Understanding Cross-Lingual Alignment—A Survey

Katharina Hämmerl[1,2] and **Jindřich Libovický**[3] and **Alexander Fraser**[2,4]

[1]Center for Information and Language Processing, LMU Munich, Germany
haemmerl[at]cis.lmu.de
[2]Munich Centre for Machine Learning (MCML), Germany
[3]Faculty of Mathematics and Physics, Charles University, Czech Republic
[4]Technical University of Munich, Germany

Abstract

Cross-lingual alignment, the meaningful similarity of representations across languages in multilingual language models, has been an active field of research in recent years. We survey the literature of techniques to improve cross-lingual alignment, providing a taxonomy of methods and summarising insights from throughout the field. We present different understandings of cross-lingual alignment and their limitations. We provide a qualitative summary of results from a large number of surveyed papers. 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, and argue that an effective trade-off between language-neutral and language-specific information is key.

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 alignment* of those representations is an underlying assumption for zero-shot cross-lingual transfer. Once the model has learned to do, e.g., a classification task in a source language, then if the representation of a target language is "aligned" to that of the source language, the model should also be able to classify target language items.

As we define it, cross-lingual alignment means that words or sentences with similar semantics: 1) Are more similar in the representation space than words or sentences with dissimilar semantics. This way of looking at alignment can be defined in "weak" or "strong" terms (see § 2). 2) Allow a prediction head trained on a source language to recognise the relevant patterns in the target language. We argue that this is related to the first view but, importantly, is more nuanced. This allows us to discuss the literature in a new way.

These two criteria are not guaranteed to be fulfilled through unsupervised pre-training, thus motivating various efforts to improve cross-lingual alignment. We survey important papers in this area between 2019 and 2023. These works propose new training objectives, new pre-trained models, contrastive fine-tuning, or post-hoc adjustments of the embedding space. The vast majority of the methods were developed for and applied to multilingual encoder models, chiefly XLM-R and mBERT.

In this paper, we provide

- a definition of cross-lingual alignment under both views, how it is typically measured—and weaknesses of these measurements—, and an important critique of "strong" alignment as stated by the first view (§ 2);
- a taxonomy of methods to increase alignment taken from a comprehensive survey of the literature (§ 3);
- a summary of findings and where the field currently stands (§ 4);
- and finally a discussion of ongoing and future research in this area as the field focuses more on generative models (§ 5). As we discuss, generative models pose new challenges for cross-lingual alignment, since we need to trade-off language-neutral with language-specific elements in new ways.

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), or as in our

Findings of the Association for Computational Linguistics: ACL 2024, pages 10922–10943
August 11-16, 2024 ©2024 Association for Computational Linguistics

case, 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 the two main views below.

View I. Similar meanings across languages have more similar representations than dissimilar meanings do. This view is particularly salient for tasks using, e.g., cosine similarity of representations. It relies on the embeddings as a whole being distributed "well". There are "weak" and "strong" definitions of cross-lingual alignment that focus on this view (cf. Roy et al., 2020; Gaschi et al., 2022; Abulkhanov et al., 2023, inter alia).

Weak alignment requires only that the nearest *target-language* neighbour of a word or sentence representation be its target-language translation. If (u_i, v_i) is a pair of encoder representations corresponding to a pair of equivalent words (or sentences) from the languages $L1$ and $L2$, s is a similarity function, and N is the number of translation pairs in a corpus:

$$\text{Align}_{\text{weak}} = s(u_i, v_i) > \max_{j \in N, j \neq i} \{s(u_i, v_j)\}. \quad (1)$$

We can then measure the proportion of pairs in the corpus where the property applies.

Strong alignment requires that the nearest neighbours *in general* of a word (or sentence) representation be its translations, *and* therefore that representations of dissimilar source-language words be more distant than the target-language translation. In terms of the same parallel corpus as above, this can be expressed as:

$$\text{Align}_{\text{strong}} = s(u_i, v_i) > \max_{j \in N, j \neq i} \{s(u_i, \boldsymbol{u}_j)\}. \quad (2)$$

Note the bolding for emphasis. These equations refer to bilingual corpora, but can be applied to multiple language pairs in order to measure alignment across a set of languages.

Strong alignment inherently requires greater distance of dissimilar meanings within a language. That said, weak alignment also benefits from increasing the distance to representations of dissimilar meanings within a language, since accurate retrieval can otherwise be hindered by hubness issues. The high anisotropy observed in Transformer models (Ethayarajh, 2019; Gao et al., 2019) may contribute to hubness issues and make it harder to distinguish similar from dissimilar representations.

To see why cross-lingual alignment, particularly "strong" alignment, under this view is hard to achieve, consider the *isomorphism assumption* as stated for cross-lingual static embeddings (cf. Vulić et al., 2020). This states that spaces of both languages should have (roughly) the same shape, measured by, for example, *relational similarity* (Vulić et al., 2020) or *eigenvector similarity* (Søgaard et al., 2018).

The isomorphism 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) emphasise that undertraining contributes significantly to non-isomorphism in static embeddings, and this may well apply to contextual models.

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. This may be intractable without also removing valuable contextual and language-specific information.

View II. 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 cross-lingual alignment in terms of simple measures such as cosine similarity, the prediction head works with the full encoder representations as its input, and can (potentially) use subspaces to that effect. Given the many constraints on the representations, it is actually very difficult to attain "cross-lingual alignment", particularly strong alignment, under View I. However, we can consider language-specific subspaces, as well as features pertaining to specific tasks. For example, some works find subspaces for morphosyntactic aspects (Hewitt and Manning, 2019; Acs et al., 2023), others find directions encoding token frequency (Rajaee and Pilehvar, 2022; Puccetti et al., 2022).

Chang et al. (2022) separate types of subspaces 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.

The prediction head is usually a linear layer added after the last encoder layer, with a softmax output. Its weights are learned during fine-tuning. For cross-lingual transfer to work, the prediction head must place more weight on features that are relevant to the task than on features that are only relevant to the specific language. Outright reducing the language component in the full encoder representations (as under View I) works to achieve this goal. However, under the second view we additionally consider (subspace) projections of the embedding space $\mathbf{E}$: We conjecture that for any task T, there exists a linear projection such that the language component $\mathbf{E}_L$ is reduced and task-relevant features $\mathbf{E}_T$ are emphasised. Cross-lingual transfer for the task should succeed if:

$$\exists \text{Proj}(\mathbf{E}) \rightarrow |\mathbf{E}_T| > |\mathbf{E}_L| \tag{3}$$

and if the prediction head is able to find such a projection. "Strong alignment" (Equation 2) implies that Equation 3 is also true, but the inverse is not the case. This second view is particularly salient for fine-tuning tasks, e.g., classification or question answering.

In summary, the first view of cross-lingual alignment implies a complex optimisation problem (similar to CLWEs), as models need to reconcile many variables in order to align representation well. As we discuss, this may well be intractable or trade off too much valuable information. The second view explains why cross-lingual transfer works anyway. Both views are of course related, and pursue the same overall goals. Importantly, though, we argue that the first view—focusing on (cosine) similarity between pairs of full vector representations—may be overly simplistic and lose sight of details.

2.2 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:

Cosine similarity is often used as the similarity function s in Equations 1 and 2. Using the notation from above, the cosine similarity of a translation pair is:

$$\mathrm{S_C}(\vec{u_i}, \vec{v_i}) = \cos(\theta) = \frac{\vec{u_i} \cdot \vec{v_i}}{\|\vec{u_i}\| \|\vec{v_i}\|}. \tag{4}$$

Note that the average cosine similarity in the space can be quite high (Ethayarajh, 2019; Rajaee and Pilehvar, 2022), so it is advisable to normalise similarities by the average cosine similarity in order to differentiate values more clearly.

Word or sentence retrieval tasks. Using cosine similarity (Equation 4), we can then count the proportion of pairs in the bilingual corpus where Equations 1 or 2 apply as a measure of cross-lingual alignment in the space. Retrieval tasks are essentially formulated in this way, since they measure whether the closest retrieved element is the correct one. The tasks may consist only of matched pairs (e.g., Tatoeba), or include "decoy" elements that do not have an equivalent (e.g., BUCC2018). Cosine similarity is commonly used, or an adjusted retrieval score such as CSLS (Lample et al., 2018).

Zero-shot transfer after fine-tuning, similarly to retrieval tasks, is both an aim in itself and a proxy for how well-aligned the representations are. This seems to be the main way that the more complex second view of cross-lingual alignment is translated into metrics. Interventions before and/or during fine-tuning have been shown to improve transfer performance. The metrics 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.

Language identification is sometimes used (e.g., Libovický et al., 2020) to measure language-specific elements of the representations. In this thinking, if a language classifier trained on the output representations performs poorly, then the model representations are highly language-neutral. This is actually an even stricter goal than "strong alignment". However, it neglects that the representations can have both language-neutral and language-specific areas, and that some language-specific information is necessary.

Visualisation. 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, 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.

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); LaBSE (Feng et al., 2022); LAPCA (Abulkhanov et al., 2023)	LASER (Artetxe and Schwenk, 2019); LASER3 (Heffernan et al., 2022); LASER3-CO (Tan et al., 2023)
Parallel, word-level	Cao et al. (2020); Wu and Dredze (2020); Joint-Align + Norm (Zhao et al., 2021); VECO (Luo et al., 2021); WEAM (Yang et al., 2021); XLM-Align (Chi et al., 2021b); WAD-X (Ahmat et al., 2023); Efimov 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); WordOT (Alqahtani 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.

3 Taxonomy of Alignment Strategies

We report on methods for improving zero-shot transfer and increasing cross-lingual alignment. Table 1 shows all included papers, organised by initialisation and data requirements of their proposed objectives. In this section, we discuss each category with examples. Additionally, we describe some strategies which are applied across different data requirements and initialisations. We note how methods fit with the two views introduced in § 2. Some methods are left out here because they are less relevant to the overall analysis, though we explain them in Appendix A for completeness.

Inclusion of Papers. We collected papers for this survey over several search iterations. We found relevant papers by searching the ACL Anthology, Semantic Scholar, and arXiv.org, as well as following the citation graph. The initial search terms were "zero-shot cross-lingual transfer" and "cross-lingual alignment". We excluded papers where we could not find a PDF version online, and papers focusing on static cross-lingual word embeddings. We prioritised papers focused on a general notion of cross-lingual alignment over papers applying the concept to a single specific task.

3.1 Objectives using Parallel Data

First, we discuss models using external parallel data—sentence-parallel or word-parallel. These make up 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, typically either masked language modelling (MLM), or minimising the distance from the original model weights (e.g., Cao et al., 2020). In some cases, a newly proposed alignment objective is combined with an existing objective such as translation language modelling (TLM).

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. To regularise the model, they minimise the distance to its initialisation. This paper clearly takes a similarity-based view of cross-lingual alignment, evaluating on word retrieval tasks but also some languages of XNLI. A number of works have followed their lead in approaching cross-lingual alignment in this way. For instance, 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". 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.

Alqahtani et al. (2021), meanwhile, formulate cross-lingual word alignment as an optimal transport problem. The mechanism of optimal transport means this is closer to the projection-based view of alignment. Their input data consists of parallel sentences, but as part of their training process they still focus on finding matched words between the source and target sentences.

Sentence-level alignment. 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 method mostly focuses on similarity scores and achieves good cross-lingual retrieval performance. 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.

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 using a student-teacher setting, where the teacher is similar to the original LASER. The follow-up also emphasises lower-resource languages, training a student for each group of similar languages.

3.2 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 increase the similarity of positive examples and the dissimilarity of negative examples jointly. In effect, contrastive learning should improve hubness issues and increase strong cross-lingual alignment as per the first view in § 2.1.

Contrastive learning can be used very effectively on the word level (see 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.

Nevertheless, contrastive learning is especially popular for sentence embedding models. Examples include OneAligner (Niu et al., 2022), which targets two sentence retrieval tasks, and 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 uses BERT-Score for similarity scoring, with in-batch normalisation and negatives.

Abulkhanov et al. (2023), for their retrieval model LAPCA, also aim for "strong" cross-lingual alignment, mining both roughly parallel positive passages and hard negatives. mSimCSE (Wang et al., 2022) is another contrastive framework using in-batch negatives, with multiple supervised and unsupervised settings. LaBSe also uses contrastive learning to achieve good sentence-embeddings, and among pre-trained models, LASER3-CO (Tan et al., 2023) extends LASER3 by adding contrastive learning to the distillation process.

3.3 Modified Pre-Training Schemes

Although many strategies rely on parallel data, 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. fine-tuning parameters (RemBERT, Chung et al., 2021). Several use training objectives that had been tested in an English-only context, such as mDeBERTaV3 (He et al., 2023) and mT5 (Xue et al., 2021). mDeBERTaV3 additionally uses *gradient-disentangled embeddings*. Meanwhile, Goyal et al. (2021) significantly scale 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, but focuses on *many-to-many* bitexts rather than only English-centric data.

3.4 Adapter Tuning

Another group of methods use adapters to modify existing models. 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.

3.5 Data Augmentation

A few 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

These methods directly transform the representation space, meaning they lean towards the subspace view of cross-lingual alignment. They may still be influenced by View I, for example in using bilingual dictionaries. 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. The in-batch normalisation used by Zhao et al. (2021) and Niu et al. (2022) is based on 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 in a translate-train setting. It adds 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 very effective for 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 designed to bridge the gap of label transfer across languages.

4 What We Do and Don't Know

In this section, we discuss broad findings both from the alignment methods summarised above and from recent related analysis papers. Future work in this area should follow from open questions.

A selection of task results achieved by the alignment methods can be found in Appendix B. Additionally, Appendix C shows which authors provide code or model downloads for reproducibility.

Contrastive training works. Contrastive training is effective for cross-lingual transfer (as well as for other problems), presumably because it reduces the hubness problem and forces models to differentiate representations. It is especially popular for retrieval-based tasks and sentence-level models.

Pre-training is not everything. Whether a model is newly-trained or modified from a pre-trained model does not appear to determine cross-lingual performance (cf. Appendix B). As models get bigger, only newly pretrained models are available, but at smaller sizes, models *modified* for cross-lingual alignment perform very well.

Related languages are more aligned. Closely related languages are more aligned within the models. In keeping with the many factors encoded by the representations, it makes sense that as more elements differ between languages, they also become more distant in the representation space. Some realignment strategies proved effective at reducing the gap to distant languages, but the pattern remains strong even then. Accordingly, some papers applying cross-lingual transfer to specific tasks leverage groups of related languages (e.g., Zeng et al., 2023). The many-to-many translation model M2M-100 (Fan et al., 2021) similarly makes use of this by grouping training languages.

Use available parallel data. While zero-shot cross-lingual transfer is an interesting research problem and can improve pre-trained models, practitioners aiming to deploy the best models should make use of available parallel data. Models in the translate-train setting consistently outperform those in the zero-shot setting (cf. Appendix B), and even a small number of training examples may help.

To what extent is strong alignment necessary? In terms of View I, 'strong' alignment is often seen as desirable, but we question whether this should be a main goal. As discussed in § 2, there is a risk of trading off language-specific information. As for downstream tasks, most cross-lingual retrieval tasks query only the target language space, ignoring the language component. Thus, strong alignment may not be necessary for these tasks. LAReQA (Roy et al., 2020) does test for strong alignment specifically, but it is the exception rather than the rule. In tasks with a prediction head, strong alignment would ensure that the condition of View II is met. Indeed, Gaschi et al. (2023) show that strong alignment correlates with better performance on downstream tasks, though they specifically look at classification tasks, not retrieval tasks. However, this does not tell us if strong alignment is *necessary* for improved transfer performance on these tasks, only that it is correlated.

Try using more source languages. There is relatively little work attempting to fine-tune models on two (or more) annotated source languages for zero-shot cross-lingual transfer. The main paradigm of zero-shot cross-lingual transfer, and a large number of relevant tasks, works with a singular source language (largely English). However, it stands to reason that learning the target task in two or more annotated source languages would encourage the model to attend more to language-agnostic components as per View II, since this would demonstrate the task as orthogonal to the source language itself. X-MIXUP and xTune are examples of fine-tuning on multiple languages using the translate-train data in the target task and are found to be quite effective. However, the translate-train approach does require fine-tuning data in all target languages, which may not always be available.

5 Multilingual Generative Models

Recently, the field has turned much attention to generative Large Language Models (LLMs). We believe multilingual capabilities of generative models will be more and more important as applications scale. Thus, we point out several areas of future research, including how cross-lingual alignment may interact with multi- and cross-lingual generation.

5.1 Multilinguality As An Afterthought?

The flagship models (e.g., Touvron et al., 2023; Jiang et al., 2023; OpenAI et al., 2024) in the generative space tend to be English-centric, with some multilingual capability. However, the proportion of multilingual data is often relatively small and a performance drop-off for mid- and low-resource languages is noticeable (e.g., Ahia et al., 2023).

Due to the significant costs of pre-training, only few models have been trained to be intentionally multilingual from the start (e.g., BLOOM, Workshop et al., 2023; XGLM, Lin et al., 2022). Unfortunately, even these skew more heavily towards English data than, e.g., XLM-R, and they tend to underperform compared to, e.g., newer GPT, Llama or Mistral models. mGPT (Shliazhko et al., 2024) has a more balanced language distribution—the model is comparable with XGLM on few-shot classification tasks, but the evaluation on generative tasks is somewhat limited.

Given this research landscape, we see room to 1) explore why efforts like XGLM or BLOOM have been less effective than retrainings of English-centric models, beyond the oft-cited 'Curse of Multilinguality' (Chang et al., 2023); and 2) be even more intentional about balancing training data, and including multiple languages. This will necessarily include further data collection efforts as well as additional work on effective multilingual pretraining.

5.2 Cross-Linguality in Generative Models

However, numerous efforts have recently extended open models to new languages. Examples include EEVE (Kim et al., 2024) for Korean, PolyLM (Wei et al., 2023) with a focus on Chinese, Andersland (2024) for Amharic, or the ongoing Occiglot project (Team, 2024) for European languages. Some models, like Tower (Alves et al., 2024) and ALMA (Xu et al., 2024) are trained with a focus on machine translation. Yet other methods, such as Tanwar et al. (2023); Huang et al. (2023); Zhang et al. (2024), rather focus on prompting strategies to achieve good cross-lingual transfer without necessarily retraining the models.

The Aya Initiative collected instruction data for over 100 languages and varieties, releasing Aya-101 (Üstün et al., 2024), initialised from mT5, and Aya-23 (Aryabumi et al., 2024), initialised from Cohere's Command R model. The latter paper appeals to the 'Curse of Multilinguality' for Aya-101 and other massively multilingual models, and focuses on training Aya-23 with those languages where a larger amount of data is available. By contrast, MaLa-500 (Lin et al., 2024) continues training from Llama-2 and emphasises training on a very broad set of languages.

Zhang et al. (2023) tune their model with instruction data and interactive translation examples in order to improve both translation and cross-lingual instruction following. Li et al. (2023) tune the hidden representations of the first layer using translated data and contrastive learning, which they combine with cross-lingual instruction tuning—this is one of few works on generative models explicitly using a notion of cross-lingual alignment. Li and Murray (2023) use two annotated source languages for cross-lingual generation.

Still, there is plenty of potential for future work in this space. Cross-lingual alignment of representations has not been explored extensively for generative models. Importantly, training schemes will need to enable the model to focus on language-neutral *and* language-specific information at the relevant times. Training for cross-lingual alignment may be combined with sparse fine-tuning approaches such as LoRA (Hu et al., 2021a). Ideally it should also be integrated by those pre-training new multilingual models. The goal is to enable better transfer of information between languages while outputting text in the relevant language.

5.3 Where to Align a Generative Model?

Cross-lingual alignment as we discussed it has been researched primarily in encoder-only models, and some encoder-decoder models. Encoder-only models transform the inputs into a latent space representation which is typically used by a shallow 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. This is clearly the case for the classification and sequence labelling tasks where encoder models shine. The task head has somewhat limited opportunity to transform the encoder representations. Intuitively, strong cross-lingual alignment can be helpful here, including more "radical" methods such as mean-centering language-specific embeddings. We have seen this born out by the proliferation and success of cross-lingual alignment methods in encoders.

In an encoder-decoder framework, there is an explicit delineation with different architectures, where the encoder still outputs a latent space representation, while the decoder predicts the next tokens one-by-one. Cross-lingual alignment techniques from an encoder context can be naturally applied to the encoder of an encoder-decoder model. However, there could be risks—"radical" alignment methods, e.g., mean-centering language-specific embeddings from the encoder, might suppress necessary language-specific information for the decoder, or "break the link" of the encoder to the decoder. Nevertheless, this is certainly a worthwile avenue to explore.

In decoder-only models, there is no such architectural delineation. Thus, there is no one obvious point at which cross-lingual alignment should be greatest. Further, the classic zero-shot cross-lingual transfer paradigm may encounter issues in generative settings, since strongly-aligned representations can lead to generation in the wrong language (Xue et al., 2021; Li and Murray, 2023).

That said, interpretability studies suggest that even in decoder-only models, early layers may perform a kind of encoding role. For instance, Wendler et al. (2024) use an early exiting technique known as *logit lens* (nostalgebraist, 2020) to illustrate at which layers in the model translating logits to tokens starts to produce reasonable outputs. They demonstrate that Llama-2 "works" in English until late in the model, and only then surfaces tokens from the target language. Meanwhile, Chuang et al. (2024) work only in English but detect hallucinated vs. correct outputs using 'differential likelihoods' of early-exited tokens throughout the model. Both papers suggest inflection points in mid-to-late layers, where the correct output language or token starts to become more likely. If we can find a kind of 'implicit encoder' in this way, this could become a target for cross-lingual alignment methods.

5.4 Evaluation of Multilingual Generation

Benchmarking multilingual generation presents unique challenges compared to both multilingual classification and monolingual generation tasks.

Several benchmarks have been proposed in the past (Asai et al., 2023; Ahuja et al., 2023; Gehrmann et al., 2022). However, these benchmarks include a number of reformulated classification tasks in addition to some translation or summarisation tasks. Classification tasks are not the main strength of generative models, and fine-tuned encoder models often do better there (Lin et al., 2022). Translation and summarisation tasks can be evaluated relatively easily using a reference text—which is likely why they were included in the benchmarks.

However, due to the open-ended nature of many generation tasks, even in high-resource languages they are non-trivial to evaluate. Increasingly, researchers use ChatGPT as a judge on open-ended generation (e.g., Liu et al., 2023), but this approach may not be reproducible due to frequent changes of the ChatGPT model, and it is less likely to work well for low-resource languages. Note also that care should be taken to avoid leaking evaluation data to proprietary models (Balloccu et al., 2024).

6 Conclusions

We have surveyed the literature around cross-lingual alignment, providing a taxonomy of methods. We clarified two main views of the concept, noting how simplistic popular measurements are. We summarised insights from the surveyed methods and related analysis papers. 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 which effectively trade-off cross-lingual semantic information with language-specific axes, allowing models to generate fluent and relevant content in many languages.

Limitations

Bilingual vs. Multilingual Alignment

Since we are talking about highly multilingual models, we are implicitly concerned with multilingual cross-lingual alignment. 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 § 2.2), this is typically also done using English as the source 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.

Acknowledgements

This publication was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder; and by the German Research Foundation (DFG; grant FR 2829/4-1). The work at CUNI was supported by Charles University project PRIMUS/23/SCI/023.

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.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

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, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

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.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *CoRR*, abs/2403.06354.

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.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

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.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

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.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2023. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. *CoRR*.

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.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. *CoRR*, abs/2309.03883.

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: Pre-training 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.

Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. 2023. The impact of cross-lingual adjustment of contextual word representations on zero-shot transfer. In *Advances in Information Retrieval*, pages 51–67, Cham. Springer Nature Switzerland.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

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.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3020–3042, Toronto, Canada. Association for Computational Linguistics.

Félix Gaschi, François Plesse, Parisa Rastin, and Yannick Toussaint. 2022. Multilingual transformer encoders: a word-level task-agnostic evaluation. *CoRR*, abs/2207.09076.

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.

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 multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

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.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and Effective Vocabulary Expansion Towards Multilingual Large Language Models.

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.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023. Align after pre-train: Improving multilingual generative models with cross-lingual alignment. *CoRR*, abs/2311.08089.

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.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *CoRR*, abs/2401.13303.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. 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.

nostalgebraist. 2020. interpreting GPT: the logit lens.

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*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, et al. 2024. Gpt-4 technical report.

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, 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.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

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.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Occiglot Team. 2024. [link].

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, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *CoRR*, abs/2307.06018.

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*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. *CoRR*, abs/2402.10588.

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.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

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.

Jiali Zeng, Yufan Jiang, Yongjing Yin, Yi Jing, Fandong Meng, Binghuai Lin, Yunbo Cao, and Jie Zhou. 2023. Soft language clustering for multilingual model pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7021–7035, Toronto, Canada. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models. *CoRR*, abs/2306.10968.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024. PLUG: Leveraging Pivot Language in Cross-Lingual Instruction Tuning. *CoRR*, abs/2311.08711.

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).

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *CoRR*, abs/2402.07827.

A Further Models Explained

We add here brief explanations of additional models which we omitted from the main body. These are methods that did not add as much to the discussion in Section 3, for example because they are quite similar to other methods.

A.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. 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. This 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.

A.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.

A.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. Efimov et al. (2023) do not tune directly with task data, but they do combine a Cao et al. (2020)-style alignment objective with task fine-tuning in a multi-task-learning objective, with mixed success.

A.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 Table 2. This model performs well on PAWS-X but does not stand out on XNLI.

A.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 takes a similarity-based view and 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.

B Evaluation of "Aligned" Models

For reference, we provide a collection of results that the surveyed models achieved on several downstream tasks. 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. A few other tasks are also popular, and we chose a small selection of word- and sentence-level tasks for this overview. Tables 2 and 3 show zero-shot transfer and translate-train results for XNLI, UD-POS (Zeman et al., 2019), PAWS-X (Yang et al., 2019b) and MLQA (Lewis et al., 2020).

Cross-lingual retrieval is also popular, although the specific tasks reported vary. We show results for Tatoeba-36 (Artetxe and Schwenk, 2019) and BUCC2018 (Zweigenbaum et al., 2018), as implemented by Hu et al. (2020), in Table 4.

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.

B.1 What works well?

First, whether the model is newly-trained or modified from a pre-trained model does not appear to determine performance. WordOT, a modified mBERT with an optimal transport objective, yields the best result in its size band. In the next size group, the newly-trained mDeBERTaV3 performs best. This is a model trained 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 element, 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. This method also improves mBERT's performance by a large margin. X-MIXUP 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.

B.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.

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 2. 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 3), 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 4) 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 BUCC, with Tien and Steinert-Threlkeld's (2022) one-pair supervision a close third.

B.3 Limitations

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 the 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.

C 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.

In Table 5, 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	XNLI	UDPOS	PAWS-X	MLQA (F1)
Zero-shot transfer					
mBERT (Hu et al., 2020)	110M	65.4	71.5	81.9	61.4
mBERT + Syntax augm.	~110M	–	–	84.3	60.3
mBERT + EPT/APT	~110M	68.4	–	**86.2**	–
DICT-MLM	~110M	68.6	71.6	84.8	–
mBERT+JointAlign+Norm	~110M	72.3	–	–	–
WordOT	~110M	**75.4**	–	–	–
AMBER	172M	71.6	–	–	–
XLM-R$_{base}$ + EPT/APT	~270M	75.8	–	87.1	–
XLM-ALIGN	~270M	76.2	**76.0**	86.8	68.1
InfoXLM$_{base}$	~270M	76.5	–	–	68.1
ERNIE-M$_{base}$	~270M	77.3	–	–	**68.7**
HiCTL$_{base}$	~270M	77.3	71.4	84.5	65.8
XLM-R+JointAlign+Norm	~270M	77.6	–	–	–
mDeBERTaV3	~276M	**79.8**	–	–	–
XLM-E$_{base}$	279M	76.6	75.6	**88.3**	68.3
mT5$_{small}$	300M	67.5	–	82.4	54.6
XY-LENT$_{base}$	447M	80.5	–	89.7	71.3
XLM-R (Hu et al., 2020)	550M	68.2	73.8	86.4	71.6
HiCTL$_{large}$	~550M	81.0	74.8	87.5	72.8
InfoXLM$_{large}$	~550M	81.4	–	–	73.6
ERNIE-M$_{large}$	~550M	82.0	–	89.5	73.7
XLM-R$_{large}$ + xTune	550M	**82.6**	**78.5**	**89.8**	**74.4**
RemBERT	575M	80.8	76.5	87.5	73.1
mT5$_{base}$	580M	75.4	–	86.4	64.6
VECO$_{out}$	662M	79.9	75.1	88.7	71.7
XLM-V	~750M	76.0	–	–	66.0
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	–	–	73.4
mT5$_{XL}$	3.7B	82.9	–	89.6	73.5
XLM-R$_{XXL}$	10.7B	83.1	–	–	74.8
mT5$_{XXL}$	13B	**85.0**	–	90.0	**76.0**

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. We draw dashed lines between models of markedly different sizes.

Model	Size	XNLI	UDPOS	PAWS-X	MLQA (F1)
Translate-train					
mBERT (Hu et al., 2020)	110M	74.6	–	86.3	65.6
mBERT + X-MIXUP	110M	**78.8**	76.5	**89.7**	**69.0**
InfoXLM$_{base}$	∼270M	80.0	–	–	–
ERNIE-M$_{base}$	∼270M	80.6	–	–	–
mDeBERTaV3	∼276M	**82.2**	–	–	–
mT5$_{small}$	300M	72.0	–	79.9	64.3
XY-LENT$_{base}$	447M	82.9	–	92.4	–
XLM-R$_{large}$ + xTune	∼550M	82.6	78.5	89.8	75.0
FILTER	∼550M	83.6	76.2	91.2	75.8
FILTER + Self-teaching	∼550M	83.9	76.9	91.5	76.2
ERNIE-M$_{large}$	∼550M	84.2	–	91.8	–
HiCTL$_{large}$	∼550M	84.5	76.8	**92.8**	74.4
XLM-R$_{large}$ + X-MIXUP	550M	**85.3**	78.4	91.8	76.5
mT5$_{base}$	580M	79.8	–	89.3	75.3
VECO$_{in}$	662M	84.3	**79.8**	**92.8**	**77.5**
XY-LENT$_{XL}$	2.1B	**87.1**	–	**92.6**	–
XLM-R$_{XL}$	3.5B	85.4	–	–	–
mT5$_{XL}$	3.7B	85.3	–	91.0	75.1
XLM-R$_{XXL}$	10.7B	86.0	–	–	–
mT5$_{XXL}$	13B	**87.1**	–	91.5	76.9

Table 3: Translate-train XNLI, 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	Tatoeba	BUCC
mBERT (Hu et al., 2020)	110M	38.7	56.7
mBERT + LSAR	∼110M	44.6	–
DICT-MLM	∼110M	47.3	–
LaBSe	∼110M	**95.0**	**89.7**
X2S-MA	∼270M	**68.1**	–
LAPCA-LM$_{base}$	∼270M	–	71.3
XLM-E$_{base}$	279M	65.0	–
XLM-R (Hu et al., 2020)	550M	57.3	66.0
HiCTL$_{large}$	∼550M	59.7	68.4
XLM-R + LSAR	∼550M	65.1	–
T&ST (unsup)	∼550M	74.2	82.4
T&ST (one-pair)	∼550M	80.4	89.6
ERNIE-M$_{large}$	∼550M	87.9	–
LAPCA-LM$_{large}$	∼550M	–	83.5
OneAligner	550M	**92.9**	90.5
mSimCSE uns.	∼550M	78.0	87.5
mSimCSE sup.	∼550M	88.3	88.8
mSimCSE NLI	∼550M	91.4	**95.2**
Kvapilíková et al. (2020)	∼570M –		75.8
VECO$_{out}$	662M	75.1	85.0

Table 4: Tatoeba-36 and BUCC performance reported by various papers, ordered by parameter counts.

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
Efimov et al. (2023)	no	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 5: 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.