A Survey of Cross-Lingual Alignment: Definitions, Methods, Future

Anonymous ACL submission

Abstract

Cross-lingual alignment in multilingual language models has been an active field of research in recent years. We survey the literature of techniques, both to train well-aligned models, and to improve the cross-lingual alignment of pre-trained encoders. Compiling evaluation results and method summaries, we give an overview of which methods work better than others. We further show how to understand cross-lingual alignment and its limitations. Finally, we discuss how these insights may be applied not only to encoder models, where this topic has been heavily studied, but also to encoder-decoder or even decoder-only models. In generative models, the focus must be on an effective trade-off between language-neutral and language-specific information.

1 Introduction

Zero-shot cross-lingual transfer using highly multilingual models has been an active subset of multilingual NLP research. In tasks like sentence classification, sequence labelling, or sentence retrieval, all of which rely on encoder representations, cross-lingual overlap of those representations is an underlying assumption. As we define it, cross-lingual alignment means that words or sentences with similar semantics are:

1. more similar in the representation space than words or sentences with dissimilar semantics.

2. similar enough that a prediction head trained on a source language will recognise the relevant patterns in the target language.

These criteria are not guaranteed to be fulfilled through unsupervised pre-training, motivating efforts to improve the cross-lingual alignment by various methods. We surveyed a number of papers in this area. These papers propose new training objectives, pre-trained new models, contrastive fine-tuning, or post-hoc adjustments of the embedding space. The vast majority of these methods were developed for and applied to multilingual encoder models, chiefly XLM-R and mBERT. We address future research on generative models in § 5.

The contributions of this paper are: a thorough review of papers in this space from the last years (§ 3, § 4), a higher-level discussion of cross-lingual alignment and the representation space (§ 2), and a discussion about cross-lingual alignment and future research in the context of generative models (§ 5).

2 Cross-Lingual Alignment

2.1 Definitions

"Alignment" is an overloaded term in NLP, referring to word alignment in machine translation (Och et al., 1999), or to desirable model behaviour in chatbot training (Ouyang et al., 2022). In our case, it refers to the meaningful similarity of multilingual representations across languages. "Cross-lingual alignment" in this sense was used in static word embeddings, and can be applied to contextual models as well. We define two main requirements:

1) Similar meanings have more similar representations than dissimilar meanings do. When querying cross-lingually, the nearest neighbour of a word representation should be its translation. This implies it is not enough that similar meanings are represented in similar ways; it is also necessary that dissimilar meanings are represented in dissimilar ways. This property is critical for tasks relying on retrieval from the space. A "stronger" cross-lingual alignment (c.f. Abulkhanov et al., 2023) would additionally require that word representations be more similar to their translations than to dissimilar words in the same language.

2) A prediction head trained on a source language should be able to find relevant patterns in the representations of a target language, and classify accordingly. Although it is tempting to think of similarity in terms of simple measures such as cosine similarity, the classification head works with the full encoder representations as its input, and can use subspaces to that effect. This property is crucial for fine-tuning tasks, i.e., classification or question answering. This requirement also implies that representations, or at least subspaces thereof, are *close to isomorphic* (see § 2.2).

Although *cross-lingual alignment* is not the only possible term, it has been commonly used in the related research and is likely to be understood. It implies a relationship between matched words across languages, at as many points as possible, and indeed a complex optimisation problem. It requires good but not necessarily perfect correspondence between spaces, as we will discuss further.

2.2 Isomorphism of Representations

As has been pointed out for cross-lingual static embeddings, alignment between two language spaces depends on an assumption of *isomorphism*, i.e., that both spaces have (roughly) the same shape and can be *linearly transformed* onto each other in such a way that equivalent tokens are consistently aligned (Vulić et al., 2020). This assumption may not always hold due to cultural-semantic differences, imperfect translation of concepts (e.g., Gibson et al., 2017), typological differences, different corpus domains, different data sizes, and more (Ormazabal et al., 2019; Vulić et al., 2020).

We can think of cross-lingual alignment as a complex optimisation problem in this light—to be completely cross-lingually aligned, the model would have to reconcile both large and small differences between many different language spaces. However, this may simply not be necessary in order to fulfill our two conditions *reasonably well*.

That said, Vulić et al. (2020) emphasise that undertraining contributes significantly to non-isomorphism in static embeddings, and this may well apply to contextual models. For example, we know that contextual models also encode token frequency (Rajaee and Pilehvar, 2022; Puccetti et al., 2022), which is implicitly related to each language's vocabulary and data size.

2.3 Subspaces

Contextual representations encode not only semantic aspects, but also morphosyntactic aspects (Hewitt and Manning, 2019; Acs et al., 2023), token frequency, and more. They are likely to pick up on many other details to some degree, including spurious attributes and noise. Among the hundreds of dimensions in the models, *subspaces* can correspond to more specific aspects. These can be found mathematically, through projections of the raw representations. For instance, Chang et al. (2022) find affine subspaces that correspond to language-sensitive as well as language-neutral information.

In effect, Chang et al. (2022) separate types of axes by how their means and variances differ in different languages. That is, if both are similar across languages, the axis is language-neutral. If the means differ between languages and/or variances are very different, the axis is language-sensitive.

2.4 Measuring Cross-Lingual Alignment

Cross-lingual alignment or language-neutrality has been measured using a range of metrics, none of which show the full picture: *Word or sentence retrieval tasks* measure the model's ability to encode the correct translation more similarly to the query than other candidate translations. Often this is measured by cosine, i.e., angular similarity, after normalising vector length, or an adjusted retrieval score such as CSLS (Lample et al., 2018).

Cosine similarity between matched words, or average cosine similarity between language spaces, has also been used more directly as a measure of cross-lingual alignment. It is important in such cases to compare against the average cosine similarity in the space, which can be quite high (Ethayarajh, 2019; Rajaee and Pilehvar, 2022).

Isomorphism between two representation spaces can be measured using *relational similarity* (Vulić et al., 2020), *eigenvector similarity* (Søgaard et al., 2018), or the *Gromov-Hausdorff distance* (Gromov, 1999; Patra et al., 2019).

Language identification is sometimes used (e.g., Libovický et al., 2020) to reveal language-specific elements of the representations. In this thinking, if a language classifier trained on the output representations performs worse, then the model outputs are more language-neutral. However, this neglects that the representations can have both language-neutral and language-specific areas (§ 2.3).

Zero-shot cross-lingual transfer, after fine-tuning, is both an aim in itself and a proxy for how well-aligned the representations are. Of course, fine-tuning will change the model again, but interventions before and/or during fine-tuning have been shown to improve transfer performance. The metrics used depend on the respective task, but a common way to highlight cross-lingual transfer is to report the *transfer gap*, i.e., the difference between source language performance and the average target language performance.

Finally, though not a metric, we mention *t-SNE* (van der Maaten and Hinton, 2008) here. This is a visualisation method where spaces are projected down into two or three dimensions for graphing, and it can be extremely helpful to get a better sense of what the space looks like. However, we must remember that due to the down-projection and selection of examples, we can see only some aspects of the representation space at any given time.

3 Strategies to Increase Alignment

We report on a number of strategies for improving zero-shot transfer and increasing cross-lingual alignment, sorting them by aspects such as continued training or full pre-training, word-level or sentence-level objectives, and more. Table 1 lists all included papers, organised by initialisation, objectives and kinds of data. In this section, we show different strategies with examples, adding categories that are not in the table as they would overlap with multiple table cells. We leave out some methods that are less relevant to the overall analysis, though we explain them in Appendix B for completeness.

3.1 Word-Level vs. Sentence-Level Objectives

First, we discuss models using external parallel data—sentence-parallel or word-parallel—which is a plurality of methods in this survey. In some cases, a sentence-parallel corpus is used and word-level alignments are induced before training. We tabulate the methods based on whether the proposed objectives focus on word-level alignments, or only sentence-level ones. “Both levels” refers mostly to methods using multiple alignment objectives. In many cases, the alignment objective is combined with a regularisation or joint objective.

Word-level alignment. Cao et al. (2020) is an influential early work in explicit cross-lingual alignment training, using parallel texts. The objective is “contextual word retrieval”, searching for word matches over the entire corpus using CSLS (Lample et al., 2018), which deals better than cosine similarity with hubness issues. As a regulariser, they keep the model similar to its initialisation. Wu and Dredze (2020) propose a similar objective with a contrastive loss, which is “strong” or “weak” based on whether negative examples are considered from both the source and target language or only from the target language. Zhao et al. (2021) also use a similar alignment process and combine it with batch normalisation, i.e., forcing “all embeddings of different languages into a distribution with zero mean and unit variance”. Alqahtani et al. (2021), meanwhile, formulate cross-lingual word alignment as an optimal transport problem. XLM-Align (Chi et al., 2021b) combines denoising word alignment with self-labelled word alignment in an EM manner.

Word- and Sentence-level. These models either use multiple objectives, or use objectives that are hard to categorise as either word- or sentence-level.

For instance, Hu et al. (2021b) propose both a *Sentence Alignment* and a *Bidirectional Word Alignment* objective inspired by MT for their AMBER model, which they train from scratch.

Among modified models, Chi et al. (2021a) propose the sentence-level cross-lingual (momentum) contrast objective for InfoXLM. However, they also emphasise the importance of MLM and TLM (translation language modelling) for token-level mutual information, casting both in information-theoretic terms. nmT5 (Kale et al., 2021) combines T5 training with a standard MT loss, which arguably targets both granularity levels. DeltaLM (Ma et al., 2021) is also an encoder-decoder model using T5-style training objectives on monolingual and parallel data. The model is initialised with InfoXLM and modified from there. Ouyang et al. (2021) propose the new objectives Cross-Attention MLM and Back-Translation MLM for ERNIE-M.

Sentence-embedding models. Models specifically targeting sentence-level tasks are typically concerned only with sentence-level alignment. One of these is multilingual Sentence-BERT (Reimers and Gurevych, 2020), an XLM-R model tuned with an English S-BERT model as a teacher. Using parallel data, the model learns to represent target language sentences similarly to the English source. This makes for strong cross-lingual alignment and good cross-lingual retrieval performance.

Among pre-trained models, LASER (Artetxe and Schwenk, 2019) is a 5-layer BiLSTM trained on machine translation, with the decoder being discarded. Its successor LASER3 (Heffernan et al., 2022) is a 12-layer Transformer model, but trained

Objectives	From Existing Model	From Scratch
Parallel, sentence-level	Multilingual S-BERT (Reimers and Gurevych, 2020); Sentence-level MoCo (Pan et al., 2021); OneAligner (Niu et al., 2022); One-pair supervised (Tien and Steinert-Threlkeld, 2022); mSimCSE supervised (Wang et al., 2022); LAPCA (Abulkhanov et al., 2023)	LASER (Artetxe and Schwenk, 2019); LASER3 (Heffernan et al., 2022); LaBSE (Feng et al., 2022); LASER3-CO (Tan et al., 2023)
Parallel, word-level	Cao et al. (2020); Weak/Strong Alignment (Wu and Dredze, 2020); Joint-Align + Norm (Zhao et al., 2021); VECO (Luo et al., 2021); WEAM (Yang et al., 2021); WordOT (Alqahtani et al., 2021); XLM-Align (Chi et al., 2021b); WAD-X (Ahmat et al., 2023)	
Parallel, both levels	Kvapilíková et al. (2020)*; InfoXLM (Chi et al., 2021a); nmT5 (Kale et al., 2021); HiCTL (Wei et al., 2021); ERNIE-M (Ouyang et al., 2021); DeltaLM (Ma et al., 2021)	ALM (Yang et al., 2020); AMBER (Hu et al., 2021b); XLM-E (Chi et al., 2022); XY-LENT (Patra et al., 2023)
Target task data	xTune (Zheng et al., 2021); FILTER (teacher model) (Fang et al., 2021); XeroAlign (Gritta and Iacobacci, 2021); Cross-Aligner (Gritta et al., 2022); X-MIXUP (Yang et al., 2022)	FILTER (student model)
Other sources	RotateAlign (Kulshreshtha et al., 2020); CoSDA-ML (Qin et al., 2020); DuEAM (Goswami et al., 2021); Syntax-augmentation (Ahmad et al., 2021); RS-DA (Huang et al., 2021); EPT/APT (Ding et al., 2022); mSimCSE NLI supervision (Wang et al., 2022)	DICT-MLM (Chaudhary et al., 2020); ALIGN-MLM (Tang et al., 2022)
Monolingual only	MAD-X (Pfeiffer et al., 2020); Adversarial & Cycle (Tien and Steinert-Threlkeld, 2022); BAD-X (Parović et al., 2022); X2S-MA (Hämmerl et al., 2022); mSimCSE unsupervised (Wang et al., 2022); LSAR (Xie et al., 2022)	RemBERT (Chung et al., 2021); XLM-R XL & XXL (Goyal et al., 2021); mT5 (Xue et al., 2021); XLM-V (Liang et al., 2023); mDeBERTaV3 (He et al., 2023);

Table 1: Proposed strategies for improved zero-shot transfer by training objectives and initialisation (training from scratch vs. modifying an existing model). *Uses only monolingual data and/or synthetic parallel data.

using a student-teacher setting, where the teacher is similar to the original LASER. This follow-up also emphasises support for lower-resource languages, training a student for each group of similar languages. By contrast, LaBSE (Feng et al., 2022) relies entirely on monolingual data and mined parallel data, but is pre-trained with standard MLM and TLM. Then, it uses translation ranking with negative sampling and additive margin softmax (Yang et al., 2019a) to train sentence embeddings.

3.2 Modified Pre-Training Schemes

Many of the proposed strategies rely on parallel data. However, several models are trained from scratch using only monolingual data while modifying specific aspects: a larger vocabulary (XLM-V, Liang et al., 2023), rebalanced pre-training vs. finetuning parameters (RemBERT, Chung et al., 2021), or using training objectives that had been tested in an English-only context, such as mDeBERTaV3 (He et al., 2023) and mT5 (Xue et al., 2021). Meanwhile, Goyal et al. (2021) improve performance by significantly scaling up model size, producing models with 3.5B and 10.7B parameters.

Like mDeBERTaV3, XLM-E (Chi et al., 2022) is pre-trained using the ELECTRA training scheme (Clark et al., 2020), but XLM-E does use both monolingual and parallel data. The later XY-LENT

(Patra et al., 2023) uses the same objectives, focusing on *many-to-many* bitexts rather than only English-centric data.

3.3 Adapter Tuning

Several other methods modify existing models using monolingual text: MAD-X (Pfeiffer et al., 2020) and BAD-X (Parović et al., 2022) are both adapter-based frameworks, combining language adapters and task adapters for improved cross-lingual transfer performance. The latter builds on the former by using 'bilingual' language adapters, which are trained on monolingual corpora of both the source and the target language. WAD-X (Ahmat et al., 2023) is another, later method that adds "word alignment adapters" using parallel text.

In a somewhat different approach, Luo et al.'s (2021) VECO uses a "plug-and-play" cross-attention module which is trained during continued pre-training, and can be used again in fine-tuning if appropriate parallel data is available.

3.4 Contrastive Learning

Contrastive learning has become popular in NLP for a variety of use cases. For cross-lingual alignment, it has also been used in several papers, since it aims to improve the similarity of positive examples and the dissimilarity of negative examples jointly. In effect, contrastive learning should help representations to fulfil our requirement number 1) as mentioned in § 2.1.

It can be used very effectively on the word level (InfoXLM, HiCTL, Wu and Dredze (2020)). For example, HiCTL (Wei et al., 2021) stands for Hierarchical Contrastive Learning, which includes both a sentence-level and a word-level contrastive loss.

Still, contrastive learning is especially popular for sentence embedding models. Examples include OneAligner (Niu et al., 2022), which targets two sentence retrieval tasks, is an XLM-R version trained on OPUS-100 data. One version uses all available English-centric pairs, another only uses the single highest-resource corpus, while setting a fixed data budget. Their training objective is based on BERT-Score, with in-batch normalisation and negatives. Abulkhanov et al. (2023), for their retrieval model LAPCA, emphasise "strong" cross-lingual alignment, mining both roughly parallel positive passages and hard negatives. mSimCSE (Wang et al., 2022) is a contrastive framework using in-batch negatives, which has multiple supervised and unsupervised settings.

Among pre-trained models, the popular LaBSe also uses contrastive learning to achieve good sentence-embeddings, and LASER3-CO (Tan et al., 2023) extends the LASER3 paradigm by adding contrastive learning to the distillation process.

3.5 Data Augmentation

Some methods create pseudo-parallel data by mining sentence pairs or machine translating monolingual text. For example, Kvapilíková et al. (2020) fine-tune XLM-100 using TLM, but they do this with 20k synthetic translation pairs, which they create for this purpose. However, there are also more complex data augmentation strategies being proposed: Yang et al.'s (2020) Alternating Language Model (ALM) uses artificially code-switched sentences constructed from real parallel data. Yang et al. (2021) propose a "cross-lingual word exchange", where representations from the source language are used to predict target language tokens.

DICT-MLM (Chaudhary et al., 2020) and ALIGN-MLM (Tang et al., 2022) both rely on a bilingual dictionary resource. DICT-MLM trains the model to predict translations of the masked tokens. ALIGN-MLM rather combines traditional MLM with an alignment loss to optimise average cosine similarity between translation pairs. CoSDA-ML (Qin et al., 2020) also uses dictionaries in a similar way, but is not trained from scratch.

3.6 Transformation of Representations

Although most models take advantage of fine-tuning techniques and deep learning, linear transformations can equally be applied to Transformer models. For instance, RotateAlign (Kulshreshtha et al., 2020) uses either dictionaries or parallel data—although parallel data is more effective—to find transformation matrices for each of the last four Transformer layers, combined with language-centering normalisation. LSAR (Xie et al., 2022) works without any parallel data, by projecting away language-specific elements of the representation space. Both Rajaee and Pilehvar (2022) and Hämmerl et al. (2023) find that mean-centering representations and forcing them to be highly isotropic can improve cross-lingual retrieval performance. And the in-batch normalisation used by Zhao et al. (2021) and (Niu et al., 2022) also targets the intuition that centering individual language-subspaces will lead to closer cross-lingual alignment.

With the fine-tuning framework X-MIXUP (Yang et al., 2022), the transformation is rather built into the fine-tuning process again, by adding MSE between source and target to the fine-tuning loss, as well as the Kullback-Leibler divergence of source and target probability distributions for classification tasks.

3.7 Tuning with Task Data

We have so far focused on methods for pre-training or continued pre-training. Some methods do fine-tuning on the task data and cross-lingual alignment in the same step, often using (translated) task data for a translate-train setting. Such methods cannot be directly compared to the zero-shot transfer setting, but they are really quite effective at achieving good transfer performance on the target tasks.

These include xTune (Zheng et al., 2021), a fine-tuning framework for cross-lingual transfer tasks which can be combined with other models. xTune also includes *consistency regularisation*, which can work without translated data. Gritta and Iacobacci's (2021) XeroAlign adds a Mean-Squared-Error (MSE) loss between the source and target sentence to the fine-tuning process. Cross-Aligner (Gritta et al., 2022) further adds a loss operating on entity level. Fang et al.'s (2021) FILTER framework first trains a teacher model in the translate-train paradigm, then a student model is trained with a self-teaching loss aimed to bridge the gap of label transfer across languages.

4 Evaluation of "Aligned" Models

There is no single metric reported by all these papers. Many report performance on XNLI (Conneau et al., 2018), in the zero-shot transfer and/or translate-train settings. We compile XNLI results in Tables 2 and 4. Cross-lingual retrieval is also popular, although the specific tasks reported vary. We show Tatoeba-36 (Artetxe and Schwenk, 2019; Hu et al., 2020) results in Table 3.

Several other tasks are reported relatively often, and we compile more results in Appendix C. For this section, we focus on XNLI, as we find that methods which work well on XNLI mostly also do well on other tasks, although the ranking changes. Unfortunately, there are a number of cases where authors report results for a task but do not use all test languages of the most commonly-used version, meaning that the average results are not comparable. We omit the results in those cases.

Additionally, App. D shows which authors provide code or model downloads for reproducibility.

Model	**Size**	**XNLI**
mBERT (Hu et al., 2020)	110M	65.4
mBERT + EPT/APT	~110M	68.4
DICT-MLM	~110M	68.6
mBERT+JointAlign+Norm	~110M	72.3
WordOT	~110M	**75.4**
AMBER	172M	71.6
XLM-R$_{base}$ + EPT/APT	~270M	75.8
XLM-ALIGN	~270M	76.2
InfoXLM$_{base}$	~270M	76.5
ERNIE-M$_{base}$	~270M	77.3
HiCTL$_{base}$	~270M	77.3
XLM-R+JointAlign+Norm	~270M	77.6
mDeBERTaV3	~276M	**79.8**
XLM-E$_{base}$	279M	76.6
mT5$_{small}$	300M	67.5
XY-LENT$_{base}$	447M	80.5
XLM-R (Hu et al., 2020)	550M	68.2
HiCTL$_{large}$	~550M	81.0
InfoXLM$_{large}$	~550M	81.4
ERNIE-M$_{large}$	~550M	82.0
XLM-R$_{large}$ + xTune	550M	**82.6**
RemBERT	575M	80.8
mT5$_{base}$	580M	75.4
VECO$_{out}$	662M	79.9
XLM-V	~750M	76.0
XLM-E$_{large}$	840M	81.3
XY-LENT$_{XL}$	2.1B	**84.8**
XLM-E$_{XL}$	2.2B	83.7
XLM-R$_{XL}$	3.5B	82.3
mT5$_{XL}$	3.7B	82.9
XLM-R$_{XXL}$	10.7B	83.1
mT5$_{XXL}$	13B	**85.0**

Table 2: Zero-shot transfer XNLI performance reported by various papers, ordered by model size. Many papers do not report exact parameter counts, so we make an estimate (~) based on the model they modify, or on hyperparameters where reported.

4.1 What works well?

The best results we see in zero-shot cross-lingual transfer are from a mix of newly-trained and modified models. WordOT, a modified mBERT with an optimal transport objective, yields the best result in its size band. In the next size group, mDeBERTaV3 performs best. This is a model trained from scratch with only monolingual data, with the ELECTRA pre-training objective and additional improvement in the form of gradient-disentangled embeddings. XLM-E, which does not have this additional ele-

Model	Size	Tatoeba
mBERT (Hu et al., 2020)	110M	38.7
mBERT + LSAR	∼110M	44.6
DICT-MLM	∼110M	47.3
LaBSe	∼110M	**95.0**
X2S-MA	∼270M	**68.1**
XLM-E$_{base}$	279M	65.0
XLM-R (Hu et al., 2020)	550M	57.3
HiCTL$_{large}$	∼550M	59.7
XLM-R + LSAR	∼550M	65.1
T&ST (unsup)	∼550M	74.2
T&ST (one-pair)	∼550M	80.4
ERNIE-M$_{large}$	∼550M	87.9
OneAligner	550M	**92.9**
mSimCSE uns.	∼550M	78.0
mSimCSE sup.	∼550M	88.3
mSimCSE NLI	∼550M	91.4
VECO$_{out}$	662M	75.1

Table 3: Tatoeba-36 performance reported by various papers, ordered by parameter counts.

ment, does markedly worse than mDeBERTaV3. Only few points behind, ERNIE-M, HiCTL and the JointAlign+Norm method sit at a near-identical performance. All modify an existing model in different ways: InfoXLM uses information theory, ERNIE-M focuses on aligning the attention parameters, whereas JointAlign+Norm looks at the output vector space. In the next group, xTune's consistency regulation proves highly effective, with ERNIE-M$_{large}$ and InfoXLM just behind.

In both zero-shot transfer and translate-train, once we cross the threshold of 1B parameters, XY-LENT$_{XL}$ is the best available method—we do not know, at this point, if this model would be outperformed by another method being scaled up. Trained from scratch, XY-LENT specifically uses a lot of parallel data that is not only English-centric, which seems to work well. XLM-R$_{XXL}$ lags behind XY-LENT$_{XL}$ and XLM-E$_{XL}$ while outperforming its own XL counterpart. Interestingly, mT5, which underperforms in smaller configurations, is competitive in XL size and does very well in XXL.

In the translate-train setting, mDeBERTaV3 again wins its size group. However, in the next larger group of models, X-MIXUP proves the most effective. It also improves mBERT's performance by a large margin. This method directly addresses representation discrepancies between different languages by linear interpolation between the hidden states of translation pairs. HiCTL, VECO, and ERNIE-M$_{large}$ come close to the performance of X-MIXUP on this task, while needing more resources. The contrastive learning approaches in these tables do well (HiCTL, InfoXLM), although they are not necessarily the most performant. We must add the caveat that not all relevant models are listed in the tables, since not all papers report the full XNLI results.

For Tatoeba, the range of results is especially large—the task has indeed been criticised for its large variability. Here, contrastive training approaches are both very common and very successful. LaBSE, OneAligner, and mSimCSE with NLI supervision attain the best overall results. LaBSE uses both negative sampling and additive margin softmax, OneAligner uses in-batch negatives, and mSimCSE follows a contrastive training approach as well, indicating the strength of these methods for the task. OneAligner additionally uses in-batch normalisation to offset the hubness problem.

4.2 What to use?

Besides the obvious conclusion that larger models usually outperform smaller ones, we recommend using (multi-directional) parallel data if available, and designing models carefully. Mined or pseudo-parallel data can fulfil that function in some cases. Use the available translated task data when optimising for a specific application. When pre-training encoder models, ELECTRA-style replaced token detection may be the way to go. Contrastive learning is popular for good reason, especially in the retrieval paradigm. Methods like OneAligner also show that models can learn from one language pair to transfer better to multiple language pairs. Representation normalisation and ensuring that language means are closer together can be very effective and make models competitive with larger ones. These could also be helpful when not enough data or resources are available for a larger training effort.

5 Multilingual Generative Models

Recently, the field has turned much attention to generative Large Language Models (LLMs). In this space, there are still fewer intentionally multilingual models (e.g., Workshop et al., 2023; Lin et al., 2022), and unfortunately they skew more heavily towards English data than models in our survey. However, we believe that multilingual generation will become increasingly relevant as applications scale. Thus, we point out several areas of future research, including how cross-lingual alignment will interact with multi- and cross-lingual generation.

There exist some efforts to benchmark multilingual generation (Asai et al., 2023; Ahuja et al., 2023; Gehrmann et al., 2022), but this presents unique challenges compared to multilingual classification, or monolingual generation tasks. Multilingual classification tasks, by contrast, are often solved considerably less well by generative methods—typically using in-context learning or zero-shot prompting—than by fine-tuned encoder models (Lin et al., 2022). Further, the zero-shot cross-lingual transfer paradigm encounters issues in generative settings, since well-aligned representations can lead to generation in the wrong language (Xue et al., 2021; Li and Murray, 2023).

Essentially, where encoder-based classification tasks must rely on language-neutral axes of representations, generative tasks must rely not only on language-neutral aspects of semantics, but also on language-specific information about the target language, such as its specific vocabulary and syntax.

5.1 Until Now: Encoder-Only Models

We have surveyed primarily encoder-only models, though we include a few encoder-decoder models. Encoder-only models transform the inputs into a latent space representation which is then used by a downstream task "head". For any tasks where the set of outputs does not depend on the language, the model needs to primarily rely on language-neutral axes of the representations. Intuitively, strong cross-lingual alignment will be helpful here, including more "radical" methods such as mean-centering language-specific subspaces.

5.2 Encoder-Decoder Models

Due to their pre-training tasks, encoders can predict the most likely tokens to fill a masked position, but encoder-decoder models are more suited to generative tasks due to their architecture. In this framework, the encoder is responsible for creating the latent space representation, while the decoder predicts the next tokens one-by-one. Conceptually, the encoder should still represent the semantics of different languages as closely aligned. However, both language-neutral and language-specific information will be present and necessary in the encodings. If we want to use established cross-lingual alignment techniques in generative models, the encoder is the most natural target for them.

The decoder, meanwhile, must learn to focus more on language-specific information in order to generate tokens in the target language. It is therefore likely that, e.g., mean-centering the language subspaces in the encoder output would be harmful. That said, language-neutral semantic information from the encoder must also be taken into account. We suggest training for both cross-lingual alignment *and* language-specificity at the decoding step. An approach such as OneAligner, where one or a few target languages are used for training along with the source language, could help to remain resource-efficient. Note that this is different from zero-shot cross-lingual transfer, where only one (source) language is used for training.

5.3 Decoder-Only Models

Many recent LLMs (Brown et al., 2020; Touvron et al., 2023) are decoder-only models, meaning they use the decoder architecture throughout the model and have no separate encoder layers. Thus, there is no one obvious point at which cross-lingual alignment should be greatest. However, any training scheme that can help an encoder-decoder model to focus on language-neutral *and* language-specific information—at the relevant times—could help here as well. Since LLMs are hard to fine-tune, this should be combined with approaches like LoRA (Hu et al., 2021a)—or integrated in a multilingual pre-training scheme. Assuming the output language should be based on the input language and/or explicit instructions in the input, this should also be explored in instruction tuning datasets.

6 Conclusions

We have surveyed the literature of cross-lingual alignment methods and compiled results in some of the most popular evaluation tasks. Our analysis confirms the strengths of methods such as contrastive learning and ELECTRA-style pre-training, as well as the importance of using available parallel data. We further collated an overview of which authors provide code or model downloads for reproducibility. Going forward, new challenges present themselves with respect to multilingual generative models: Simply maximising cross-lingual alignment can lead to wrong-language generation. We thus call for methods that effectively trade-off cross-lingual semantic information with language-specific axes, allowing models to generate fluent and relevant content in many languages.

Limitations

Evaluation Tasks

XNLI is reported in a plurality of papers in our survey, more often than any other single task. The relative prevalence of XTREME (Hu et al., 2020) means that this and several other tasks, including UD-POS, MLQA, PAWS-X, Tatoeba, BUCC2018 and NER, are frequently reported in specific configurations. Most of these tasks are also popular on their own. Unfortunately, despite this, many papers do not report results for the full range of "standard" target languages, a problem that is more common the more target languages appear in a task. This particularly limits our ability to compare models across lower-resource languages, and we strongly urge researchers to report results for all standard languages when evaluating on a task.

We chose here to compile results from XTREME, focusing on XNLI and Tatoeba, since the former is so common among fine-tuning tasks, and the latter functions under a different paradigm. We simply omit papers that do not report these results, or do not report all target languages, from our tables. We do not calculate such missing results ourselves. Thus, our picture of model performance is admittedly narrow, with multiple kinds of tasks—particularly word-level ones—missing entirely, and many models missing from the ranking, even if they perform well on the tasks they do report.

Bilingual vs. Multilingual Alignment

Since we are talking about highly multilingual models, we are implicitly concerned with multilingual cross-lingual alignment. Many-to-many cross-lingual alignment makes for a hard optimisation problem, as mentioned in § 2.2. However, most of the parallel data involved in (re-)aligning the models or measuring transfer performance are parallel with English. Thus, in practice, bilingual alignments with English as a pivot language are the most common. To the extent that alignment in the models is measured (see § 2.4), this is typically also done between English and some target language, and less often between a non-English source and non-English target language. These circumstances significantly limit the training and evaluation of many-to-many cross-lingual alignment.

Multimodality

Alignment of representations between modalities adds further complexities compared to cross-lingual alignment. We omit multimodal models from this survey, but note that cross-modal alignment should be similarly examined in future work.

References

Dmitry Abulkhanov, Nikita Sorokin, Sergey Nikolenko, and Valentin Malykh. 2023. Lapca: Language-agnostic pretraining with cross-lingual alignment. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2098–2102, New York, NY, USA. Association for Computing Machinery.

Judit Acs, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2023. Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*, pages 1–40.

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. 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 4538–4554, Online. Association for Computational Linguistics.

Ahtamjan Ahmat, Yating Yang, Bo Ma, Rui Dong, Kaiwen Lu, and Lei Wang. 2023. Wad-x: Improving zero-shot cross-lingual transfer via adapter-based word alignment. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(9).

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. *CoRR*, abs/2303.12528.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919, Punta Cana, Dominican Republic. 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.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. *CoRR*, abs/2305/14857.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, (NeurIPS)*.

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

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. DICT-MLM: Improved multilingual pre-training using bilingual dictionaries. *CoRR*, abs/2010.12566.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. 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 3418–3430, Online. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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

Kunbo Ding, Weijie Liu, Yuejian Fang, Weiquan Mao, Zhe Zhao, Tao Zhu, Haoyan Liu, Rong Tian, and Yiren Chen. 2022. A simple and effective method to improve zero-shot cross-lingual transfer learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4372–4380, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. FILTER: An enhanced fusion method for cross-lingual language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12776–12784.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez Beltrachini, Leonardo F . R.

Ribeiro, Lewis Tunstall, Li Zhang, Mahim Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. 2017. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. 2022. CrossAligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4048–4061, Dublin, Ireland. Association for Computational Linguistics.

Milan Gritta and Ignacio Iacobacci. 2021. XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 371–381, Online. Association for Computational Linguistics.

Mikhael Gromov. 1999. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152 of *Progress in Mathematics*. Birkhäuser Boston Inc.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2022. Combining static and contextualised multilingual embeddings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2316–2329, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021b. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mihir Kale, Aditya Siddhant, Rami Al-Rfou, Linting Xue, Noah Constant, and Melvin Johnson. 2021. nmT5 - is parallel data still relevant for pre-training massively multilingual 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 2: Short Papers)*, pages 683–691, Online. Association for Computational Linguistics.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. *CoRR*, abs/2301.10472.

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

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. 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 3980–3994, Online. Association for Computational Linguistics.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736.

Tong Niu, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2022. OneAligner: Zero-shot cross-lingual transfer with one rich-resource language pair for low-resource sentence retrieval. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2869–2882, Dublin, Ireland. Association for Computational Linguistics.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of* the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Barun Patra, Saksham Singhal, Shaohan Huang, Zewen Chi, Li Dong, Furu Wei, Vishrav Chaudhary, and Xia Song. 2023. Beyond English-centric bitexts for better multilingual language representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15354–15373, Toronto, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. 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.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1286–1304, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. CoSDA-ML: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, page 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

Sara Rajaee and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490, Dubrovnik, Croatia. Association for Computational Linguistics.

Henry Tang, Ameet Deshpande, and Karthik Narasimhan. 2022. ALIGN-MLM: Word embedding alignment is crucial for multilingual pre-training. *CoRR*, abs/2211.08547.

Chih-chan Tien and Shane Steinert-Threlkeld. 2022. Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8696–8706, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel

Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2022. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5617–5633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9386–9393.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, page 5370–5378. International Joint Conferences on Artificial Intelligence Organization.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019b. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, and Shijin Wang. 2021. Bilingual alignment pre-training for zero-shot cross-lingual transfer. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 100–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, et al. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. 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 3403–3417, Online. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

A Inclusion of Papers

The majority of this literature review was done in early 2023, and we found papers by searching the ACL Anthology, Semantic Scholar, and arXiv.org, as well as following the citation graph. We focused on the initial search terms “zero-shot cross-lingual transfer” and “cross-lingual alignment”.

B Further Models Explained

We add here brief explanations of additional models which we omitted from the main body. These are methods that are either quite similar to ones described in Section 3, or did not fit well into one of the larger categories which we exemplify. Several of these papers do not report results for the tasks we looked at.

B.1 More Word- and Sentence-Level methods

Pan et al. (2021) propose a sentence-level momentum contrast objective, which they combine with TLM to train mBERT. This seems to be a similar idea to InfoXLM. Unfortunately, the paper does not include all languages on XNLI or MLQA results.

B.2 Further Sentence-embedding models

Tien and Steinert-Threlkeld (2022) propose two different methods, one supervised by a single language pair not unlike OneAligner, and one unsupervised approach. Their unsupervised approach uses an adversarial loss encouraging language distributions to become indistinguishable, and a Cycle loss to keep them from degenerating. In both cases, they freeze the parameters of XLM-R and only train a linear mapping. Their one-pair supervised model is competitive with OneAligner on BUCC2018, but lags further behind on Tatoeba-36, which contains more languages.

B.3 More Tuning with Task Data

DuEAM (Goswami et al., 2021) uses data from the XNLI dataset while targeting semantic textual similarity and bitext mining tasks. The objectives used are Word Mover's Distance and a translation mining loss. The model performs reasonably well but does not reach the performance of S-BERT.

B.4 More Data Augmentation

RS-DA (Huang et al., 2021) is "randomised smoothing with data augmentation"—a kind of robustness training during fine-tuning, using synonym sets to create the augmented (English) data. Ding et al. (2022) build on the idea of robust regions and synonym-based data augmentation, adding three objectives to 'push' and 'pull' the embeddings and attention matrices appropriately (EPT/APT in Tables 2 and 5). This model performs well on PAWS-X but does not stand out on XNLI.

B.5 Other Approaches

X2S-MA (Hämmerl et al., 2022) is an approach using monolingual data to first distill static embeddings from XLM-R, which are then aligned post-hoc and used to train the model for similarity with the aligned static embeddings. This model works well on Tatoeba.

Ahmad et al. (2021), meanwhile, augment mBERT with syntax information using dependency parses. They employ a graph attention network to learn the dependencies, which they then mix using further parameters with some attention heads in each layer.

C More Task Results

Table 7 compiles BUCC2018 (Zweigenbaum et al., 2018) performance, as implemented by Hu et al. (2020). Tables 5 and 6 show zero-shot transfer and translate-train results for UD-POS (Zeman et al., 2019), PAWS-X (Yang et al., 2019b) and MLQA (Lewis et al., 2020).

Results on zero-shot transfer overall show a similar picture to XNLI, although details change. For example, XLM-ALIGN's performance stands out on UD-POS but is "only" competitive on the other tasks. HiCTL, meanwhile, is fairly competitive in zero-shot XNLI performance but falls a bit further behind in Table 5. The authors of mDeBERTaV3 do not report any of these other tasks, leaving XLM-ALIGN, XLM-E$_{base}$, and ERNIE-M$_{base}$ to take the top spots: they all perform well on these three tasks but alternately take the lead.

In the translate-train setting (Table 6), VECO$_{in}$ performs best on all three tasks, with HiCTL$_{large}$ on par for PAWS-X but not UD-POS or MLQA. For XNLI, the best translate-train performance was attained by X-MIXUP, which still does well on these tasks. Again, overall trends are very similar as for the XNLI task.

Finally, BUCC2018 (Table 7) also conveys a similar picture as Tatoeba, although the variation is smaller, likely due the larger datasets and smaller selection of relatively high-resource languages. mSimCSE with NLI supervision performs best on this task—it also proved effective on Tatoeba-36. OneAligner is the second most effective on BUCC2018, with Tien and Steinert-Threlkeld's (2022) one-pair supervision a close third.

D Reproducibility

In order to reproduce a method, or apply it to a new use case, detailed instructions and ease of reuse are vital. Providing implementation code is the most straightforward way to ensure that *all* necessary details are conveyed to a reader, and they do not waste time reimplementing them. Similarly, model downloads save time and make further experimentation much easier. The larger the model in question, the more important model downloads become, since re-training them requires more time, effort, and compute.

Model	Size	XNLI
mBERT (Hu et al., 2020)	110M	74.6
mBERT + X-MIXUP	110M	**78.8**
InfoXLM$_{base}$	$\sim$270M	80.0
ERNIE-M$_{base}$	$\sim$270M	80.6
mDeBERTaV3	$\sim$276M	**82.2**
mT5$_{small}$	300M	72.0
XY-LENT$_{base}$	447M	82.9
XLM-R$_{large}$ + xTune	$\sim$550M	82.6
FILTER	$\sim$550M	83.6
FILTER + Self-teaching	$\sim$550M	83.9
ERNIE-M$_{large}$	$\sim$550M	84.2
HiCTL$_{large}$	$\sim$550M	84.5
XLM-R$_{large}$ + X-MIXUP	550M	**85.3**
mT5$_{base}$	580M	79.8
VECO$_{in}$	662M	84.3
XY-LENT$_{XL}$	2.1B	**87.1**
XLM-R$_{XL}$	3.5B	85.4
mT5$_{XL}$	3.7B	85.3
XLM-R$_{XXL}$	10.7B	86.0
mT5$_{XXL}$	13B	**87.1**

Table 4: Translate-train XNLI performance reported by various papers, ordered by model size.

In Table 8, we list all papers that provide their code, a model download, or both. Some of these are well documented, some not so much. Some are well-maintained, some not at all. We did not test the provided code and links, simply checked that they are online and contain what looks to be the promised artifacts. Papers where we did not find any artifacts are omitted.

Model	**Size**	**UDPOS**	**PAWS-X**	**MLQA (F1)**
Zero-shot transfer				
mBERT (Hu et al., 2020)	110M	71.5	81.9	61.4
mBERT + Syntax augm.	∼110M	–	84.3	60.3
mBERT + EPT/APT	∼110M	–	**86.2**	–
DICT-MLM	∼110M	71.6	84.8	–
XLM-R$_{base}$ + EPT/APT	∼270M	–	87.1	–
XLM-ALIGN	∼270M	**76.0**	86.8	68.1
ERNIE-M$_{base}$	∼270M	–	–	**68.7**
HiCTL$_{base}$	∼270M	71.4	84.5	65.8
InfoXLM$_{base}$	∼270M	–	–	68.1
XLM-E$_{base}$	279M	75.6	**88.3**	68.3
mT5$_{small}$	300M	–	82.4	54.6
XY-LENT$_{base}$	447M	–	89.7	71.3
XLM-R (Hu et al., 2020)	550M	73.8	86.4	71.6
HiCTL$_{large}$	∼550M	74.8	87.5	72.8
ERNIE-M$_{large}$	∼550M	–	89.5	73.7
InfoXLM$_{large}$	∼550M	–	–	73.6
XLM-R$_{large}$ + xTune	550M	**78.5**	**89.8**	**74.4**
RemBERT	575M	76.5	87.5	73.1
mT5$_{base}$	580M	–	86.4	64.6
VECO$_{out}$	662M	75.1	88.7	71.7
XLM-V	∼750M	–	–	66.0
XL-LENT$_{XL}$	2.1B	–	91.0	**75.4**
XLM-R$_{XL}$	3.5B	–	–	73.4
mT5$_{XL}$	3.7B	–	89.6	73.5
XLM-R$_{XXL}$	10.7B	–	–	74.8
mT5$_{XXL}$	13B	–	90.0	**76.0**

Table 5: Zero-shot transfer UD-POS, PAWS-X, and MLQA performance reported by various papers, ordered by model size. Many papers do not report exact parameter counts, so we make an estimate based on the model they modify, or based on hyperparameters where reported. We mark the estimates with a tilde (∼). We draw dashed lines between models of markedly different sizes.

Model	**Size**	**UDPOS**	**PAWS-X**	**MLQA (F1)**
Translate-train				
mBERT (Hu et al., 2020)	110M	–	86.3	65.6
mBERT + X-MIXUP	110M	76.5	**89.7**	**69.0**
mT5$_{small}$	300M	–	79.9	64.3
XY-LENT$_{base}$	447M	–	92.4	–
XLM-R large + xTune	$\sim$550M	78.5	89.8	75.0
FILTER	$\sim$550M	76.2	91.2	75.8
FILTER + Self-teaching	$\sim$550M	76.9	91.5	76.2
ERNIE-M$_{large}$	$\sim$550M	–	91.8	–
HiCTL$_{large}$	$\sim$550M	76.8	**92.8**	74.4
XLM-R$_{large}$ + X-MIXUP	550M	78.4	91.8	76.5
mT5$_{base}$	580M	–	89.3	75.3
VECO$_{in}$	662M	**79.8**	**92.8**	**77.5**
XL-LENT$_{XL}$	2.1B	–	**92.6**	–
mT5$_{XL}$	3.7B	–	91.0	75.1
mT5$_{XXL}$	13B	–	91.5	76.9

Table 6: Translate-train UD-POS, PAWS-X, and MLQA performance reported by various papers, ordered by model size. Many papers do not report exact parameter counts, so we make an estimate based on the model they modify, or based on hyperparameters where reported. We mark the estimates with a tilde ($\sim$). We draw dashed lines between models of markedly different sizes.

Model	**Size**	**BUCC**
mBERT (Hu et al., 2020)	110M	56.7
LaBSe	$\sim$110M	**89.7**
LAPCA-LM$_{base}$	$\sim$270M	71.3
XLM-R$_{large}$ (Hu et al., 2020)	550M	66.0
HiCTL$_{large}$	$\sim$550M	68.4
Tien and Steinert-Threlkeld (2022) (unsup)	$\sim$550M	82.4
Tien and Steinert-Threlkeld (2022) (one-pair)	$\sim$550M	89.6
LAPCA-LM$_{large}$	$\sim$550M	83.5
OneAligner	550M	90.5
mSimCSE uns.	$\sim$550M	87.5
mSimCSE sup.	$\sim$550M	88.8
mSimCSE NLI	$\sim$550M	**95.2**
Kvapilíková et al. (2020)	$\sim$570M	75.8
VECO$_{out}$	662M	85.0

Table 7: BUCC performance reported by various papers, ordered by model size.

Model Name	**Code Available**	**Model Download**
Syntax Augmented mBERT (Ahmad et al., 2021)	yes	no
LASER (Artetxe and Schwenk, 2019)	yes	yes, fairseq
XLM-Align (Chi et al., 2021b)	yes	yes, HF
InfoXLM (Chi et al., 2021a)	yes	yes, HF
RemBERT (Chung et al., 2021)	no	yes, HF
EPT/APT (Ding et al., 2022)	yes	no
FILTER (Fang et al., 2021)	yes	no
LaBSE (Feng et al., 2022)	no	yes, TFH, HF
X2S-MA (Hämmerl et al., 2022)	yes	no
XLM-R$_{XL/XXL}$ (Goyal et al., 2021)	yes	yes, fairseq, HF
XeroAlign (Gritta and Iacobacci, 2021)	yes	no
CrossAligner (Gritta et al., 2022)	yes	no
mDeBERTaV3 (He et al., 2023)	yes	yes, HF
LASER3 (Heffernan et al., 2022)	yes	yes, fairseq
XLM-V (Liang et al., 2023)	no	yes, HF
XGLM (Lin et al., 2022)	no	yes, fairseq, HF
VECO (Luo et al., 2021)	no*	yes, fairseq
ERNIE-M (Ouyang et al., 2021)	yes	yes, HF
BAD-X (Parović et al., 2022)	yes	yes, AdapterHub
MAD-X (Pfeiffer et al., 2020)	no	yes, AdapterHub
Multilingual S-BERT (Reimers and Gurevych, 2020)	yes	yes, HF
ALIGN-MLM (Tang et al., 2022)	yes	no
Tien and Steinert-Threlkeld (2022)	yes	no
mSimCSE (Wang et al., 2022)	yes	yes, HF
Wu and Dredze (2020)	yes	no
LSAR (Xie et al., 2022)	yes	no
mT5 (Xue et al., 2021)	yes	yes, custom, HF
X-MIXUP (Yang et al., 2022)	yes	no
JointAlign + Norm (Zhao et al., 2021)	yes	yes
xTune (Zheng et al., 2021)	yes	no

Table 8: A list of those surveyed papers that provide code and/or model downloads. We do not test the provided code, only making sure it remains online at time of writing. We sort by first author last name. *VECO has a repository online that includes only fine-tuning code.