge Scholarship. The work of Goran Glavaš has been supported by the Multi2ConvAI project of MWK Baden-Württemberg. The work of Edoardo Maria Ponti's work has been supported by the Facebook CIFAR AI Chair program during his affiliation with Mila – Quebec AI Institute. Nigel Collier kindly acknowledges grant-in-aid funding from ESRC (grant number ES/T012277/1). We would like to thank Yaoyiran Li for interesting insights and discussions.

We also thank the three anonymous reviewers for their extremely substantial and in-depth reviews with a large number of helpful suggestions.

Limitations

This work focuses on lexical specialization of multilingual encoders, off-the-shelf LMs (experiments with mBERT and XLM-R Base) and, in particular multilingual encoders specialized for sentence-level semantics (experiments with LaBSE and xMPNET). While these are all widely used models, they are arguably among the smaller pretrained multilingual encoders. Due to computational constraints, we have not evaluated the effectiveness of the proposed cross-lingual lexical specialisation for larger multilingual LMs, e.g., XLM-R-Large (Conneau et al., 2020) or mT5 (Large, XL, and XXL) (Xue et al., 2021). It is possible that these larger multilingual LMs would close (some of) the performance gap w.r.t. multilingual SEs. Such large LMs, however, are effectively available to fewer researchers and practitioners. Our work includes less resource-demanding LMs and SEs, making their lexically specialized variants that we offer more widely accessible.

Lexical input (i.e., words or phrases) are provided to each multilingual encoder fully *"in isolation"* (see §2), without any surrounding context. However, the alternative of using external corpora and *averaging-over-context* (Litschko et al., 2022), which we have not evaluated in this work for clarity and space constraints, might yield slightly improved task performance. Nonetheless, the 'in isolation' approach has been verified in previous work (Vulić et al., 2021a; Litschko et al., 2022; Li et al., 2022) as very competitive, and is more lightweight: **1)** it disposes of any external text corpora and is not impacted by the external data; **2)** it encodes words more efficiently due to the absence of context. Moreover, it allows us to directly study the richness of cross-lingual information stored in the encoders' parameters, and its interaction with additional cross-lingual signal from bilingual lexicons.

The contrastive cross-lingual lexical fine-tuning we proposed in this is work is *bilingual*. It leverages a small bilingual dictionary $\mathcal{D}$ for each language pair and specializes the multilingual encoders (LMs and SEs) independently for each language pair. Assuming interest in cross-lingual lexical tasks between all pairs of N_L languages, this entails $\frac{N_L \cdot (N_L - 1)}{2}$ fine-tuning procedures and as many resulting bilingual models. Although our contrastive fine-tuning is relatively fast and lightweight, given that it leverages at most 5k translation pairs, for large N_L it could easily exceed the computational and time budget for most users. On a high level, our work again outlines the advantages as well as disadvantages between **1)** more versatile massively multilingual models that serve multiple languages without any further adaptation, and **2)** better-performing but typically less modular and less versatile models adapted (i.e., *bilingually specialized*) from the multilingual models (Bapna and Firat, 2019; Parović et al., 2022).

Intuitively, for each bilingually fine-tuned model, we evaluate the performance for that respective language pair. Currently, we do not investigate the spillover effects that a bilingual lexical fine-tuning of multilingual encoders could have on lexical representations of other languages. Such an analysis, planned for future work, would be particularly interesting in the context of low-resource languages, unseen from the point of view of cross-lingual lexical fine-tuning, and in particular closely related low-resource languages. For instance, if we are doing cross-lingual lexical fine-tuning for language pairs involving Turkish, are there spillover benefits for low(er)-resource Turkic languages such as Uyghur or Kazakh?

Finally, we acknowledge that our choice of lexical tasks as probing tasks is non-exhaustive: we put focus on standard tasks from previous work on (multilingual) lexical semantics that are especially convenient as cross-lingual lexical probes: such tasks directly test and compare the quality of cross-lingual lexical representations obtained via different methods; see §4.1 again.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL 2018*, pages 789–798.

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

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of EMNLP*, pages 1538–1548.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

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

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of ACL 2020*, pages 4758–4781.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of EMNLP 2018*, pages 169–174.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of ACL 2022*, pages 4702–4712.

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 ACL 2020*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS 2019*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR 2018*.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of NAACL-HLT 2022*, pages 3610–3623.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP 2019*, pages 55–65.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of ACL 2022*, pages 878–891.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP 2021*, pages 6894–6910.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of ACL 2020*, pages 7548–7555.

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 ACL 2019*, pages 710–721.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of EMNLP 2020*, pages 2161–2174.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of ACL 2019*, pages 5392–5404.

Geert Heyman, Bregt Verreet, Ivan Vulić, and Marie-Francine Moens. 2019. Learning unsupervised multilingual word embeddings with incremental multilingual hubs. In *Proceedings of NAACL-HLT 2019*, pages 1890–1902.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of ACL 2019*, pages 3651–3657.

David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of LREC 2014*, pages 3145–3150.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of EMNLP 2020*, pages 9119–9130.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2022. Improving word translation via two-stage contrastive learning. In *Proceedings of ACL 2022*, pages 4353–4374.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of EMNLP 2020*, pages 1663–1674.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval*, 25:149–183.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021a. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *Proceedings of EMNLP 2021*, pages 2396–2406.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021b. Self-alignment pretraining for biomedical entity representations. In *Proceedings of NAACL-HLT 2021*, pages 4228–4238.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021c. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of EMNLP 2021*, pages 1442–1459.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021d. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*, pages 565–574.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *Proceedings of ICLR 2018*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint, CoRR*, abs/1309.4168.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Proceedings of ECCV 2020*, volume 12370, pages 681–699.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of NAACL-HLT 2022*, pages 1791–1799.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP 2019*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of EMNLP 2020*, pages 4512–4525.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: what we know about how BERT works. *Transactions of the ACL*, 8:842–866.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Peter H Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pretraining for language understanding. In *Proceedings of NeurIPS 2020*, pages 16857–16867.

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

Thomas Viklands. 2006. *Algorithms for the weighted orthogonal Procrustes problem and other least squares problems*. Ph.D. thesis, Datavetenskap.

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, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP-IJCNLP 2019*, pages 4407–4418.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021a. LexFit: Lexical fine-tuning of pretrained language models. In *Proceedings of ACL-IJCNLP 2021*, pages 5269–5283.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. In *Proceedings of EMNLP 2020*, pages 7222–7240.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021b. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of EMNLP 2021*, pages 1151–1168.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS 2019*, pages 3261–3275.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of ACL 2022*, pages 863–877.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT 2021*, pages 483–498.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of ACL-IJCNLP 2021*, pages 5065–5075.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of ACL 2020: System Demonstrations*, pages 87–94.

Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. 2022. Prix-LM: Pretraining for multilingual knowledge base construction. In *Proceedings of ACL 2022*, pages 5412–5424.

Languages in: GT-BLI, Multi-SimLex, XL-BEL	
EN	English
DE	German
TR	Turkish
FI	Finnish
HR	Croatian
RU	Russian
IT	Italian
FR	French
Languages in: PanLex-BLI	
BG	Bulgarian
CA	Catalan
ET	Estonian
HE	Hebrew
KA	Georgian

Table 5: Languages and their ISO 639-1 codes.

A List of Languages

The list of languages used in this work, along with their ISO 639-1 codes, is available in Table 5.

B BLI Results across Individual Language Pairs

Additional experiments and analyses over individual language pairs and other λ values, which further support the main claims of the paper, have been relegated to the appendix for clarity and compactness of the presentation in the main paper:

Table 6. It provides results over all 28 language pairs in GT-BLI with 2 multilingual LMs and 2 multilingual SEs in the *noCL* variant without contrastive fine-tuning.

Table 7. It provides results over all 28 language pairs in GT-BLI with 2 multilingual LMs and 2 multilingual SEs in the *+CL* variant with contrastive fine-tuning.

Table 8. It provides results over all 28 language pairs in GT-BLI and across different λ values with the LABSE +noCL variant.

Table 9. It provides results over all 28 language pairs in GT-BLI and across different λ values with the LABSE +CL variant.

C XLSIM Results across Individual Language Pairs

Table 10. It provides results over selected 6 language pairs from Multi-SimLex with 2 multilingual LMs and 2 multilingual SEs in the *noCL* variant without contrastive fine-tuning.

Table 11. It provides results over selected 6 language pairs from Multi-SimLex with 2 multilingual LMs and 2 multilingual SEs in the *+CL* variant with contrastive fine-tuning.

Table 12. It provides results over selected 6 language pairs from Multi-SimLex and across different λ values with the XMPNET +noCL variant.

Table 13. It provides results over selected 6 language pairs from Multi-SimLex and across different λ values with the XMPNET +CL variant.

D Models and Evaluation Data

URLs to the models used in this paper are provided in Table 14. Training and test data for all three tasks (BLI, XLSIM, XEL) is available online:

- GT-BLI is available here: https://github.com/codogogo/xling-eval
- PanLex-BLI: https://github.com/cambridgeltl/panlex-bli
- Multi-SimLex [XLSIM]: https://multisimlex.com/
- XL-BEL [XEL]: https://github.com/cambridgeltl/sapbert

Our code is based on PyTorch, and relies on the following two widely used repositories:

- sentence-transformers: www.sbert.net
- huggingface.co/transformers/

		mBERT		XLM-R		LaBSE		xMPNET	
Pair ↓ / Config →	VecMap	+noCL (1.0)	+noCL (0.3)	+noCL (1.0)	+noCL (0.3)	+noCL (1.0)	+noCL (0.3)	+noCL (1.0)	+noCL (0.3)
EN–DE	55.6	15.6	50.7	12.7	45.5	25.4	54.1	22.0	47.7
EN–TR	40.4	7.2	34.9	6.0	28.9	23.6	42.1	16.1	33.7
EN–FI	45.6	7.9	38.7	6.6	33.4	19.3	45.1	14.8	39.5
EN–HR	37.5	8.9	31.8	6.8	25.6	24.7	45.3	18.5	38.9
EN–RU	45.6	3.2	40.7	0.9	34.7	24.7	49.9	17.6	41.7
EN–IT	60.2	12.3	57.1	9.3	53.0	26.4	62.3	23.5	58.8
EN–FR	64.1	25.2	62.5	19.5	56.6	34.2	67.5	29.2	61.7
DE–TR	32.5	9.1	28.9	6.9	24.4	17.7	33.0	13.1	29.7
DE–FI	39.7	9.2	34.1	7.3	30.1	16.0	37.4	13.3	34.7
DE–HR	33.3	11.5	31.0	9.7	25.7	19.2	38.5	14.9	34.1
DE–RU	40.0	4.2	36.4	0.9	32.4	14.3	41.5	9.5	37.7
DE–IT	49.5	10.9	45.9	8.1	42.7	19.4	51.4	18.3	49.1
DE–FR	50.0	15.8	49.7	10.5	42.3	22.5	53.2	20.2	49.8
TR–FI	31.3	6.7	26.2	5.2	22.0	15.5	31.7	11.3	30.9
TR–HR	25.4	10.8	24.3	8.3	20.5	18.7	33.1	14.4	28.2
TR–RU	32.9	2.6	29.1	0.8	25.5	14.1	36.9	11.3	33.5
TR–IT	37.1	7.9	34.7	5.5	27.4	17.0	38.9	14.8	38.3
TR–FR	39.4	7.8	37.3	5.7	30.9	20.9	43.1	16.6	39.4
FI–HR	30.4	7.2	26.6	5.5	22.8	17.4	36.4	12.2	32.7
FI–RU	38.2	2.6	34.0	0.9	30.5	15.1	41.0	9.0	37.1
FI–IT	39.9	7.9	36.7	6.8	30.4	18.1	43.4	16.4	41.8
FI–FR	42.8	7.5	38.9	5.9	32.2	18.6	45.9	16.2	42.2
HR–RU	40.6	6.0	35.8	1.6	30.4	24.5	45.7	16.3	41.4
HR–IT	40.4	11.2	39.0	8.4	31.3	24.5	47.9	22.4	44.5
HR–FR	43.6	9.7	42.3	6.1	30.6	25.7	50.0	19.8	43.9
RU–IT	46.6	3.1	42.2	1.5	36.4	21.8	47.6	18.4	46.5
RU–FR	48.7	4.1	44.3	1.5	38.4	26.6	50.9	18.9	47.5
IT–FR	64.1	16.6	62.8	9.3	58.6	33.9	65.6	27.5	63.1
Average	42.7	9.0	39.2	6.4	33.7	21.4	45.7	17.0	41.7

Table 6: Individual P@1 scores (×100%) for all 28 language pairs in the GT-BLI dataset of Glavaš et al. (2019), with multilingual LMs and SEs used 'off-the-shelf' *without contrastive fine-tuning* (§2). See §3 for the description of different model configurations/variants. $|\mathcal{D}| = 5k$. The number in the parentheses denotes the value for λ (see §3): the value of 1.0 effectively means 'no interpolation' with static VecMap CLWEs.

		mBERT		XLM-R		LaBSE		xMPNET	
Pair ↓ / Config →	VecMap	+CL (1.0)	+CL (0.3)	+CL (1.0)	+CL (0.3)	+CL (1.0)	+CL (0.3)	+CL (1.0)	+CL (0.3)
EN–DE	55.6	26.4	59.2	24.5	56.3	31.6	61.1	30.2	58.7
EN–TR	40.4	17.4	39.3	19.2	42.0	33.1	50.1	30.4	45.7
EN–FI	45.6	18.6	45.6	20.2	45.7	30.8	53.3	28.7	50.4
EN–HR	37.5	23.6	44.3	21.8	44.3	36.9	53.9	31.8	49.6
EN–RU	45.6	23.9	48.9	28.3	48.8	46.4	55.8	37.9	53.1
EN–IT	60.2	26.8	64.0	25.4	61.7	33.3	66.9	33.0	65.3
EN–FR	64.1	34.4	67.7	33.0	65.4	42.1	71.2	39.5	68.5
DE–TR	32.5	19.5	32.8	15.5	33.5	24.4	37.6	22.7	36.2
DE–FI	39.7	20.3	38.5	19.0	37.5	25.0	43.3	24.4	42.1
DE–HR	33.3	24.2	37.3	22.4	36.9	27.9	42.9	27.2	41.8
DE–RU	40.0	21.2	42.0	19.1	41.5	27.5	45.6	26.8	44.1
DE–IT	49.5	23.0	49.9	19.0	48.6	24.6	52.5	25.6	51.4
DE–FR	50.0	27.4	52.5	24.3	51.5	31.3	54.5	30.0	53.9
TR–FI	31.3	18.0	29.6	15.9	31.7	22.2	34.9	22.4	35.9
TR–HR	25.4	21.8	30.2	19.0	30.8	27.4	36.8	26.0	36.4
TR–RU	32.9	16.2	33.9	15.7	33.5	24.7	37.4	24.7	37.8
TR–IT	37.1	17.4	37.8	15.4	36.8	22.8	41.6	22.2	41.0
TR–FR	39.4	18.4	40.4	17.4	38.9	29.4	43.9	26.3	43.6
FI–HR	30.4	20.2	32.7	17.4	32.5	27.4	39.7	24.3	38.4
FI–RU	38.2	17.6	37.3	18.0	38.7	27.6	42.4	27.6	41.3
FI–IT	39.9	18.4	40.5	17.6	40.0	25.2	45.3	24.0	44.9
FI–FR	42.8	17.5	43.0	18.2	41.9	26.4	47.6	26.1	45.8
HR–RU	40.6	26.4	41.4	29.5	43.9	36.1	46.4	33.3	46.4
HR–IT	40.4	24.6	44.0	22.8	42.8	32.3	49.7	30.2	49.2
HR–FR	43.6	24.1	46.5	23.4	44.1	35.0	51.8	29.3	50.7
RU–IT	46.6	22.0	46.9	19.4	45.0	32.0	50.2	28.4	51.1
RU–FR	48.7	22.1	49.1	20.5	47.6	37.0	53.5	28.9	51.2
IT–FR	64.1	33.4	65.5	32.9	64.3	42.2	66.0	39.0	66.4
Average	42.7	22.3	44.3	21.2	43.8	30.8	49.1	28.6	47.9

Table 7: Individual P@1 scores (×100%) for all 28 language pairs in the GT-BLI dataset of Glavaš et al. (2019), with model variants *with contrastive fine-tuning* (§2). $|\mathcal{D}| = 5k$.

		LaBSE +noCL						
Pair $\downarrow$ **/** $\lambda = \rightarrow$	VecMap(0.0)	0.1	0.2	0.3	0.4	0.5	0.7	1.0
EN–DE	55.6	57.3	56.3	54.1	50.9	47.8	40.6	25.4
EN–TR	40.4	42.0	41.9	42.1	41.6	40.2	35.9	23.6
EN–FI	45.6	46.1	46.1	45.1	44.4	42.8	34.4	19.3
EN–HR	37.5	39.8	43.0	45.3	46.5	46.5	41.3	24.7
EN–RU	45.6	46.5	48.0	49.9	50.5	50.2	46.1	24.7
EN–IT	60.2	61.6	63.0	62.3	61.3	60.1	49.9	26.4
EN–FR	64.1	66.0	67.1	67.5	65.9	64.8	55.3	34.2
DE–TR	32.5	33.6	33.4	33.0	32.3	30.2	24.4	17.7
DE–FI	39.7	39.7	39.3	37.4	35.9	34.3	25.8	16.0
DE–HR	33.3	35.8	37.0	38.5	38.8	36.1	29.2	19.2
DE–RU	40.0	40.9	41.2	41.5	41.1	38.4	29.6	14.3
DE–IT	49.5	51.3	51.3	51.4	49.4	46.3	36.4	19.4
DE–FR	50.0	52.2	53.2	53.2	51.8	49.1	37.8	22.5
TR–FI	31.3	32.3	31.6	31.7	31.2	29.8	24.8	15.5
TR–HR	25.4	28.3	31.2	33.1	34.1	33.6	28.9	18.7
TR–RU	32.9	34.9	35.5	36.9	36.5	34.0	29.0	14.1
TR–IT	37.1	38.7	39.6	38.9	38.8	37.7	31.5	17.0
TR–FR	39.4	41.4	42.1	43.1	43.4	41.9	34.6	20.9
FI–HR	30.4	32.3	34.7	36.4	36.9	36.3	28.6	17.4
FI–RU	38.2	39.5	40.0	41.0	40.5	37.6	28.9	15.1
FI–IT	39.9	42.9	42.9	43.4	44.2	41.5	34.0	18.1
FI–FR	42.8	44.7	45.9	45.9	46.1	43.6	34.6	18.6
HR–RU	40.6	41.8	43.9	45.7	45.8	45.0	39.0	24.5
HR–IT	40.4	43.5	46.0	47.9	48.6	47.6	41.8	24.5
HR–FR	43.6	46.8	48.6	50.0	50.1	47.9	42.0	25.7
RU–IT	46.6	48.1	48.1	47.6	46.8	44.5	38.7	21.8
RU–FR	48.7	50.2	51.0	50.9	50.0	49.0	43.6	26.6
IT–FR	64.1	64.9	65.8	65.6	64.9	63.0	54.0	33.9
Average	42.7	44.4	45.3	45.7	45.3	43.6	36.5	21.4

Table 8: Individual P@1 scores ($\times 100\%$) for all 28 language pairs in the GT-BLI dataset of Glavaš et al. (2019), across different values for λ. The model variant is LaBSE +noCL (see §3); similar patterns are observed with another multilingual SE in our evaluation (xMPNET). $|\mathcal{D}| = 5k$.

		LaBSE +CL						
Pair $\downarrow$ **/** $\lambda = \rightarrow$	VecMap(0.0)	0.1	0.2	0.3	0.4	0.5	0.7	1.0
EN–DE	55.6	59.4	61.2	61.1	60.1	56.8	46.5	31.6
EN–TR	40.4	43.9	46.8	50.1	51.0	49.9	45.4	33.1
EN–FI	45.6	48.6	51.0	53.3	54.1	53.9	46.1	30.8
EN–HR	37.5	44.1	49.7	53.9	56.7	57.1	49.7	36.9
EN–RU	45.6	50.1	52.5	55.8	58.5	59.0	56.2	46.4
EN–IT	60.2	62.5	65.3	66.9	67.2	66.2	55.7	33.3
EN–FR	64.1	66.3	69.5	71.2	71.1	69.5	59.5	42.1
DE–TR	32.5	35.4	36.5	37.6	36.7	35.6	31.8	24.4
DE–FI	39.7	41.6	42.5	43.3	42.2	39.5	32.7	25.0
DE–HR	33.3	37.9	40.8	42.9	43.1	41.6	35.9	27.9
DE–RU	40.0	43.1	44.0	45.6	45.9	44.7	37.3	27.5
DE–IT	49.5	51.3	52.0	52.5	50.9	47.7	38.9	24.6
DE–FR	50.0	52.4	53.7	54.5	53.6	50.3	42.7	31.3
TR–FI	31.3	33.3	34.3	34.9	35.0	33.7	30.2	22.2
TR–HR	25.4	30.4	34.2	36.8	38.5	38.2	35.8	27.4
TR–RU	32.9	35.2	36.6	37.4	37.5	36.1	32.5	24.7
TR–IT	37.1	39.0	41.2	41.6	41.5	39.9	34.4	22.8
TR–FR	39.4	41.8	43.0	43.9	43.8	43.3	38.6	29.4
FI–HR	30.4	34.0	37.5	39.7	41.2	40.2	36.4	27.4
FI–RU	38.2	40.2	41.1	42.4	42.6	40.2	36.2	27.6
FI–IT	39.9	43.3	44.3	45.3	45.9	44.7	38.2	25.2
FI–FR	42.8	44.5	46.3	47.6	47.8	46.0	40.0	26.4
HR–RU	40.6	42.0	44.9	46.4	47.5	48.3	44.7	36.1
HR–IT	40.4	43.3	47.7	49.7	50.6	50.4	46.1	32.3
HR–FR	43.6	47.2	49.0	51.8	52.0	51.0	46.4	35.0
RU–IT	46.6	48.8	49.6	50.2	50.4	48.7	44.9	32.0
RU–FR	48.7	51.0	51.9	53.5	53.1	52.1	47.1	37.0
IT–FR	64.1	64.9	65.9	66.0	66.0	64.6	56.9	42.2
Average	42.7	45.6	47.6	49.1	49.4	48.2	42.4	30.8

Table 9: Individual P@1 scores ($\times 100\%$) for all 28 language pairs in the GT-BLI dataset of Glavaš et al. (2019), across different values for λ. The model variant is LaBSE +CL (see §3); similar patterns are observed with another multilingual SE in our evaluation (xMPNET). $|\mathcal{D}| = 5k$.

		mBERT		XLM-R		LaBSE		xMPNET	
Pair ↓ / Config →	VecMap	+noCL	+noCL	+noCL	+noCL	+noCL	+noCL	+noCL	+noCL
		(1.0)	(0.5)	(1.0)	(0.5)	(1.0)	(0.5)	(1.0)	(0.5)
EN–FI	42.2	1.1	30.9	1.4	21.4	46.8	51.6	47.7	53.4
EN–RU	39.7	5.2	30.1	2.5	21.3	53.5	52.3	57.5	55.2
EN–FR	64.8	11.3	52.8	2.2	29.6	68.5	74.2	64.7	73.9
FI–RU	33.0	4.3	25.3	3.2	20.7	38.4	41.6	42.7	45.6
FI–FR	46.7	3.6	35.3	0.3	22.8	43.1	52.6	42.9	53.2
RU–FR	48.2	8.6	37.9	0.8	25.3	52.2	57	52.3	58.4
Average	45.8	5.7	35.4	1.7	23.5	50.4	54.9	51.3	56.6

Table 10: Individual Spearman's ρ correlation scores ($\times 100$) on the XLSIM task (Multi-SimLex) for a subset of language pairs in our evaluation, with multilingual LMs and SEs used 'off-the-shelf' *without contrastive fine-tuning* (§2). See §3 for the description of different model configurations/variants. $|\mathcal{D}| = 5k$, with XLSIM test pairs removed from the dictionary. The number in the parentheses denotes the value for λ (see §3).

		mBERT		XLM-R		LaBSE		xMPNET	
Pair ↓ / Config →	VecMap	+noCL	+noCL	+noCL	+noCL	+noCL	+noCL	+noCL	+noCL
		(1.0)	(0.5)	(1.0)	(0.5)	(1.0)	(0.5)	(1.0)	(0.5)
EN–FI	42.2	32.4	43.5	42.3	48	45.6	50.6	48.1	51.5
EN–RU	39.7	34.8	42.1	45.8	47.3	47.1	49.2	50.5	50.4
EN–FR	64.8	56.5	67.4	57.9	69.2	64	72	64.1	71.9
FI–RU	33.0	28.4	35.9	38.3	40.8	38.3	42.3	40.1	43.1
FI–FR	46.7	34.7	47.3	41.7	50.6	45.9	53.5	46.6	54.2
RU–FR	48.2	43.6	52.1	50.4	55.1	51.8	56.8	48.5	55.6
Average	45.8	38.4	48.1	46.1	51.8	48.8	54.1	49.6	54.5

Table 11: Individual Spearman's ρ correlation scores ($\times 100$) on the XLSIM task (Multi-SimLex) for a subset of language pairs in our evaluation, with model variants *with contrastive fine-tuning* (§2). $|\mathcal{D}| = 5k$.

	xMPNET +noCL								
Pair ↓ / $\lambda = \rightarrow$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	1.0
EN–FI	42.2	45.2	48.2	50.7	52.6	53.4	53	52.2	47.7
EN–RU	39.7	43.1	46.6	49.9	52.9	55.2	56.7	57.5	57.5
EN–FR	64.8	67.6	70.1	72.2	73.5	73.9	73.2	71.7	64.7
FI–RU	33	36.1	39.3	42.5	44.7	45.6	45.5	45.4	42.7
FI–FR	46.7	49.2	51.5	53.1	53.6	53.2	51.5	49.6	42.9
RU–FR	48.2	51.1	53.8	56.2	57.7	58.4	58.1	57.2	52.3
Average	45.8	48.7	51.6	54.1	55.8	56.6	56.3	55.6	51.3

Table 12: Individual Spearman's ρ correlation scores ($\times 100$) on the XLSIM task (Multi-SimLex) for a subset of language pairs in our evaluation, across different values for λ. The model variant is xMPNET +noCL (see §3); similar patterns are observed with another multilingual SE in our evaluation (LaBSE). $|\mathcal{D}| = 5k$.

	xMPNET +CL								
Pair ↓ / $\lambda = \rightarrow$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	1.0
EN–FI	42.2	44.2	46.4	48.6	50.4	51.5	51.8	51.3	48.1
EN–RU	39.7	41.6	43.9	46.3	48.6	50.4	51.6	52	50.5
EN–FR	64.8	66.9	69	70.7	71.8	71.9	71	69.5	64.1
FI–RU	33.0	35.1	37.5	39.9	41.9	43.1	43.5	43	40.1
FI–FR	46.7	48.9	51.2	53.1	54.2	54.2	53.3	51.7	46.6
RU–FR	48.2	50.1	52.1	53.8	55.2	55.6	55.1	53.8	48.5
Average	45.8	47.8	50	52.1	53.7	54.5	54.4	53.6	49.6

Table 13: Individual Spearman's ρ correlation scores ($\times 100$) on the XLSIM task (Multi-SimLex) for a subset of language pairs in our evaluation, across different values for λ. The model variant is xMPNET +CL (see §3); similar patterns are observed with another multilingual SE in our evaluation (LaBSE). $|\mathcal{D}| = 5k$.

Name	**URL**
mBERT	huggingface.co/bert-base-multilingual-uncased
XLM-R	huggingface.co/xlm-roberta-base
LaBSE	huggingface.co/sentence-transformers/LaBSE
xMPNET	huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Table 14: URLs of the multilingual Transformer models used in this work.