



PDF Download  
3764112.pdf  
30 December 2025  
Total Citations: 0  
Total Downloads: 641

Latest updates: <https://dl.acm.org/doi/10.1145/3764112>

SURVEY

## Lost in Alignment: A Survey on Cross-Lingual Alignment Methods for Contextualized Representation

FILIPPO PALLUCCHINI, University of Milano-Bicocca, Milan, MI, Italy

LORENZO MALANDRI, University of Milano-Bicocca, Milan, MI, Italy

FABIO MERCORIO, University of Milano-Bicocca, Milan, MI, Italy

MARIO MEZZANZANICA, University of Milano-Bicocca, Milan, MI, Italy

Open Access Support provided by:

University of Milano-Bicocca

Published: 20 November 2025

Online AM: 26 August 2025

Accepted: 11 August 2025

Revised: 23 May 2025

Received: 31 October 2023

[Citation in BibTeX format](#)

# Lost in Alignment: A Survey on Cross-Lingual Alignment Methods for Contextualized Representation

**FILIPPO PALLUCCHINI**, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy

**LORENZO MALANDRI**, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy

**FABIO MERCORIO**, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy

**MARIO MEZZANZANICA**, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy

---

Cross-lingual word representations allow us to analyse word meanings across diverse language settings. It is crucial in aiding cross-lingual knowledge transfer when constructing natural language processing (NLP) models for languages with limited resources. This survey presents a comprehensive classification of cross-lingual contextual embedding models. We assess their data requirements and objective functions, and we introduce a taxonomy for categorising these approaches. Then, we present a comprehensive table containing a set of hierarchical criteria to compare them better, along with information regarding the availability of code and data to enable replication of the research. Furthermore, we delve into the evaluation methodologies employed for cross-lingual embeddings, exploring their practical applications and addressing their current associated challenges.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine translation*;

Additional Key Words and Phrases: Embedding alignment, cross-lingual alignment

## ACM Reference Format:

Filippo Pallucchini, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. 2025. Lost in Alignment: A Survey on Cross-Lingual Alignment Methods for Contextualized Representation. *ACM Comput. Surv.* 58, 5, Article 116 (November 2025), 34 pages. <https://doi.org/10.1145/3764112>

---

## 1 Introduction

Word embeddings are dense, low-dimensional vector representations of words that encode semantic and syntactic information based on their distributional properties in large text corpora. They are typically categorized into two main types: static embeddings (e.g., word2vec [66], GloVe [74]), which assign a single, context-independent vector to each word based on co-occurrence statistics or

---

Authors' Contact Information: Filippo Pallucchini, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy; e-mail: [filippo.pallucchini@unimib.it](mailto:filippo.pallucchini@unimib.it); Lorenzo Malandri, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy; e-mail: [lorenzo.malandri@unimib.it](mailto:lorenzo.malandri@unimib.it); Fabio Mercorio, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy; e-mail: [fabio.mercorio@unimib.it](mailto:fabio.mercorio@unimib.it); Mario Mezzanzanica, Statistics and Quantitative Method, University of Milan-Bicocca, Milano, Italy and CRISP Research Centre, University of Milano-Bicocca, Milano, Italy; e-mail: [mario.mezzanzanica@unimib.it](mailto:mario.mezzanzanica@unimib.it).



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

© 2025 Copyright held by the owner/author(s).

ACM 0360-0300/2025/11-ART116

<https://doi.org/10.1145/3764112>

shallow neural networks; and contextual embeddings (e.g., ELMo [75], BERT [23]), which generate dynamic, context-sensitive vectors using deep neural architectures such as LSTMs or Transformers. While static embeddings capture general word similarity, contextual embeddings can disambiguate word senses depending on their usage in a sentence.

The cross-lingual alignment problem is a specific aspect of the broader challenge of cross-lingual NLP, introduced by Mikolov et al. [65] in 2013. It refers to mapping or aligning embeddings or representations from different languages into a shared or common vector space such that the representations of words or phrases in these languages are directly comparable. The potential to align embedding spaces is relevant in the field of computational language for multiple reasons: (i) it enables us to compare the meanings of words across languages, a critical aspect for numerous NLP tasks, including, but not limited to, **Bilingual Lexicon Induction (BLI)** [102], machine translation [38], mining parallel corpora [53], cross-lingual information retrieval [49]; (ii) it facilitates model transfer across languages, bridging the gap between resource-rich and low-resource languages [82]. Some examples of applications are text classification [51], sentiment analysis [105], and dependency parsing [4]. Finally, (iii) even embeddings trained on the same corpus but in different phases produce misaligned vector spaces, making it impracticable to compare word vectors trained with different hyperparameter settings or at different times [84].

*Motivating Example.* To clarify the matter, we propose an example in the context of **Labour Market Intelligence (LMI)** since it is a domain that could benefit from these techniques and shows increasing use of embedding models, e.g., in detecting new emerging occupations from **Online Job Advertisements (OJAs)** or comparing them [33, 61, 93]. The ability to align embeddings could play a crucial role in advancing **Latent Semantic Shift (LSC)** detection—a growing field in NLP that focuses on automatically identifying changes in word meanings across time, domains, or even languages [62]. For instance, D’Amico et al. [21] employ an innovative approach to align embeddings across countries, enabling a more nuanced comparison of professions and skills. Instead of relying solely on standard taxonomies, this approach analyses job advertisements published directly by companies, capturing real differences in skill requirements across countries.

#### Motivating Example: Aligning multilingual labour market vector space models

Consider two different embeddings generated from two different corpora of OJAs. The first is on Italian OJAs generated in Italian (IT); the second is on UK OJAs generated in English (EN). The vectors generated on occupations and skills contain context-specific semantic and lexical information, i.e. they depend on the country (with its specific labour market features), the language (with its specific lexical features), and the corpus generation year. As such, vectors generated from different embeddings (from different countries or different time periods) are not truly comparable. One possible solution is to align the embeddings using common general words as anchors. For instance, the occupation that is called *Data Scientist* in the EN, might have required skills and mansions that are requested for a *Data Engineer* in IT. Or it could be that recruiters’ use of the term *Digital Manager* is not the same as 5 years ago. How can we effectively compare the two labour markets or job ads from different years using their embeddings? Opting for common general words as anchors might be biased, as they could be used in entirely different contexts.

As illustrated in Figure 1, adapted from Reference [21], it is possible to compute the most similar skills for given occupations using cosine similarity between their embedding vectors. This method leverages both standardised classifications and corpus-specific mappings across countries. In this

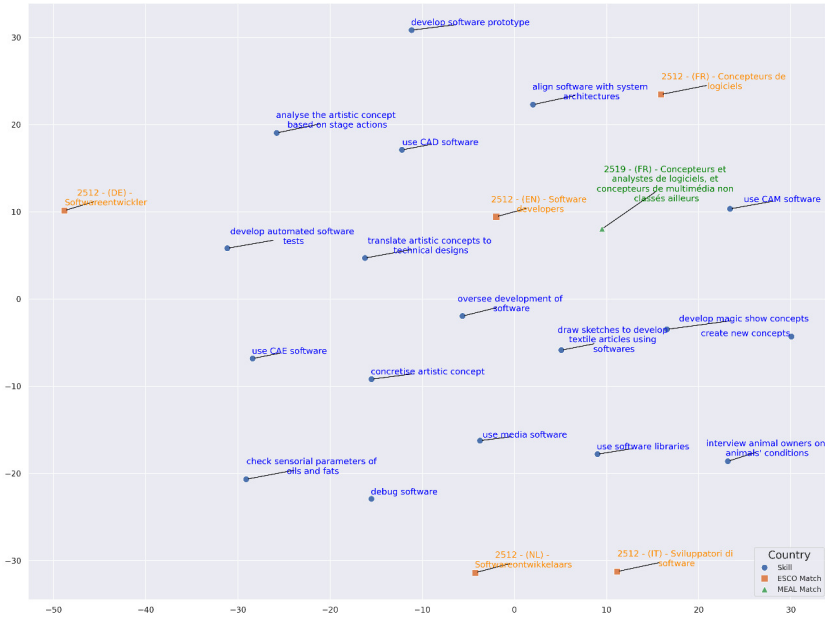


Fig. 1. Example of cross-country occupation mapping based on embedding alignment. Starting from a set of skills frequently required for *Software Developers* in the UK, we identify the most similar occupations and associated skills across countries. Orange squares represent occupations, blue dots represent skills, and green triangles highlight the closest matching occupation in France—*Concepteurs et analystes de logiciels*. This example, adapted from [21], demonstrates how embedding alignment enables comparison of models trained on heterogeneous corpora from different national contexts.

specific case, the analysis starts from a set of skills frequently associated with the occupation *Software Developer* in the UK. Through embedding alignment, we map these to corresponding occupations and skills in other countries. Interestingly, in the French embedding space, the occupation that emerges as most semantically similar is *Concepteurs et analystes de logiciels* rather than the direct translation, *Développeur de logiciels*. This discrepancy reflects how different labour markets associate distinct skill sets with otherwise similar job titles—emphasizing the added value of alignment techniques when comparing embeddings trained on heterogeneous datasets. The visualization thus enables a more nuanced cross-national comparison of occupational semantics, going beyond standard taxonomies such as ESCO.<sup>1</sup>

Instead of aligning already trained embedding, it is also possible to directly train a multilingual model; numerous researchers proposed cross-lingual word representations [51, 65] that create a shared embedding space for words across two (Bilingual Embeddings, BE) or more languages (Multilingual Embeddings, ME). However, multilingual static models often contain inaccuracies, primarily due to the lack of context consideration. Despite this, there is some work applied to static embedding that is important to mention because it lays the foundation for some methods developed later, like [6, 53]. Most of these approaches for static embedding alignment primarily focus on constructing a “seed lexicon”, which is essentially a collection of words that share the same meaning in both corpora. These words, commonly referred to as anchors, serve as reference points for learning a transformation that maps embeddings from one space to another by minimising

<sup>1</sup><https://esco.ec.europa.eu/en/about-esco/what-esco>

their distances. Context-free embeddings, like word2vec [66], generate a fixed representation for each word regardless of its context, capturing only general meanings. In contrast, contextualised embeddings, introduced with models like ELMo [76] and later advanced with BERT, adapt word representations based on surrounding words, allowing for nuanced interpretations of each word's meaning within specific contexts. With the advent of contextual embedding, the drawbacks of multilingual models decrease significantly because each language carries a different context that this model has the ability to embed. Indeed, the contextual nature of the embeddings means that they capture the meaning of individual words and the meaning of the surrounding context, making them particularly useful for natural language processing tasks [85]. So, researchers have started proposing contextualised multilingual models. One such instance is Multilingual BERT, referred to as mBERT and introduced by Devlin et al. in 2019 [23]. It is a single language model pre-trained on the combination of monolingual Wikipedia corpora sourced from 104 different languages. Multilingual BERT simplifies zero-shot cross-lingual model transfer; Pires et al. [77] fine-tuned the model on task-specific supervised data from one language and evaluated it in another, showcasing how the model generalises across languages effectively. Their experiments demonstrated mBERT's remarkable cross-lingual generalisation capabilities, even across languages with no lexical overlap. This suggests that mBERT can effectively capture multilingual representations. In contrast to monolingual BERT, which lacks zero-shot transfer capabilities, multilingual BERT distinguishes itself in two main ways: (1) during pre-training, specifically in the masked word prediction task, each batch comprises sentences from all languages, and (2) it adopts a unified vocabulary generated by applying WordPiece to the combined monolingual corpora [23]. But can multilingual contextual embeddings effectively address the cross-lingual problem? These methods can only learn implicitly the correspondence between words and structures across languages in a purely unsupervised manner. Indeed, Blevins et al. [9] could even quantify how *multilingual* English pre-trained models are. While those multilingual models may demonstrate significant empirical effectiveness [10], the weaknesses have emerged constantly in the literature. Indeed, those models work best for typologically similar languages, indicating that they can map learned structures to new vocabularies but do not systematically transform those structures to accommodate languages with different word orders. This latter insight is reinforced by the work of Wu et al. [97] who show that language representations are not correctly aligned in mBERT, but can be linearly re-mapped; in particular, they built a contrastive alignment objective that can better utilise bitext signal. Furthermore, these models may encounter challenges related to polysemy and homonymy. This entails the difficulty of managing words with multiple meanings (polysemy) or words that share the same spelling but possess different meanings (homonymy) in the context of multilingual applications. Some approaches directly utilise multilingual parallel corpora ([18, 27, 44, 63, 86]), which implicitly provide some level of supervision for aligning words in the two languages. However, the pressure on the model to learn clear correspondences between the contextualised representations in the two languages is still implicit and somewhat weak [42]. This is due to the absence of explicit alignment supervision, requiring the model to deduce the connections between representations from the available training data. Consequently, several subsequent studies (e.g., References [12, 85, 95]) have introduced methods that utilise word alignments from parallel corpora as supervision signals to align multilingual contextualised representations in a post-hoc manner or directly during the training ([15, 42, 96]). In most cases, these alignment procedures enhance multilingual language models and address many of their systematic deficiencies, particularly in the **Natural Language Inference (NLI)** task [26]. However, it is well established that language dependency persists, making tasks like **Question Answering (QA)** and **Named Entity Recognition (NER)** still challenging.

## 1.1 Contribution

In this work, we report all the most important cross-lingual alignment methods for contextualised representations. We can summarise the contributions of the article in four points:

- (1) We provide a comprehensive classification of alignment models. As far as we know, no other works in the literature provide a taxonomy.
- (2) We present a categorisation of research works in cross-lingual alignment based on 6 features to gain a deeper understanding of the authors' rationale and to enhance their comprehensibility
- (3) Starting from the analysed models, we identify the primary challenges related to the problem of cross-lingual alignment, and we discuss them.
- (4) We create a repository containing codes and data for reproducing the methods, if provided, available to the whole community on Git.<sup>2</sup>

## 2 Cross-Lingual Alignment Taxonomy

Cross-lingual alignment of contextual embeddings refers to the process of mapping embeddings from different languages into a shared space or direct cross-lingual adjustment of the multilingual model, such that semantically similar words are located close to each other in the shared space. This allows for transfer learning across languages, where models trained on one language can be applied to another language without additional training data. One of the main goals of cross-lingual alignment is to take an existing model that has been trained on a resource-rich language and align the contextual embeddings from a less-resourced language into the vector space of the resource-rich language. This alignment allows input in the less-resourced language to be mapped into the resource-rich language, making it possible to classify it using existing models in that language. This is achievable because words with equivalent meanings in both languages exhibit highly similar vectors after undergoing the cross-lingual alignment process [91]. We build a taxonomy that classifies the most important methods into 6 macro-clusters as a result of a study of the State-Of-The-Art, as reported in Figure 2

Here we report the description of each building block present in the Figure 2:

**N Monolingual Corpora** It refers to the N monolingual corpora used as input for the embedding model.

**Train N Embeddings** It refers to training N contextualised language models that need to be aligned. NB: numerous authors do not train models but use already existing pre-trained models.

**Anchors Selection** It describes the procedure for selecting anchors, which are reference points to guide the mapping of one embedding into the other or directly create a common space for different corpora. This selection process may involve pairs of words or pairs of sentences. It can also involve a standalone model for pair extraction (e.g., a word alignment method) or an off-the-shelf list of pairs.

**Cross-Lingual Alignment** It refers to a linear or non-linear transformation that maps source embeddings to the target language or to a shared space.

**Train Multilingual Embeddings** It is related to the process of training a contextualised multilingual language model.

**Adjust Multilingual Embeddings** It refers to fine-tuning methods used to adjust an existing multilingual model, creating a more effective cross-lingual language model. The main difference with respect to the cross-lingual alignment box is that in this case, languages are already in the same embedding space, but there is a further modification of it to align them.

---

<sup>2</sup><https://gitlab.com/crisp1/lost-in-alignment>

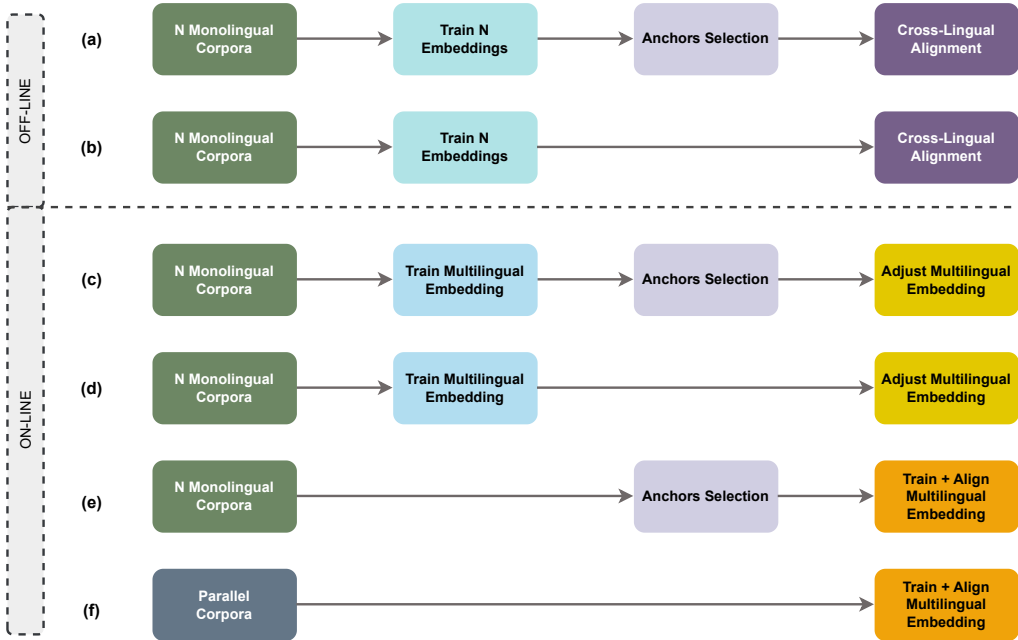


Fig. 2. A taxonomy of the cross-lingual alignment methods.

**Parallel Corpora** It refers to both the monolingual corpus and its respective corpus translated into another language. These are used as input for the multilingual embedding model.

**Train+Align Multilingual Embeddings** It is related to the process of training a multilingual language model that utilises objective functions during the training phase to represent cross-linguality.

The combination of those building blocks defines the following strategies:

(a) - **Offline Independent Embeddings w Alignment via anchor selection**: This category includes methods that attempt to align  $N$  different monolingual language models with the assistance of anchor sets extracted from monolingual or parallel corpora.

(b) - **Independent Embeddings with Alignment w/o direct anchor selection**: Research in this category bears a resemblance to that in class (a), but they do not rely on anchors for alignment.

(c) - **Online Multilingual Embedding Adjustment w Anchors**: In this category, we find works that adjust an existing multilingual language model by modifying the embedding remapping process, where the alignment occurs directly during the training phase through signals derived from bilingual dictionaries or parallel corpora.

(d) - **Direct Multilingual Embedding Adjustment**: Similar to the (b) case concerning (a), this category encompasses models that employ methods from the (c) category without the need for explicit anchor supervision.

(e) - **Online Anchored Multilingual Embedding Training and Alignment**: Papers in this category perform pre-training for cross-lingual language models from the ground up, selecting anchors to facilitate alignment during the training process.

(f) - **Direct Parallel Training and Alignment**: These methods initiate pre-training directly from parallel corpora without the prerequisite of pre-existing anchor sets as signals.

Methods (a), (b), (c), and (d) optionally utilise parallel corpora solely for alignment purposes, rather than for embedding training. They may employ parallel source data either for anchor selection or as input for fine-tuning multilingual embeddings. Methods like (d) and (f) omit the use of an anchor selection module, as they create a word-level alignment objective that encourages the model to independently identify word alignment patterns from the parallel corpus during an end-to-end training process. This approach helps mitigate the risk of accumulating potential errors at distinct stages of the pipeline. Some methods in category (f) also incorporate monolingual corpora as input for training the multilingual embedding model.

### 3 Main Challenges Related to Cross-Lingual Language Models

There are some key challenges embedded in cross-lingual language models that should be noted to better understand if and how the state-of-the-art methods address them. Mikolov et al. in 2013 initially observed that word vectors pre-trained on monolingual data exhibit comparable topological structures across various languages [65]. This similarity enables the alignment of embedding spaces through a straightforward linear mapping [65]. Nevertheless, this assumption also poses a substantial limitation, as the diverse structural characteristics, such as morphology and syntax, present challenges for embeddings in adhering to this hypothesis. In the next subsections, we will address the challenges that have arisen over time when confronting the cross-alignment dilemma.

#### 3.1 Isomorphism, Isometry, and Isotropy

Isomorphism, isometry, and isotropy are all related concepts in the context of cross-lingual embedding space mapping, but they refer to different aspects of the embedding spaces.

**Isomorphism** refers to the degree to which two embedding spaces have similar topological structures. In other words, it measures how well the two spaces preserve the relationships between words and concepts. Although contextual embeddings are designed to offer distinct representations of the same word in various contexts, Schuster et al. [85] discovered that the contextual embeddings of different senses of a single word exhibit much greater similarity compared with embeddings of different words. This phenomenon contributes to the anisomorphic distribution of embeddings in different languages and poses challenges for cross-lingual alignment. For instance, aligning the English word “bank” with its Italian translations, “banca” and “sponda”, corresponding to its two different senses (“banca” as “financial institution” and “sponda” as “land at river’s edge”), becomes difficult due to the contextual embeddings of the different senses of “bank” being close to each other, while those of “banca” and “sponda” are distant from each other [69]. We report a figure from Reference [87], Figure 3, where they affirm that considering the top  $k$  most frequent English nouns and their translations, the graphs are not isomorphic; see Figure 3(c)- 3(d). Even if we consider the top  $k$  most frequent English words and their translations into German, the nearest neighbour graphs are not isomorphic. Figure 3(a)- 3(b) shows the nearest neighbour graphs of the top 10 most frequent English words on Wikipedia, and their German translations.

**Isometry**, on the other hand, refers to the degree to which the distances between points in the two spaces are preserved. It measures how well the two spaces preserve the relative distances between words and concepts. Two spaces are considered isometric when the relative Euclidean distances among vectors remain the same between these spaces. An orthogonal mapping is a transformation that preserves distances and guarantees isometric isomorphism, ensuring that the Euclidean distance between two vectors remains unchanged after the mapping process. Consequently, aligning two semantic language spaces becomes more straightforward when the relative distances among their vectors exhibit similarity [100]. To elucidate the concept, we present a figure (Figure 4) sourced from the work of Conneau et al. [19] that illustrates the process of aligning two embedding spaces while

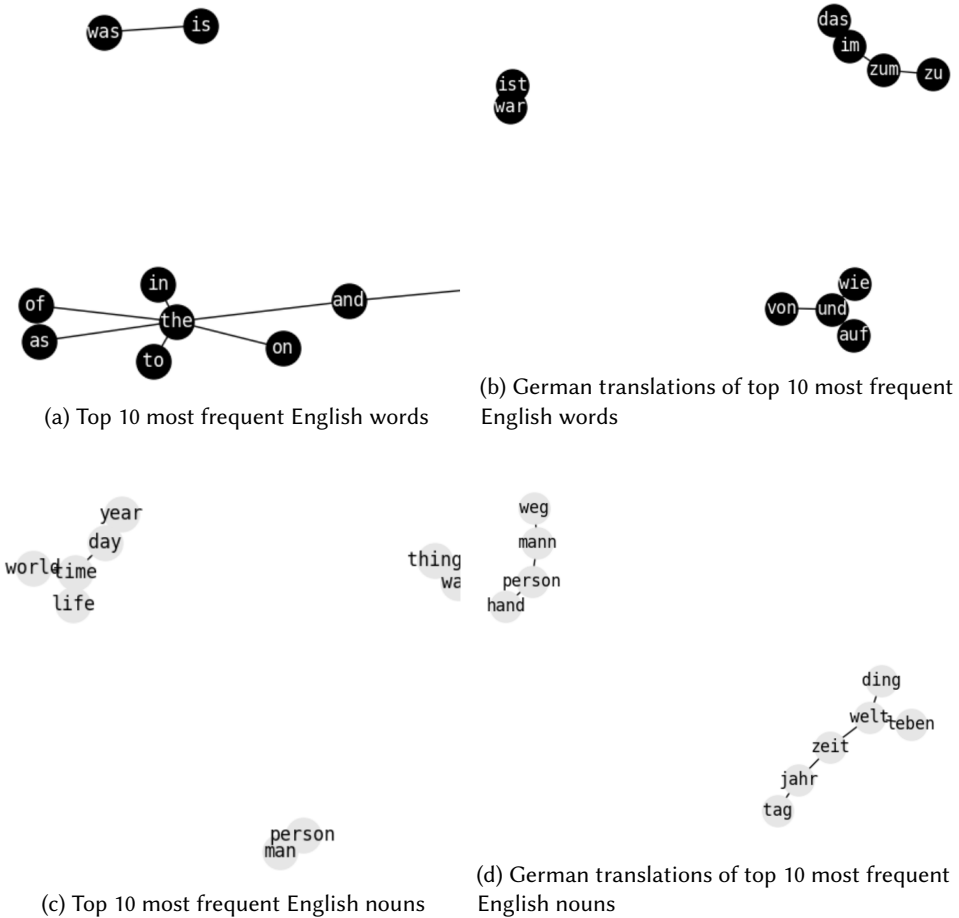


Fig. 3. Nearest neighbour graphs from [87]: Authors select the top 10 most frequent words in English, and the respective translation in German, and build nearest neighbour graphs for English and German using the monolingual embeddings used in Conneau et al. [19], the graphs are of course very different. (a), (b) represent the graph of the 10 most frequent words in English, and the respective German translations; (c), (d) represent the graph of the 10 most frequent nouns in English, and the respective German translations of.

retaining the original distances between terms. The authors employed the Procrustes measure to achieve this alignment, which involves calculating an orthogonal matrix. This matrix preserves the dot product of vectors and their distances and functions as an isometry in the Euclidean space, akin to a rotation. More specifically, the authors align English  $X$  and Italian word embeddings  $Y$  through an adversarial learning process and subsequent refinement. First, a rotation matrix  $W$  is learned to roughly align the distributions of each language's embeddings. Then, Procrustes refinement further optimises  $W$  using frequent words as anchors. Finally,  $W$  is applied to all words in the dictionary, with a distance metric expanding high-density areas to improve the separation of common words.

**Isotropy** refers to how symmetrically vectors are distributed across an embedding space. High isotropy indicates that vectors are evenly spread in all directions, while low isotropy suggests clustering in certain directions. An embedding space is considered isotropic when vector directions are uniformly distributed. Unfortunately, contextual word representations often exhibit anisotropy [78]:

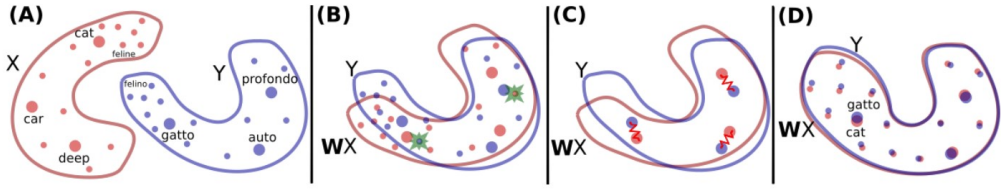


Fig. 4. Figure from [19] where: (A) Within this context, there exist two sets of word embeddings: English words represented in red and denoted as  $X$ , and Italian words represented in blue and denoted as  $Y$ . The primary objective is to align or translate these embeddings. Each data point (depicted as a dot) in the space corresponds to a word, with the size of each dot proportional to the word's frequency in the training corpus of its respective language. (B) An adversarial learning process acquires a rotation matrix  $W$ , aiming to align the two distributions roughly. The green stars mark randomly selected words used for discrimination, determining whether the embeddings from the two sets originate from the same distribution. (C) The mapping  $W$  undergoes further refinement through the Procrustes method. This involves the utilisation of frequently occurring words, already aligned in the previous step, as anchor points. The optimisation process minimises an energy function akin to a spring system between these anchor points. The refined mapping is then employed to map all words in the dictionary. (D) Ultimately, translation is achieved by employing the mapping  $W$  and a distance metric that expands the space, particularly in regions with high point density (such as around the word “cat”). This expansion serves to separate “hubs” (e.g., the word “cat”) from other word vectors to a greater extent compared to the original state (as seen in panel (A)).

normalised embeddings tend to cluster within a narrow, cone-like region on a hypersphere rather than spreading evenly. Different languages commonly display varying anisotropy levels [100]. Ethayarajh et al. [28] found that contextualised embeddings from models like BERT, ELMo, and GPT-2 exhibit this property, with vectors concentrating in specific regions due to their sensitivity to contextual nuances. This clustering effect seems intrinsic to contextualization itself. Recent work by Godey et al. [34] shows that anisotropy emerges independently of token frequencies or vocabulary size, challenging prior assumptions that only these factors drive it. Geometrically, anisotropy manifests as vectors confined within a cone, with the degree of anisotropy corresponding to the cone's narrowness, as Mimno et al. [67] noted. This effect is observed across most layers in models like BERT, ELMo, and GPT-2, where word representations form similar conical clusters, as visually depicted by Ethayarajh [28] and shown in Figure 5. The Figure 5 shows that in GPT-2, the average cosine similarity between randomly sampled words stays around 0.6 in layers 2 to 8 but then increases sharply in layers 8 through 12, with near-perfect similarity in the final layer. A similar trend occurs in BERT and ELMo, although BERT's second-to-last layer has higher anisotropy than its final layer.

Those concepts create issues in the mapping process, in particular due to the inconsistent relative distances between embeddings in the source and target embedding spaces. As a response, the paper proposed by Gao et al. [31] introduced regularisation techniques to improve isotropy by expanding the “aperture” of the embedding space, thus mitigating this degeneration. Zhao et al. [104] sought to reduce (or increase) the degree of anisotropy (isometry) to alleviate its adverse effects. However, it is impractical to match the anisotropic characteristics of spaces precisely. Instead, they introduced an **iterative normalisation (IN)** preprocessing technique. This method is applied to transform anisotropic contextual embedding spaces, making them approximately isotropic by redistributing vectors evenly across the surface of a unit hypersphere. This significantly enhances the degree of isometry, making relative embedding distances across language spaces more similar. The

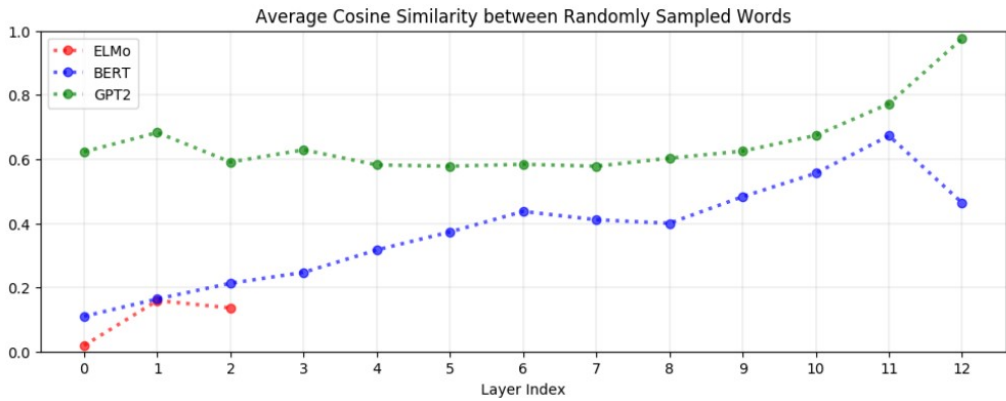


Fig. 5. Nearest neighbour graphs from [28]: In the majority of layers within BERT, ELMo, and GPT-2, the word representations exhibit anisotropy, meaning they are not directionally uniform. Specifically, when you calculate the average cosine similarity between words that are randomly selected uniformly, you find that this similarity is not zero. One notable exception is ELMo’s input layer. This exception is expected since it generates character-level embeddings independently of context. Typically, as you move to higher layers, the word representations become even more anisotropic than those in lower layers.

iterative preprocessing method enforces vectors to become zero-mean and normalised as part of this transformation [100], showing the positive effect of isotropic space on the degree of isomorphism, which in turn results in improved performance in cross-lingual alignment algorithms [79]. Zhao et al. [104] continued this line of research and introduced enhancements to the model alignment approach, including (1) the application of z-normalisation to vectors, and (2) text normalisation to align the input more closely with the structural characteristics of English training data.

### 3.2 Language Neutrality

Language-neutrality is about to what extent similar phenomena are represented similarly across languages [58]. It refers to the ability of a language model to represent language in a way independent of any specific language. This means that the model can capture linguistic features and meanings common across different languages, allowing it to be used for multilingual tasks without requiring language-specific training data. Based on the findings of Pires et al. [77], it can be postulated that a sentence’s representation in multilingual language models, without a specific alignment objective, consists of two components: one that is specific to the language of the sentence, responsible for identifying the language, and another that is language-neutral, capturing the sentence’s meaning in a way that transcends language boundaries. Libovicky et al. [58] demonstrated that the representations in mBERT can be decomposed into a language-specific element and a language-agnostic element. They found that the language-neutral component is quite versatile in modelling semantic information, enabling accurate word alignment and sentence retrieval. However, it falls short regarding the more challenging task of assessing machine translation quality. The fact that mBERT has shown limited success, particularly for certain language pairs, in zero-shot scenarios and necessitates explicit bilingual projection for optimal performance, as observed by Pires et al. [77], Wu and Dredze [98], and Ronnqvist et al. [81], also underscores its limited language neutrality. By conducting a series of semantically oriented tasks that demand explicit cross-lingual semantic representations, Libovicky et al. [58] demonstrated that mBERT’s contextual embeddings do not consistently represent similar semantic phenomena in a way that can be directly applied to zero-shot cross-lingual tasks.

## 4 Alignment Methods for Contextual Embeddings

Cross-lingual approaches can be categorised into several groups.

- (1) The first group of methods employs monolingual embeddings, supplemented by bilingual dictionaries or parallel corpora, to align the embeddings in a linear or non-linear fashion.
- (2) The second group of approaches optionally uses bilingually aligned (comparable or even parallel) corpora or dictionaries to fine-tune a pre-trained multilingual model. Here, alignment is performed as a separate step after the initial pre-training phase.
- (3) The third group focuses on large pre-trained multilingual masked language models, such as mBERT [23], which are modified to incorporate cross-lingual alignment during training. These models are simultaneously trained on multiple languages with a specific objective function for alignment, allowing alignment to occur as part of the training process itself. Similar to the previous groups, bilingual corpora or dictionaries may be used for additional supervision.

### 4.1 Off-Line Linear/Non-Linear Alignment

These methods belong to the categories (a) and (b) in the taxonomy of Figure 2. linear: Mikolov et al. [65] were the first that proposed to learn a linear transformation to project an embedding in the source language to its translation (since they used anchors) in the target language. Now we are going to present one of the most commonly used methods. Let's consider two embeddings as  $e_s$  for the source ( $s$ ) and  $e_t$  for the target ( $t$ ). Given a dictionary containing pairs of source and target elements represented as  $x_s, y_t$  and matrix representations  $X_s$  and  $Y_t$  where the columns represent vector representations of the corresponding dictionary entries, the objective is to discover an orthogonal transformation matrix denoted as  $W$ . This matrix aims to minimise the discrepancies between the transformed vectors in  $WX_s$  and  $Y_t$ . In a formal sense, this can be expressed as

$$\arg \min_{\hat{W}} \|\hat{W}X_s - Y_t\| \quad s.t. \quad \hat{W}^T \hat{W} = I. \quad (1)$$

Here, the notation  $\|\cdot\|$  denotes the Frobenius norm. The orthogonality constraint ensures that the relative distances between pairs of vectors in the original source vector space remain unchanged following the transformation. These methods imply an off-line alignment. Indeed, for computing the matrix, it is necessary just to have the embeddings of the source and target language without further training the model. Rotating embeddings relies on the controversial assumption, explained above, that the embedding spaces exhibit approximate isometry [87]. This assumption may not be valid for contextual pre-trained models since they encompass more than just a word's type, including context and syntax, which are less likely to exhibit isomorphism across languages.

Some authors try to align two contextual embedding models in a non-linear way. A common practice in performing this task is to use **Generative Adversarial Networks (GANs)** [35]. GANs consist of two interconnected neural models: a generator and a discriminator. These two models are concurrently trained through an adversarial process. The discriminator's role is to distinguish whether the input data it receives is real or artificially generated (i.e., fake). Simultaneously, the generator aims to create synthetic data that can deceive the discriminator. GANs operate as a zero-sum game, where the discriminator's success implies the generator's failure and vice versa. Through simultaneous training, both networks enhance their performance. GANs find extensive applications in image generation, where this described process can yield impressive newly generated images. In their 2022 work, Ulčar et al. [91] introduce an innovative supervised nonlinear mapping approach that utilizes bidirectional GANs. In particular, they built a system where the generator module contains two generators that map vectors from one vector space to the other; one generator will map from  $s$  to  $t$ , and the second will map from  $t$  to  $s$ . The two generators are completely independent

of one another, and they do not share data during training. The discriminator module contains two discriminators. The first discriminator tries to predict whether a given pair of vectors  $\langle x_s, y_t \rangle$  represent the same token, and the second vector represents the translation of the word  $x$  in  $t$  (i.e.,  $y_t$ ). The second discriminator attempts to learn the difference between the direction of mapping. For a given pair of vectors, it predicts whether they are a vector from  $s$  and its mapping to  $t$  or a vector from  $t$  and its mapping to  $s$ . GANs could also be used to learn linear mapping.

## 4.2 Off-Line Fine-Tuning

These methods belong to categories (c) and (d) in the taxonomy shown in Figure 2. The key distinction concerning the previous category is related to the fact that alignment is carried out on an already pre-trained multilingual embedding model. In this case, there are no longer two distinct embedding models, but rather a single model with a multilingual representation. Some authors have suggested fine-tuning the entire encoder as an alternative to utilising source and target embeddings as static features. The concept involves directly modifying the multilingual model to bring the representations of semantically related words in various languages closer to each other, thereby encapsulating alignment in the loss function. This idea stemmed from the observation that embedding spaces for different languages may not always exhibit isometric properties [87], and as a result, they may not always be easily aligned through rotation. Considering the notation in Section 4.1, we can represent the multilingual embedding of the word  $m$  as  $x_m$ , where  $x_{s_m}$  is associated with source language terms and  $x_{t_m}$  with target ones. Fine-tuning aims to minimise the distance between the contextual representations of matched word or sentence pairs in parallel corpora, as measured by the Frobenius norm, at each layer of the model:

$$L_{align}^i = \min \sum_m \|x_{s_m}^i - x_{t_m}^i\|, \quad (2)$$

where  $i$  is the ME layer. Nevertheless, fine-tuning solely based on the objective above would result in the loss of semantic information that ME acquired during pre-training. This is because a straightforward solution to Equation (2) would be to make all embeddings identical. To tackle this issue, researchers introduced a regularisation loss; for instance, Cao et al. [12] proposed one that restricts the source language embeddings from straying too far from their original locations within the pre-trained mBERT model. This regularisation loss functions in the following manner:

$$L_{regularise}^i = \min \sum_m \|x_{s_m}^i - c_{s_m}^i\|, \quad (3)$$

where  $c_{s_m}^i$  is a copy of the original pre-trained  $x_{s_m}^i$ , with its parameters remaining frozen. Both the alignment and regularisation losses are integrated and jointly optimised to align the two language subspaces while preserving the informativeness of the embeddings:

$$L_{finetune} = \min \sum_{i=n_s}^{n_e} L_{align}^i + L_{regularise}^i \quad (4)$$

Here  $n_s$  to  $n_e$  is the range of ME layers aligned. These objectives could be used for word alignment and sentence alignment, as demonstrated in the case of Pan et al. [73]. It is important to mention the framework built by Wang et al. [95] that applies cross-lingual alignment methods mentioned in the previous Section 4.2 to a multilingual embedding model.

## 4.3 On-Line Alignment

These methods belong to the categories (e) and (f) in the taxonomy of Figure 2. Online methods combine monolingual and cross-lingual objectives to acquire cross-lingual embeddings concurrently.

These approaches require more data and are computationally less efficient alternatives to the fine-tuning approach, as they involve pretraining a large multilingual dictionary. Each word in the dictionary must have ducting alignment as an afterthought. This line of research suggests contextual pre-training procedures that exhibit greater cross-lingual awareness. Joint training methods in general have the following objective ( $L_j$ , where  $j$  stand for joint):

$$L_j = L_1 + L_2 + R(L_1, L_2), \quad (5)$$

where  $L_1$  and  $L_2$  are monolingual objectives and  $R(L_1, L_2)$  is a cross-lingual regularization term. An example is the method developed by Chi et al. [15], in which they introduce a new pre-training task called **denoising word alignment (DWA)**. This task involves training a model to predict the correct alignment of noisy word pairs across parallel sentences in different languages. The DWA task serves as an explicit alignment objective during the pre-training of the cross-lingual language model, representing the  $R(L_1, L_2)$  component mentioned above. The goal of DWA is to predict the word alignments from the perturbed version of the input translation pair (constructed by randomly replacing the tokens with masks); so, more specifically, the training objective is to minimize the cross-entropy between the alignment probabilities from the perturbed version and the self-labelled word alignments (semantic similarity of original tokens). Additionally, the authors employ a **masked language modelling (MLM)** objective, which is a standard pre-training task for language models and corresponds to the  $L_1$  and  $L_2$  components.

## 5 Sources of Cross-Lingual Supervision

In this section, we will explain how previous studies use two distinct cross-lingual indicators, namely bilingual dictionaries and parallel corpora, to oversee the alignment process. Furthermore, we will assess the pros and cons associated with each option.

### 5.1 Bilingual Dictionary

A bilingual dictionary provides translations between two languages. In this survey's research, a bilingual dictionary is used as a form of cross-lingual supervision to align the embeddings of words in different languages. These dictionaries can be made using a translation tool, like the widespread Google Translate, or generated from parallel corpora. To effectively employ a bilingual dictionary for supervising the alignment of embeddings, each word in the dictionary must have a single representation, avoiding possible degradation in the mapping phase. However, the same word can have multiple representations depending on its context in the vector space of contextualised language models. According to Schuster et al. [85], the contextual embeddings of the same word form a cohesive cluster or "word cloud," and the centroid of this word cloud is distinct and separable for individual words, so it can be used as a candidate anchor. Nevertheless, there are limitations to this approach. When averaging over multiple contextual embeddings, some contextual information is inevitably lost for both the source and target language words. Some works like the one proposed by Zhang et al. [103], discovered that multi-sense words, such as "bank," which can refer to either a financial organization or the edge of a river based on the context, also have clearly distinguishable clusters within their respective word clouds for each word sense. Bilingual dictionaries suffer from several weaknesses. For instance, Schuster et al. [85] conducted an experiment where they computed the average of ELMo embeddings for each word selected as anchor. Their research unveiled that the average cosine distance between contextual embeddings of polysemous words and their associated anchors was notably smaller than the average distance between those anchors. This implies that the embeddings representing different senses of a single word are relatively closer to each other compared with embeddings representing different words, as already mentioned in Section 3.1. Liu et al. [60] observed a similar pattern with BERT embeddings as well. This discovery suggests that

the sense clusters of a multi-sense word's occurrences are not well separated in the embedding space.

## 5.2 Parallel Corpora

A parallel corpus is a corpus that consists of texts in multiple languages, wherein the texts are translations of each other. In the field of research of this survey, parallel corpora are employed as a form of cross-lingual signal to align contextual embeddings or to extract a bilingual dictionary. This approach allows us to leverage not just individual words in the source and target languages, but also their contextual information. The cross-lingual alignment process can be guided effectively by utilizing the contextual information in the parallel corpus [2, 96]. Some authors apply word-alignment techniques to obtain the silver-aligned token pairs. As summarised by Xu et al. [100], using parallel corpora instead of a dictionary offers three advantages for mapping purposes:

- (1) Parallel corpora offer a broader and more comprehensive range of translation pairs compared with a dictionary;
- (2) Embeddings of translation token pairs preserve the same contextual information;
- (3) Tokens in each parallel sentence are already aligned, and their embeddings are aligned as well. Thus, mappings can be created by aligning the embeddings directly, eliminating the need for word alignment using a dictionary.

It would seem that parallel corpora are much better than bilingual dictionaries, but they entail some challenges: First, these kinds of sources are not easy to obtain for specific domains or languages. Second, word-alignment annotations (often needed in addition to aligned pairs of sentences) are not commonly available in parallel corpora, but they can be automatically generated using off-the-shelf tools, which we will discuss in the upcoming subsection.

*Word alignment methods.* Word alignment is a critical issue in statistical machine translation. While pursuing advanced models that offer more nuanced interpretations of parallel corpora is a prominent research endeavour, it is equally important to have simple and efficient models that can scale effectively. Such models are crucial in various scenarios, including parallel data mining and rapid large-scale experimentation. They also serve as subcomponents in other models or training and inference algorithms [25]. The IBM models [11] are statistical models used for representing the translation process and extracting word alignments between pairs of sentences. Numerous word alignment models have been developed based on the IBM models, including those proposed by Och and Ney [70], Mermer and Sarac¸olar [64], Dyer et al. [25], and stling and Tiedemann [72]. Recent studies have revealed that word alignments can also be extracted from neural machine translation models ([32, 52, 57]) or from pre-trained cross-lingual LM ([68, 83]). The predominant method utilised in the research reported in this article is `fast_align` [25], which is based on a modified version of the lexical translation models initially proposed by Brown et al. [11]. In particular, `fast_align` adopts a log-linear reparameterization of IBM Model 2 [70]. The lexical translation process operates as follows: when presented with a source sentence, it generates an alignment indicating the correspondence between each target word and its respective source word (or null token) as a translation. Naturally, utilising such models to detect anchor pairs for alignment introduces some errors, which must be considered in addition to the errors inherent in the cross-lingual alignment method. Importantly, an off-the-shelf tool's word alignment error rate decreases as the number of parallel sentences increases. Consequently, parallel corpus supervision is particularly advantageous for language pairs with a lot of parallel data available.

## 6 Classification of the Models

In this section, we report the Table 1 with all the research analysed in the survey and the hierarchical criteria used to compare them. Now we describe the set of hierarchical criteria reported in the Table 1:

- (1) **Reproducibility:** Reproducibility is related to the code availability and dataset necessary to replicate the experiments.
- (2) **Alignment Approach:** We categorize methods into two sub-classes similar to the ones introduced by Wang et al. [95]: (1) off-line alignment, where distinct independently trained monolingual representations are mapped into a shared space, and (2) on-line alignment, which involves the simultaneous learning of unified multilingual representations through both monolingual and cross-lingual objectives.
- (3) **Training signal:** It refers to the type of signal used for performing the cross-lingual alignment.
- (4) **Word-alignment method:** It is specified whether the author utilises an off-the-shelf method or performs the word alignment task during the cross-lingual alignment procedure.
- (5) **Level of alignment:** It specifies whether the alignment is conducted at the word level, with the objective being to align specific identical words considering the context or not, or at the sentence level, with the objective being to align entire pairs of sentences directly.
- (6) **Taxonomy:** Specifies to which categories, introduced in Section 2, the model belongs.

We conducted a comprehensive literature review to identify the most commonly used cross-lingual alignment methods, and we framed the works in the categories just introduced, as reported in Table 1. As advised by Reference [50], we comprehensively searched electronic databases<sup>3</sup>. We reviewed 39 papers on cross-lingual alignment of contextual embeddings. Papers were included if they met the following quality criteria, or if they were recognised as particularly impactful within the research community:

- (1) for journal papers, to be either *Q1* or *Q2* of SCImago journal ranking in any computer science-related topics in the of publication.
- (2) for conference papers, to be classified as *A/B* for all those rankings: (i) CORE Conference Rating, (ii) LiveSHINE, and (iii) Microsoft Academic.

### 6.1 Category (a) - Offline Independent Embeddings w Alignment via Anchor Selection

Approaches classified under category (a) follow the process depicted in Figure 6. Aldarmaki et al. [2] employ parallel text to align independently trained contextual embeddings across languages, employing both word-level and sentence-level mapping techniques. Similarly, Wang et al. [94] generate cross-lingual contextualised embeddings using publicly available pre-trained BERT models, utilizing word pairs from parallel corpora for word-level alignment. Schuster et al. [85] propose

<sup>3</sup>The databases utilised for this search were as follows:

ACL (<https://aclanthology.org/>)  
 Springer ([www.springerlink.com](http://www.springerlink.com))  
 ACM Digital Library ([www.acm.org/dl](http://www.acm.org/dl))  
 ScienceDirect ([www.sciencedirect.com](http://www.sciencedirect.com))  
 Wiley Interscience ([www.interscience.wiley.com](http://www.interscience.wiley.com))  
 Google Scholar ([www.scholar.google.co.in](http://www.scholar.google.co.in))  
 IEEE eXplore ([www.ieeexplore.ieee.org](http://www.ieeexplore.ieee.org))  
 Taylor Francis Online ([www.tandfonline.com](http://www.tandfonline.com))  
 PubMed (<https://pubmed.ncbi.nlm.nih.gov/>)  
 SemEval (<https://semeval.github.io/>)

Table 1. Mapping Selected Papers to Our Roadmap

Paper	Reproducibility		Alignment Approach		Training Signal		Training Signal		Level of Alignment		Taxonomy According to Figure 2
	Code	Dataset	Off-line	On-line	Bilingual Dictionary	Parallel Corpora	off-the-shelf methods	During Alignment	Word-level	Sentence-level	
[21] Aldarmaki et al. 2019	☐	☑	●	○	○	●	●	○	●	●	(a)
[85] Schuster et al. 2019	☐	☑	●	○	○	○	○	●	●	○	(a)
[85] Schuster et al. 2019	☐	☑	●	○	○	○	○	●	●	○	(b)
[96] Wieting et al. 2019	☐	☑	○	○	○	○	○	●	○	○	(f)
[94] Wang et al. 2019 - CLBT	☐	☑	○	○	○	○	○	●	○	○	(a)
[71] Artetxe et al. 2019 - LASER	☐	☑	○	○	○	○	○	●	○	○	(f)
[12] Cao et al. 2020	☐	☑	○	○	○	○	○	○	○	○	(c)
[95] Wang et al. 2020	☐	☑	○	○	○	○	○	○	○	○	(c)
[99] Wu et al. 2020	☐	☑	○	○	○	○	○	○	○	○	(c)
[13] Chi et al. 2020 - XNLI	☐	☑	○	○	○	○	○	○	○	○	(f)
[80] Reimers et al. 2020	☐	☑	○	○	○	○	○	○	○	○	(f)
[73] Pan et al. 2021	☐	☑	○	○	○	○	○	○	○	○	(d)
[42] Hu et al. 2021 - AMBER	☐	☑	○	○	○	○	○	○	○	○	(f)
[3] Alqahtani et al. 2021	☐	☑	○	○	○	○	○	○	○	○	(d)
[14] Chi et al. 2021 - infoXLM	☐	☑	○	○	○	○	○	○	○	○	(f)
[104] Zhao et al. 2021	☐	☑	○	○	○	○	○	○	○	○	(c)
[36] Goswami et al. 2021 - DuEAM	☐	☑	○	○	○	○	○	○	○	○	(e)
[37] Gratta et al. 2021 - XeroAlign	☐	☑	○	○	○	○	○	○	○	○	(f)
[91] Ul'car et al. 2022 - Vecmap/MUSE	☐	☑	○	○	○	○	○	○	○	○	(a)
[91] Ul'car et al. 2022 - ELMOGAN	☐	☑	○	○	○	○	○	○	○	○	(a)
[60] Liu et al. 2022 - Bi-SaELMO	☐	☑	○	○	○	○	○	○	○	○	(e)
[90] Tien et al. 2022	☐	☑	○	○	○	○	○	○	○	○	(d)
[29] Feng et al. 2022 - LaBSE	☐	☑	○	○	○	○	○	○	○	○	(f)
[24] Ding et al. 2022 - EAR	☐	☑	○	○	○	○	○	○	○	○	(d)
[41] Heffernan et al. 2022 - LASER3	☐	☑	○	○	○	○	○	○	○	○	(b), (f)
[39] Hämmerl et al. 2022	☐	☑	○	○	○	○	○	○	○	○	(a), (d)
[26] Efimov et al. 2023	☐	☑	○	○	○	○	○	○	○	○	(c)
[11] Abulkhanov et al. 2023 - LAPCA	☐	☑	○	○	○	○	○	○	○	○	(f)
[89] Tan et al. 2023 - LASER3-CO	☐	☑	○	○	○	○	○	○	○	○	(b), (f)
[56] Li et al. 2023	☐	☑	○	○	○	○	○	○	○	○	(f)
[54] Li et al. 2024 - AFP	☐	☑	○	○	○	○	○	○	○	○	(d)
[92] Vasilyev et al. 2024	☐	☑	○	○	○	○	○	○	○	○	(f)
[48] Jiang et al. 2024 - CLASS	☐	☑	○	○	○	○	○	○	○	○	(c)
[8] Bakos et al. 2025 - AlignFreeze	☐	☑	○	○	○	○	○	○	○	○	(c)
[90] Feng et al. 2025 - IDA	☐	☑	○	○	○	○	○	○	○	○	(d)
[59] Liu et al. 2025	☐	☑	○	○	○	○	○	○	○	○	(d)
[88] Sundar et al. 2025	☐	☑	○	○	○	○	○	○	○	○	(d)
[47] Jha et al. 2025 - vec2vec	☐	☑	○	○	○	○	○	○	○	○	(b)
[22] Deng et al. 2025	☐	☑	○	○	○	○	○	○	○	○	(d)

(Code)→ Not provided: ☐, Provided no documentation: ☐, Provided with documentation: ☐ (Dataset) → Not mentioned: ☐, Private dataset: ☐, Public dataset: ☐; (Rest of features) → Not mentioned: ○, Applied: ●



Fig. 6. Category (a) - Offline Independent Embeddings w Alignment via anchor selection.



Fig. 7. Category (b)—Independent embeddings with alignment w/o direct anchor selection.

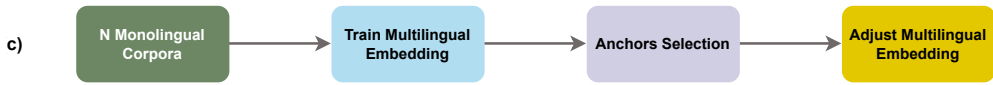


Fig. 8. Category (c)—Online multilingual embedding adjustment w anchors.

aligning two language models at the word level, relying on bilingual dictionaries as a source of supervision; to do this, they used the averaged contextual word embeddings as an anchor for each word type. These approaches incorporate anchors to solve an optimisation problem aimed at discovering an appropriate linear transformation, as detailed in Section 4.1.

## 6.2 Category (b)—Independent Embeddings with Alignment w/o Direct Anchor Selection

Approaches classified under category (b) follow the process depicted in Figure 7. In this category, we highlight only two papers. The first, by Schuster et al. [85], proposes aligning two contextual embedding models using the adversarial MUSE framework introduced by Lample et al. [53], without relying on parallel text. In MUSE, anchor points are typically selected by comparing the similarity of semantic distributions among terms in the corpora to be aligned. However, in this case, no anchor points are required. Within category (d), some studies adopt the same approach, but focus on aligning multilingual embeddings rather than multiple monolingual ones. The second article, by Jha et al. [47], presents an unsupervised post-hoc mapping method that aligns embedding spaces via a “universal” latent geometry—a universal semantic structure postulated by the Platonic Representation Hypothesis [45]. Two encoders map each source space into the universal space and back, using a GAN-based strategy, enabling cross-lingual alignment without anchors or parallel data by leveraging the assumed shared semantic geometry.

## 6.3 Category (c) - Online Multilingual Embedding Adjustment w Anchors

Approaches classified under category (c) follow the process depicted in Figure 8. These approaches begin with models that are potentially not trained on parallel data, so they undergo a remapping process. In the cases of Cao et al. [12] and Zhao et al. [104], alignment is integrated into the loss function. Their strategy adopts more expressive alignment methods than rotation, mitigating the isomorphism and isometry challenge. They fine-tune their models using parallel data and leverage it to align the contextual embeddings across different languages, ensuring that words with similar meanings have corresponding embeddings in the multilingual space. Similar to Cao et al. [12], Efimov et al. [26] adapt the pre-trained multilingual model mBERT by using a limited parallel corpus to enhance its cross-lingual transfer capabilities. They work with one language pair at a

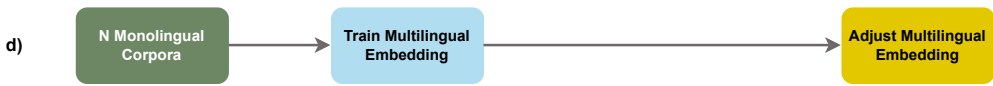


Fig. 9. Category (d)—Direct multilingual embedding adjustment.

time, whereas Cao adapted mBERT for five languages simultaneously. Wang et al. [95] adjust the parameters of an unsupervised joint training method, such as mBERT, which does not explicitly employ any dictionaries or alignment mechanisms. Consequently, the resulting embedding set may be coarse and misaligned in the shared vector space. To refine alignments across non-shared embedding sets, they apply an off-the-shelf alignment method as a final step. Wu et al. [99] perform adjustments using a novel contrastive alignment objective. Instead of optimising for the absolute distance between source and target terms, the contrastive loss offers more flexibility by encouraging source and target words to be closer to each other than any other hidden states. Finally, CLASS method [48] involves online pre-training of a multilingual model where cross-lingual links and anchor texts serve as explicit supervisory signals to guide the alignment of representations.

#### 6.4 Category (d) - Direct Multilingual Embedding Adjustment

Approaches classified under category (d) follow the process depicted in Figure 9. Pan et al. [73] introduce a straightforward approach to aligning multilingual contextual embeddings as a post-pretraining step, enhancing the zero-shot cross-lingual transferability of pre-trained models. They utilise parallel data to perform alignment at both the word level, using the **Translation Language Modeling (TLM)** objective, and at the sentence level through contrastive learning and random input shuffling. Alqahtani et al. [3], on the other hand, suggest using **Optimal Transport (OT)** as an alignment objective during fine-tuning to enhance multilingual contextualised representations further. Notably, this approach does not necessitate an explicit predefined set of anchors. Tien et al., as described in their work [90], have created unsupervised models trained on unpaired sentences and supervised models trained on bitexts. Ding et al. [24] propose three strategies to enhance cross-lingual embeddings without relying on word-alignment pairs during fine-tuning. The first, **Embedding-Push**, moves English embeddings closer to other language clusters. The second, **Attention-Pull**, maintains relative word positions to preserve meaning. The final strategy, **Robust Target**, introduces a Virtual Multilingual Embedding to create a stable embedding space. The **Align aFter Pre-training (AFP)** [54] framework improves multilingual generative models by aligning sentence representations across languages after the pre-training phase. It leverages translation pairs and contrastive learning to align internal representations, enhancing cross-lingual capabilities without requiring model retraining. This alignment is achieved by encouraging isomorphic representations, without relying on explicit geometric assumptions. These models are built upon the unsupervised language model XLM-RoBERTa (XLM-R) [17], with the model’s parameters kept consistent. AlignFreeze [8] also builds on these models, applying a “realignment” strategy and finding that freezing the lower layers helps prevent performance degradation in cross-lingual transfer. Feng et al. [30] also focus on the alignment of different layers within **large language models (LLMs)**, demonstrating that word-level embeddings in the hidden layers are isomorphic across languages. They show that the hidden states corresponding to inputs in different languages can be aligned at the word level using an orthogonal transformation. Similarly, Liu et al. [59] propose a fine-tuning approach that alternates between a task loss and a contrastive alignment loss applied to middle layers of LLMs. Using parallel sentence pairs, the alignment loss encourages cross-lingual consistency in hidden states while preserving task performance. The loss proposed maximises the



Fig. 10. Category (e) - Online anchored multilingual embedding training and alignment.



Fig. 11. Category (f) - Direct parallel training and alignment.

similarity between translations while minimising similarity between non-translations. Completely unlike the methods mentioned above, Sundar et al. [88] propose a model-intervention strategy to enhance cross-lingual alignment without parallel data by selectively modifying neuron activations in multilingual LLMs. By identifying and steering “expert” neurons, they reshape the embedding space to achieve tighter multilingual alignment. In Reference [22] the authors introduce a rewiring strategy that adapts LLMs to cross-lingual tasks. The method leverages the autoregressive structure of embeddings, compressing and re-aligning them to better capture language-invariant patterns. By modifying internal connections rather than relying on parallel data or external dictionaries, it enables unsupervised cross-lingual transfer in a lightweight and efficient manner.

### 6.5 Category (e) - Online Anchored Multilingual Embedding Training and Alignment

Approaches classified under category (e) follow the process depicted in Figure 10. Liu et al. [60] developed an innovative approach to align contextual embeddings at the sense level using cross-lingual signals derived solely from bilingual dictionaries, thereby eliminating the need for parallel corpora. They achieve this by introducing a novel sense-aware cross-entropy loss that explicitly models different senses of a word based on its context. This is made possible through a clustering analysis that identifies distinct word senses, effectively addressing the anisomorphism’s challenge. In their cross-lingual model pre-training, they incorporate a sense alignment objective combined with this sense-aware cross-entropy loss. Goswami et al. [36] use a multitask loss function to capture semantic similarity and relatedness between sentences, training a dual-encoder model to map different languages into a shared vector space. Their dual-encoder architecture leverages word-level semantic similarity scores, embedding these into unified sentence-level vectors.

### 6.6 Category (f) - Direct Parallel Training and Alignment

Approaches classified under category (f) follow the process depicted in Figure 11. Wieting et al. [96] constructed a model to acquire paraphrastic sentence embeddings directly from bilingual text data. Their training dataset comprises pairs of sentences in source and target languages. For each sentence pair, they randomly select a non-matching target sentence during training. The primary objective is to make source and target sentences more similar than source and negative target examples by a specific margin. To understand this category, we need first to describe XLM proposed by Conneau et al. [18], as it serves as the baseline for most of the papers in this field. Conneau et al. [18] explore cross-lingual language model pre-training using three approaches: **Causal Language Modeling (CLM)**, MLM, or MLM combined with TLM (Translation Language Modeling). The TLM objective extends MLM by concatenating parallel sentences and randomly masking words in both the source and target sentences. To predict a masked word in a source sentence, the model can

attend to nearby source words or the target translation, promoting alignment between source and target representations. The work is not included in our set of selected papers because it lacks an explicit objective for aligning contextual embeddings. In their work [42], Hu and colleagues put forward a training approach to acquire contextualised word representations. This approach fosters symmetry in training, encompassing both word and sentence levels. Like Conneau et al. [18], they employ MLM but also introduce a sentence alignment objective to encourage the model to predict the correct translation of a target sentence when provided with a source sentence. Additionally, they incorporate a word alignment objective by leveraging the Transformer model's attention mechanism. Artetxe et al. (LASER) [7] employ a BiLSTM encoder to transform input sentences into fixed-length vector representations, initialising the decoder LSTM responsible for generating target sentences. They train both the encoder and decoder using parallel corpora with a translation objective. In contrast to Conneau et al. [18], their approach is designed to scale to a larger number of languages. Chi et al. [13] propose pre-training both the encoder and decoder of a sequence-to-sequence model under both monolingual and cross-lingual conditions. They adopt the MLM approach used by Hu et al. [42] and Conneau et al. [18] but also incorporate a **denoising auto-encoding (DAE)** objective to reconstruct the original text from the corrupted text. They apply these two objectives to both monolingual and parallel input datasets. In their 2021 work (infxlm), Chi et al. [14] introduce a novel pre-training task based on contrastive learning. They consider a bilingual sentence pair as two viewpoints of the same meaning and promote the similarity of their encoded representations compared with negative examples. By using both monolingual and parallel corpora, they collectively train these preliminary tasks to improve the cross-lingual transfer capabilities of pre-trained models. Feng et al. [29] introduced an innovative approach that enhances translation ranking performance through a unique blend of pre-training and dual-encoder fine-tuning. Specifically, they merged dual encoders, which were trained using a translation ranking loss to maximise the similarity of translation pairs within a common embedding space, with encoders that were initialised using large pre-trained language models. Abulkhanov et al. [1] introduce a novel approach to cross-lingual retrieval, utilising cross-lingual pre-training and fine-tuning for cross-lingual information retrieval tasks with loosely aligned data automatically mined from Wikipedia. Gritta and Iacobacci [37] propose XeroAlign, a method for task-specific alignment of cross-lingual pre-trained transformers like XLM-R. XeroAlign incorporates an auxiliary training objective that leverages translated data to improve target language performance, bringing it closer to that of the source (labelled) language. Li et al. [56] propose a method based on Cross-Lingual Representation Similarity (XLRs) that aligns multilingual representations during training to enhance zero-shot generation. By using multiple source languages, the method regularises representation similarity, helping to prevent language-specific errors. Interestingly, the study finds that neutral representations can actually degrade performance on generation tasks—challenging the common assumption that such invariance universally benefits cross-lingual transfer across all downstream tasks. Finally, Vasilyev et al. [92] consider a simple linear cross-lingual mapping as a possible improvement of the multilingual embeddings.

## 6.7 Composition of Categories

Heffernan et al. [41] and Reimers & Gurevych [80] take opposite approaches to the teacher-student framework for multilingual embeddings. Reimers et al. [80] use monolingual embeddings as a teacher model to align multilingual embeddings, aiming to improve cross-lingual knowledge transfer and reduce linguistic bias, thereby making the representation space more isotropic. So, this approach is a combination of (f) and (b). In contrast, Heffernan et al. use multilingual embeddings as the teacher to align monolingual embeddings, with a focus on scaling encoder training and bitext mining for low-resource languages that are not well-covered by existing models. So, in this case,

the combination is the opposite (b) and (f). Tan et al. [89] follow Heffernan et al.'s approach [41] but extend it by integrating contrastive learning into their distillation method, making it more effective for training encoders for low-resource languages. Additionally, H"ammerl et al. [39] are classified as a combination of two taxonomy categories. They blend the advantages of static and contextual models, exploring their mutual benefits. Specifically, they extract static embeddings for 40 languages from XLM-R, validate them with cross-lingual word retrieval, and align them using VecMap [5]. They further apply a novel continued pre-training approach to XLM-R, leveraging the high-quality alignment from static embeddings to better align XLM-R's representation space. Through this intuition, the method addresses the challenges posed by isometry and isomorphism assumptions in contextualised embeddings, where such issues are particularly problematic.

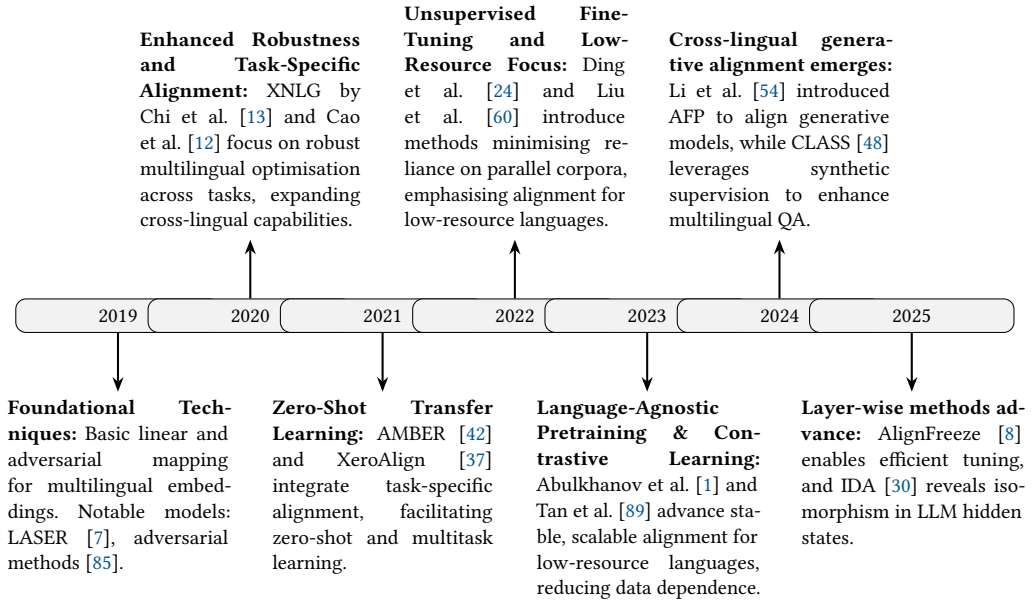
## 6.8 Timeline

Here's a breakdown of the main characteristics and evolutionary trends for each year in cross-lingual alignment from 2019 to 2025:

- Before 2019:** Looking at the technique for align embedding like the one proposed by Lample et al. [53], S"ogaard et al. [87] question the assumption that monolingual embedding spaces exhibit approximate isometry, showing that this is not valid for many language pairs.
- 2019:** Initial focus on establishing foundational cross-lingual alignment methods with tools like LASER for multilingual embeddings [7] and adversarial alignment approaches by Schuster et al. [85]. This year introduced basic linear and adversarial mapping techniques to bring embeddings from different languages into a shared space, setting the groundwork for more sophisticated transformations and pre-training models.
- 2020:** Emphasis on expanding model robustness and task-specific alignment. Cao et al. [12] and Chi et al. [13] developed alignment methods using multilingual objectives, exploring ways to optimise models across both monolingual and cross-lingual tasks. Pre-training large multilingual models like XLM-R also became a focal point, marking a shift towards enhancing performance in multiple languages simultaneously.
- 2021:** Advances in embedding alignment for complex tasks and zero-shot applications. New methods like AMBER [42] and XeroAlign [37] emphasised task-specific alignment goals such as zero-shot transfer learning and semantic similarity across languages. This year marked a push towards fine-grained semantic alignment and transfer capabilities, integrating multitask learning to capture semantic similarity.
- 2022:** Focus on fine-tuning and unsupervised approaches for enhanced transfer learning. Ding et al. [24] and Liu et al. [60] introduced methods targeting unsupervised alignment (e.g., bilingual dictionaries for sense-aware embeddings), which reduced reliance on parallel corpora. This shift reflected the industry's interest in scaling models for low-resource languages by minimising data requirements and optimising alignment through cross-lingual objectives.
- 2023:** Growth in language-agnostic pre-training and contrastive learning integration. Abulkhanov et al. [1] and Tan et al. [89] integrated contrastive learning in LASER3 for low-resource languages, enhancing alignment stability and transferability in limited-data contexts. This year emphasised scalability and refinement of cross-lingual alignment with techniques like LAPCA, as models evolved to address language diversity with minimal supervised data.
- 2024:** Emergence of cross-lingual generation alignment. Li et al. [54] introduce AFP to enhance multilingual generative models without retraining, while Jiang et al. [48] develop CLASS using cross-lingual links as explicit supervisory signals. Vasilyev et al. [92] demonstrate that simple linear mappings remain effective for sentence embeddings.
- 2025:** Layer-specific alignment advancements include Bakos et al.[8], who introduce Align-Freeze with layer-freezing strategies to mitigate performance degradation, and Feng et al.[30],

who reveal word-level isomorphism in LLM layers through iterative decomposition. New adaptation strategies also emerge, such as Deng et al. [22], which rewrites autoregressive embeddings for unsupervised cross-lingual transfer, and Jha et al. [47] present Vec2Vec, an unsupervised post-hoc mapping method that leverages a universal latent geometry and GAN-based projections to align embedding spaces without anchors or parallel data.

Each year shows a trend toward increasing model sophistication, scalability, and data efficiency, particularly in handling low-resource languages and advancing zero-shot capabilities.



## 6.9 Comprehensive Discussion of Methodological Challenges

Although alignment methods have achieved significant success, they still have some notable drawbacks. As mentioned by Hämmerl et al. [40] it is possible to consider 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. Offline methods in categories (a) and (b), as well as online models in categories (c) and (d), rely on two separately trained embeddings. In contrast, recent research in online joint training highlights the advantages of word sharing during the training phase. This distinction may result in information loss in the final embedding and impact the effectiveness of fine-tuning aligned embeddings for downstream tasks. The absence of cross-lingual objectives during fine-tuning can lead to suboptimal results, unlike jointly trained models where shared words can facilitate this role. Off-line alignment methods hinge on the assumption of isomorphism in monolingual embedding spaces. However, studies such as those by Sogaard et al. [87], have contested this assumption, revealing that it doesn't hold true for many language pairs. Notably, Ormazabal et al. [71] suggest that this limitation arises from the independent training of the two sets of monolingual embeddings. Conversely, the on-line joint training methods that belong to categories (e) and (f), are simpler and avoid the disadvantages above of off-line alignment methods. Nevertheless, it carries its own set of limitations. These methods assume that all shared words between two languages implicitly serve as anchors and need not be aligned with other words. However, this assumption does not

always hold true, leading to misalignment. For instance, the English word “the” will likely appear in the Spanish training corpus, but ideally, it should align with Spanish words like “el” and “la” instead of aligning with itself. This issue is referred to as oversharing. The methods discussed in the survey involve joint training with the incorporation of an explicit cross-lingual task, enabling the alignment of words that are not shared between languages. Methods that adapt an already multilingual model trained in this manner strike a perfect balance, as they consider words within their language domain without sharing during the pre-training phase but remap only those words present in the source used as a signal. In contrast, sources used during the adjustment phase do not have as comprehensive an impact as when they are used in the pre-training phase.

## 7 Evaluation Tasks

In this section, we will describe the downstream task utilised by researchers to evaluate the proposed models. This information can assist researchers in identifying a benchmark task that is suitable for evaluating their models among the ones that are widely recognised in the community. Moreover, in Table 2 we report the evaluation tasks used by the surveyed papers. This can be useful for comparison. In our view, the greater the variety of tasks used to evaluate the authors’ proposed method, the more robust the method appears to be.

**Sentence Translation Retrieval (STR):** STR entails retrieving the accurate translation from the target side of a test parallel corpus by employing the nearest neighbour search based on cosine similarity.

**Bilingual Lexicon Induction (BLI):** BLI task is meant to quantify vocabulary induction performance, that is, given a benchmark set of words in a source language, find the translation in the target language.

**Dependency parsing (DP):** DP is a task that involves analysing the grammatical structure of a sentence to determine the relationships between words. In this task, each word in a sentence is assigned a syntactic label describing its relationship to other words. The resulting structure is typically represented as a tree, where each word is a node and the relationships between words are represented as edges.

**Part-of-speech (POS) Tagging:** POS tagging is the process of assigning a grammatical category (such as noun, verb, adjective, etc.) to each word in a sentence. Differently from dependency parsing, it does not analyse the relationships or dependencies between words in a sentence.

**Paraphrase Detection (PD):** PD identifies whether two sentences or phrases convey the same or similar meaning, even though they may be expressed in distinct linguistic forms.

**Natural Language Inference (NLI):** NLI is the process of establishing the connection between two given sentences in a specific order and categorising them as either showing entailment, contradiction, or having “no relation” between them.

**Sentiment Analysis (SA):** In the field of SA, the goal is to automatically identify and categorise the emotional tone expressed in a piece of text. This task typically involves classifying a sentence, a short passage, or even a full document into one of several predefined categories—most commonly positive, negative, or neutral—based on the sentiment conveyed.

**Named Entity Recognition (NER):** NER involves information extraction to identify and categorise named entities within unstructured text into predefined categories. These categories may include person names, organisations, locations, medical codes, time expressions, quantities, monetary values, and more.

**Semantic Textual Similarity (STS):** STS involves comparing pairs of sentences and assigning a score to each pair that reflects the degree of similarity between the two sentences. Human judges typically assign the scores based on their subjective assessment of the sentence

Table 2. Downstream Tasks Used by Authors for Evaluating their Models

Paper	Downstream Tasks													
	NLP			Text Similarity			Translation				Classification		Reasoning	
	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[2] Aldarmaki et al. 2019	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[85] Schuster et al. 2019	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[96] Wieting et al. 2019	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[94] Wang et al. 2019 - CLBT	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[7] Artetxe et al. 2019 - LASER	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[12] Cao et al. 2020	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[95] Wang et al. 2020	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[99] Wu et al. 2020	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[13] Chi et al. 2020 - XNLG	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[80] Reimers et al. 2020	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[73] Pan et al. 2021	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[42] Hu et al. 2021 - AMBER	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[3] Alqahtani et al. 2021	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[14] Chi et al. 2021 - infoXLM	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[104] Zhao et al. 2021	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[36] Goswami et al. 2021 - DuEAM	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[37] Gritta et al. 2021 - XeroAlign	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[91] Ulčar et al. 2022 - Vecmap/MUSE	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[91] Ulčar et al. 2022 - ELMOGAN	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[60] Liu et al. 2022 - Bi-SaELMO	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[90] Tien et al. 2022	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[29] Feng et al. 2022 - LaBSE	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[24] Ding et al. 2022 - EAR	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[41] Heffernan et al. 2022 - LASER3	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[39] Hämmerl et al. 2022	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[26] Efimov et al. 2023	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[1] Abul Khanov et al. 2023 - LAPCA	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[89] Tan et al. 2023 - LASER3-CO	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[56] Li et al. 2023	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[54] Li et al. 2024 - AFP	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[92] Vasilyev et al. 2024	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[48] Jiang et al. 2024 - CLASS	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[8] Bakos et al. 2025 - AlignFreeze	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[30] Feng et al. 2025 - IDA	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[59] Liu et al. 2025	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[88] Sundar et al. 2025	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[47] Jha et al. 2025 - vec2vec	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	QA
[22] Deng et al. 2025	NER	POS	DP	STS	PD	SS	STR	BM	PSM	RFEval	DC	SA	NLI	MT

Task categories: NLP (Named Entity Recognition (NER), Part-of-Speech Tagging (POS), Dependency Parsing (DP)), Text Similarity (STS, Paraphrase Detection (PD), Similarity Search (SS)), Translation (Sentence Translation Retrieval (STR), Bitext Mining (BM), Parallel Sentence Matching (PSM), Reference-free machine translation evaluation (RFEval)), Classification (Document Classification (DC), Sentiment Analysis (SA)), and Reasoning (Natural Language Inference (NLI), Question Answering (QA)).

similarity. The evaluation metric for the STS task is usually Pearson's correlation coefficient between the predicted scores and the gold standard scores assigned by human judges.

**Similarity Search (SS):** SS consists of finding similar documents or text segments based on a query, often using vector representations of text. This task is crucial for applications like information retrieval, recommendation systems, and content-based filtering.

**Bitext Mining (BM):** The BM task aims to identify parallel sentence pairs within two large monolingual corpora in different languages. These sentence pairs are commonly referred to as “gold” alignments when used as ground truth for evaluation.

**Bilingual Terminology Alignment (BTA):** The task of BTA involves aligning terms found in two distinct languages between two candidate term lists.

**Question Answering (QA):** QA task in the cross-lingual domain involves answering questions in different languages based on a given context.

**Machine Reading Comprehension (MRC):** MRC is a variation of the QA task where the system is tasked with responding to the form of a continuous sequence of tokens from a text paragraph, given a question and that paragraph.

**Document Classification (DC):** DC is a task where a model is trained to classify a given document into one of several predefined categories. So, in the cross-lingual domain, document classification refers to the task of classifying documents in different languages into the same set of categories.

**Reference-free machine translation evaluation (RFEval):** RFEval task assesses the quality of translation, specifically measuring the similarity between a sentence in the target language and the corresponding sentence in the source language.

**Parallel Sentence Matching (PSM):** PSM consists of identifying sentence pairs in different languages that are mutual translations or convey equivalent meanings.

Many authors used the Xtreme [43]<sup>4</sup> benchmark for evaluating their methods on different tasks as the one listed above.

In Table 2 we report the downstream tasks used by authors classified in this review. The performance of the cross-lingual alignment methods described is presented in Table 3. We present the most comprehensive results, encompassing the widest range of languages and the highest level of generalisation. Specifically, we report average results across all languages used in the evaluation. Additionally, for each proposed model, only the version with the highest recorded performance is included; if a study proposed multiple versions of the same model, we have reported only the best-performing one.

## 8 Application of Cross-Lingual Alignment

Finally, in this section, we report some examples of research that use the methods classified in the article for solving complex tasks. Yi et al. [101] introduce a highly efficient method for webpage snippet extraction, termed DeepQSE, to identify a concise set of sentences that effectively summarise the webpage content within the context of a given input query. They employ XML-RoBERTa [14] for the initial configuration of their language model. Also, Cho et al. [16] employ infoXLM [14] and other models to construct a multimodal intelligent document processing framework. This framework combines a pre-trained deep learning model with traditional robotic process automation techniques commonly used in banking to automate business processes based on real-world financial document images. Dadu and colleagues [20] present a cross-lingual inductive method for detecting offensive language in tweets, leveraging the contextual word embeddings provided by XLM-R. [17]. Ils et al. [46] utilise models from Zhao et al. [104] and Cao et al. [12] to examine the shifts in European solidarity discourses occurring before and after the global declaration of the COVID-19 outbreak as a pandemic. According to Li et al. [55], developing intelligent solutions to help e-commerce sellers provide local products to a global consumer base is paramount. To address this need, they leverage the methodology introduced by Schuster et al. [85] and embark on a novel endeavour in cross-lingual information retrieval. Specifically, they focus on cross-lingual

<sup>4</sup><https://github.com/google-research/xtreme>

Table 3. Performance of Cross-Lingual Alignment Methods

Paper	Task	Dataset	Metric used	Metric values (%) with comments	
[39]	POS	UD-POS	F1 Score	72.1	
[42]			F1 Score	70.5	
[99]		Universal Dependencies v2.3	F1 Score	81.5 ± 0.6 <sup>a</sup>	
[8]		Universal Dependencies	Accuracy	81.7 <sup>b</sup>	
[94]	DP	UD Treebanks v2.2	LAS	63.54 <sup>c</sup>	
[85]		UD v2.0 Treebanks	LAS	77.3 <sup>d</sup>	
[91]		Universal Dependencies	UAS / LAS	57.2 <sup>e</sup> / 33.5 <sup>e</sup>	
[91]			UAS / LAS	75.0 <sup>e</sup> / 54.7 <sup>e</sup>	
[91]			UAS / LAS	73.4 <sup>e</sup> / 53.1 <sup>e</sup>	
[99]		Universal Dependencies v2.6	LAS	57.4 ± 0.5 <sup>f</sup>	
[60]		CoNLL-2002, CoNLL-2003	F1 Score	75.11 <sup>g</sup>	
[99]		Pan et al., 2017	F1 Score	66.2 ± 1.0 <sup>h</sup>	
[39]	NER	PAN-X	F1 Score	62.73	
[26]		Wikiann	F1 Score	69.1 <sup>i</sup>	
[91]			F1 Score	46.6 <sup>j</sup>	
[91]			F1 Score	52.4 <sup>j</sup>	
[91]			F1 Score	38.3 <sup>j</sup>	
[8]			F1 Score	84.8 <sup>k</sup>	
[41]			SS	Tatoeba	F1 Score
[80]		Accuracy		53.4	
[96]	STS	SemEval STS 2017	Pearson Corr.	79.62 <sup>l</sup>	
[29]			Pearson Corr.	72.8	
[36]		STS Benchmark	Pearson Corr.	74.5 <sup>m</sup>	
[80]			Spearman Corr.	83.7 <sup>n</sup>	
[22]			Spearman Corr.	83.81 <sup>o</sup>	
[104]			Pearson Corr.	48.1 <sup>p</sup>	
[56]		WMT <sup>13</sup>	ROUGE-L	27.7 <sup>q</sup>	
[59]		MTG	BLEU/COMET	17/80.7 <sup>r</sup>	
[88]		WMT <sup>23</sup>	Accuracy	32 <sup>s</sup>	
[88]		FLORES200	Accuracy	32 <sup>s</sup>	
[47]	NQ / TweetTopic / MIMIC / Enron	Cosine Similarity (×100)	86.0 <sup>t</sup>		

(Continued)

<sup>a</sup>Average and Standard Deviation of 10 runs<sup>b</sup>Average POS tagging accuracy over 34 languages using XLM-R Base with ALIGNFREEZE (front-freezing); see Table 2.<sup>c</sup>Average on 18 languages<sup>d</sup>Average on De Es Fr It Pt Sv<sup>e</sup>Average on 20 language pairs across 9 languages: En Hr Et Fi Lv Lt Ru Sl Sv<sup>f</sup>Average and Standard Deviation of 10 runs<sup>g</sup>Average on 3 languages, es, nl, and de<sup>h</sup>Average and Standard Deviation of 10 runs<sup>i</sup>Adj+cont model average on 4 languages: Es Hi Ru Vi<sup>j</sup>Average for 19 language pairs across 9 languages: En Hr Et Fi Lv Lt Ru Sl Sv<sup>k</sup>F1 Score on WikiANN averaged over 34 languages using XLM-R Base with ALIGNFREEZE (front-freezing); see Table 2 of the paper.<sup>l</sup>Average on 4 languages: Ar Es En Tr<sup>m</sup>Average on 5 languages: En De Tr Es Fr<sup>n</sup>Average on 8 languages: En Ar De Tr Es Fr It Nl<sup>o</sup>Average on 10 STS datasets: STS12, STS13, STS14, STS15, STS16, STS17, STS22, STS-B, BIOSSES, SICK-R<sup>p</sup>Average for 19 languages grouped in 4 groups of typologically similar ones<sup>q</sup>avg on en, es, de, fr, zh<sup>r</sup>avg on en, he, ja, uk using LLaMA 3<sup>s</sup>improvements in top-1 retrieval accuracy. Median on 22 languages using Aya-8b model, for the intervention on Spanish<sup>t</sup>Average cosine similarity across in-distribution (NQ) and out-of-distribution datasets (TweetTopic, MIMIC, Enron)<sup>ab</sup>Result related to Tatoeba Dataset. Average of all languages supported by Tatoeba

Table 3. Continued

Paper	Task	Dataset	Metric used	Metric values (%) with comments
[37]	PD	PAWS-X	Accuracy	93.6 <sup>u</sup>
[42]			Accuracy	89.2
[54]			Accuracy	57.3 <sup>xy</sup>
[56]			ROUGE-L	29.5 <sup>v</sup>
[29]		SentEval MRPC	F1 Score	74.4
[1]	BM	BUCC	F1 Score	83.5 <sup>w</sup>
[7]			F1 Score	93.9 <sup>x</sup>
[36]			F1 Score	81.7 <sup>x</sup>
[80]			F1 Score	88.6 <sup>y</sup>
[90]			F1 Score	92.8 <sup>z</sup>
[96]			F1 Score	77.15 <sup>aa</sup>
[41]		FLORES	xsim error rate	0.6 <sup>ab</sup>
[95]		MUSE	precision	74.5 <sup>ac</sup>
[89]		Paracrawl	BLEU	9.28
[29]			BUCC, Tatoeba	Accuracy/F1 Score
[14]	STR	Tatoeba	Accuracy	79.2 <sup>af</sup>
[39]			Accuracy	68.1
[42]			Accuracy	87.9
[90]			Accuracy	80.4
[95]		WMT	BLEU	22.59 <sup>ag</sup>
[2]		WMT'13	Accuracy	84.0 <sup>ah</sup>
[1]	PSM	XOR-Retrieve	Recall@2000 / Recall@5000	65.0 <sup>w</sup> / 70.5 <sup>w</sup>
[48]				71.6 <sup>ai</sup> / 78.2 <sup>ai</sup>
[47]		NQ / TweetTopic / MIMIC / Enron	Accuracy	82.0 <sup>aj</sup>
[36]		Tatoeba	Accuracy	77.7 <sup>ak</sup>
[92]	WikiNews	Cosine Improvement (fC)	0.991 <sup>al</sup>	
[92]	Tatoeba	Distance Reduction (dD)	0.192 <sup>am</sup>	
[30]	MUSE	Precision@1	74.5 <sup>an</sup>	
[7]	MLDoc	Accuracy	72.8 <sup>ao</sup>	
[37]	DC	MultiATIS++	Accuracy/F1 Score	96.0 <sup>ap</sup> / 81.2 <sup>ap</sup>
[24]		XNLI	Accuracy	68.7 <sup>aq</sup>

(Continued)

<sup>u</sup>Average on 7 languages: En De Es Fr Ja Ko Zh<sup>v</sup>Average on 5 countries: En, Es, De, Fr, Zh<sup>w</sup> Result for  $LAPCA - LM + XPAQ_{large}$  average across 8 languages: Ar Bn Fi Ja Ko Ru Te En<sup>x</sup>Average for 5 languages: En De Ru Fr Zh<sup>y</sup>Average on 5 languages: De En Fr Ru Zh<sup>z</sup>Average on 4 datasets : De Fr Ru Zh<sup>aa</sup>Result related to BUCC Dataset. Average on 3 languages : En De Fr<sup>ab</sup>Average on 12 languages:Amh Be Ga Hy Ka Kk Km Sw Ta Te Ur Uz<sup>ac</sup>Average on 7 languages: En Es Fr De It Ru Zh<sup>ad</sup>Result related to Tatoeba Dataset. Average of all languages supported by Tatoeba<sup>ae</sup>Result related to BUCC Dataset. Average for 5 languages: En De Fr Ru Zh<sup>af</sup>Average 14 languages covered by parallel data (both directions): Ar Bg Zh De El Fr Hi Ru Es Sw Th Tr Ur Vi.<sup>ag</sup>Average on 3 language pairs: En Fr De<sup>ah</sup>Result for ELMO (sent) average across 3 languages: En Es De<sup>ai</sup>Results on XOR-Retrieve dev set. CLASS model with full supervised training.<sup>aj</sup>Average retrieval accuracy across in-distribution (NQ) and out-of-distribution datasets (TweetTopic, MIMIC, Enron)<sup>ak</sup>Average on 10 languages: De, Hi, Zh, El, Af, Te, Tl, Ga, Ka, Am<sup>al</sup>Highest fC on title-text pairs in Russian<sup>am</sup>Maximum distance gain; see Table 1 of the paper<sup>an</sup>Bilingual Lexicon Induction interpreted as PSM evaluation<sup>ao</sup>Average on 8 languages: En De Es Fr It Ja Ru Zh<sup>ap</sup>Average on 8 languages: De Es Fr Tr Hi Zh Pt Ja<sup>aq</sup>Average for 15 languages: En Ar Bg De El Es Fr Hi Ru Sw Th Tr Ur Vi Zh

Table 3. Continued

Paper	Task	Dataset	Metric used	Metric values (%) with comments		
[60]	SA	Amazon Reviews	Accuracy	75.32 <sup>at</sup>		
[29]		SentEval SST	Accuracy	83.8		
[91]		Twitter Sentiment	Accuracy	11.2 <sup>as</sup>		
[91]			Accuracy	10.3 <sup>as</sup>		
[91]			Accuracy	11.2 <sup>as</sup>		
[56]	NLI	WikiLingua	ROUGE-L	22.7 <sup>at</sup>		
[3]		XNLI	F1 Score	67.8 <sup>au</sup>		
[7]			Accuracy	69.9 <sup>av</sup>		
[12]			Accuracy	65.9 <sup>aw</sup>		
[14]			Accuracy	81.4 <sup>ax</sup>		
[26]			Accuracy	71.0 <sup>i</sup>		
[42]			Accuracy	71.6		
[54]			Accuracy	48.0 <sup>ay</sup>		
[73]			Accuracy	66.8 <sup>az</sup>		
[99]			F1 Score	76.1 ± 0.4 <sup>ba</sup>		
[104]			Accuracy	77.6 <sup>bb</sup>		
[8]			Accuracy	73.6 <sup>bc</sup>		
[14]			QA	MLQA	F1 Score / EM	73.6 <sup>bd</sup> / 55.2 <sup>bd</sup>
[73]					F1 Score	78.2 <sup>bc</sup>
[13]				SQuAD 1.1 (English-English QG)	BLEU / Meteor / ROUGE	22.4 <sup>bt</sup> / 24.3 <sup>bt</sup> / 49.2 <sup>bt</sup>
[1]	XOR-Full	F1 Score / EM / BLEU		47.8 <sup>w</sup> / 38.7 <sup>w</sup> / 35.5 <sup>w</sup>		
[3]	XQuAD	F1 Score / EM		63.8 <sup>bg</sup> / 48.8 <sup>bg</sup>		
[26]	XQuAD, MLQA	F1 Score		66.9 <sup>bh</sup>		
[39]	XQuAD	Accuracy		70.9 <sup>bi</sup>		
[48]	XOR-Full	F1 Score / EM		50.1 <sup>bj</sup> / 41.8 <sup>bj</sup>		
[8]	XQuAD	F1 Score / EM		67.2 <sup>bk</sup> / 52.3 <sup>bk</sup>		

set-to-description retrieval within cross-border e-commerce. This task entails aligning product attribute sets in the source language with compelling product descriptions in the target language.

<sup>i</sup> Adj+cont model average on 4 languages: Es Hi Ru Vi

<sup>w</sup> Result for *LAPCA – LM + XPAQ<sub>large</sub>* average across 8 languages: Ar Bn Fi Ja Ko Ru Te En

<sup>ar</sup> Average on Bi-SaELMo across German and Japanese in 6 domains

<sup>as</sup> Average on 38 language pairs across 9 languages: En Hr Et Fi Lv Lt Ru Sl Sv

<sup>at</sup> Average on 17 languages: Ar, Zh, Cs, Nl, En, Fr, Hi, Id, It, Ja, Ko, Pt, Ru, Es, Th, Tr, Vi

<sup>au</sup> Represents the average of both seen and unseen 15 languages: En Bg De El Es Fr Ar Hi Ru Sw Th Tr Ur Vi Zh

<sup>av</sup> Average on 15 languages: En, Fr, Es, De, El, Bg, Ru, Tr, Ar, Vi, Th, Zh, Hi, Sw, Ur

<sup>aw</sup> Average on 6 languages: En, Bg, De, El, Es, Fr

<sup>ax</sup> *INFOXML<sub>LARGE</sub>* Average on 15 languages: En, Fr, Es, De, El, Bg, Ru, Tr, Ar, Vi, Th, Zh, Hi, Sw, Ur

<sup>ay</sup> Considering LLama\_7B + AFP. Average result on 2 languages: En, Zh

<sup>az</sup> Average on 6 languages: En, Ar, De, Es, Hi, Zh. The biggest model, 2M parallel training sentences, is considered

<sup>ba</sup> Average and Standard Deviation of 10 runs

<sup>bb</sup> Average for 19 languages grouped in 4 groups of typologically similar ones

<sup>bc</sup> Average accuracy on 12 languages using XLM-R Base with ALIGNFREEZE (front-freezing); see Table 2 of the paper.

<sup>bd</sup> Average on 7 languages: En Es De Ar Hi Vi Zh

<sup>be</sup> Average on En Fr Es De Bg Ar Zh Hi. The biggest model, 2M parallel training sentences, is considered

<sup>bf</sup> Average on 2 languages: En Zh

<sup>bg</sup> Represents the average on both seen and unseen 11 languages: En De El Es Ar Hi Ru Th Tr Vi Zh

<sup>bh</sup> We report results for Adj+cont model using XQuAD dataset since it embeds all the languages as for the other experiments.

The number reported is an average on 4 languages: Es Hi Ru Vi

<sup>bi</sup> We report results for the dataset XQuAD

<sup>bj</sup> Average on 6 languages: En Ar Es Hi Vi Zh

<sup>bk</sup> Evaluation on XQuAD over 11 languages: En, De, Es, El, Ar, Hi, Ru, Th, Tr, Vi, Zh

## 9 Conclusions

This survey comprehensively overviews cross-lingual embedding models and their applications in various natural language processing tasks. We have classified these models based on their data requirements, objective functions, and alignment methods, and discussed their strengths and limitations. We have also reviewed the evaluation methodologies for cross-lingual embeddings and proposed future research avenues. Overall, our survey highlights the importance of cross-lingual alignment methods for contextualised representation in multilingual settings, where often the availability of labelled data is limited. We have shown that cross-lingual embeddings can improve the performance of downstream tasks such as machine translation, sentiment analysis, and named entity recognition. They can be trained on various data types, including parallel corpora, comparable corpora, and monolingual corpora. However, we have also identified some challenges and open issues, such as the lack of standardised evaluation metrics, the domain adaptation problem, and the need for more fine-grained alignment methods. The main contributions are the provision of a comprehensive taxonomy of cross-lingual embedding models, categorising research works in cross-lingual alignment, and identifying the primary challenges related to cross-lingual language models.

## References

- [1] 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*. 2098–2102.
- [2] Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. 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)*. 3906–3911.
- [3] 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*. 3904–3919.
- [4] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. arXiv:1602.01925. Retrieved from <https://arxiv.org/abs/1602.01925>
- [5] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. 5012–5019.
- [6] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Ed.). 789–798.
- [7] 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 (2019), 597–610.
- [8] Steve Bakos, David Guzmán, Riddhi More, Kelly Chutong Li, Félix Gaschi, and En-Shiun Annie Lee. 2025. AlignFreeze: Navigating the impact of realignment on the layers of multilingual models across diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. 562–586.
- [9] Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explain the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 3563–3574.
- [10] Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm’s translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- [11] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, et al. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19, 2, (1993), 263–311.
- [12] Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.
- [13] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7570–7577.

- [14] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. 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*. 3576–3588.
- [15] Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, He-Yan Huang, and Furu Wei. 2021. 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)*. 3418–3430.
- [16] Seongkuk Cho, Jihoon Moon, Junhyeok Bae, Jiwon Kang, and Sangwook Lee. 2023. A framework for understanding unstructured financial documents using RPA and multimodal approach. *Electronics* 12, 4 (2023), 939.
- [17] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451. DOI : <https://doi.org/10.18653/v1/2020.acl-main.747>
- [18] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 7059–7069.
- [19] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.
- [20] Tanvi Dadu and Kartikey Pant. 2020. Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 2183–2189.
- [21] Simone D’Amico, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanatica, and Filippo Pallucchini. 2024. Alignment of multilingual embeddings to estimate job similarities in online labour market. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.
- [22] Jingcheng Deng, Zhongtao Jiang, Liang Pang, Liwei Chen, Kun Xu, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Following the autoregressive nature of LLM embeddings via compression and alignment. *CoRR* abs/2502.11401 (February 2025). Retrieved from <https://doi.org/10.48550/arXiv.2502.11401>
- [23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI : <https://doi.org/10.18653/v1/N19-1423>
- [24] Kunbo Ding, Weijie Liu, Yuejian Fang, Weiyan 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*. 4372–4380.
- [25] Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 644–648.
- [26] 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 *European Conference on Information Retrieval*. Springer, 51–67.
- [27] Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. arXiv:1809.04686. Retrieved from <https://arxiv.org/abs/1809.04686>
- [28] 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)*. Association for Computational Linguistics.
- [29] 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)*. 878–891.
- [30] Zihao Feng, Hailong Cao, Wang Xu, and Tiejun Zhao. 2025. Word-level cross-lingual structure in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 2026–2037.
- [31] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tiejun Zhao. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.
- [32] Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to?. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 30–39.

- [33] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. 2020. NEO: A tool for taxonomy enrichment with new emerging occupations. In *The Semantic Web—ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II* 19. Springer, 568–584.
- [34] Nathan Godey, Eric Villemonte de La Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *EACL 2024-18th Conference of the European Chapter of the Association for Computational Linguistics*. 35–48.
- [35] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*. 2672–2680.
- [36] Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Franssen, and John Philip McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9099–9113.
- [37] Milan Gritta and Ignacio Iacobacci. 2021. XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 371–381.
- [38] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. arXiv:1802.05368. Retrieved from <https://arxiv.org/abs/1802.05368>
- [39] Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2022. Combining static and contextualised multilingual embeddings. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2316–2329.
- [40] Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual alignment—a survey. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 10922–10943. DOI: <https://doi.org/10.18653/v1/2024.findings-acl.649>
- [41] 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*. 2101–2112.
- [42] Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. 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*. 3633–3643.
- [43] 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 generalisation. In *International Conference on Machine Learning*. PMLR, 4411–4421.
- [44] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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)*. 2485–2494.
- [45] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987* (2024).
- [46] Alexandra Ils, Dan Liu, Daniela Grunow, and Steffen Eger. 2021. Changes in european solidarity before and during covid-19: Evidence from a large crowd-and expert-annotated twitter dataset. 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)*. 1623–1637.
- [47] Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. 2025. Harnessing the universal geometry of embeddings. arXiv:2405.07987. Retrieved from <https://arxiv.org/abs/2405.07987>
- [48] Fan Jiang, Tom Drummond, and Trevor Cohn. 2024. Pre-training cross-lingual open domain question answering with large-scale synthetic supervision. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 13906–13933.
- [49] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. arXiv:2004.13005. Retrieved from <https://arxiv.org/abs/2004.13005>
- [50] Staffs Keele et al. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).
- [51] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. 1459–1474.
- [52] Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*. Association for Computational Linguistics, 28–39.
- [53] Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.
- [54] Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American*

- Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 8051–8069.
- [55] Juntao Li, Chang Liu, Jian Wang, Lidong Bing, Hongsong Li, Xiaozhong Liu, Dongyan Zhao, and Rui Yan. 2020. Cross-lingual low-resource set-to-description retrieval for global e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8212–8219.
- [56] 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*. 12461–12476.
- [57] Xintong Li, Guanlin Li, Lemaio Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1293–1303.
- [58] 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*. 1663–1674.
- [59] Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 15979–15996. DOI : <https://doi.org/10.18653/v1/2025.acl-long.778>
- [60] Linlin Liu, Thien Hai Nguyen, Shafiq Joty, Lidong Bing, and Luo Si. 2022. Towards multi-sense cross-lingual alignment of contextual embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*. 4381–4396.
- [61] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2021. MEET-LM: A method for embeddings evaluation for taxonomic data in the labour market. *Computers in Industry* 124 (2021), 103341. <https://doi.org/10.1016/j.compind.2020.103341>
- [62] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Filippo Pallucchini. 2025. SeNSE: Embedding alignment via semantic anchors selection. *International Journal of Data Science and Analytics* 20, 1 (2025), 167–181.
- [63] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. 6294–6305.
- [64] Coşkun Mermer and Murat Saraçlar. 2011. Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 182–187.
- [65] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv e-prints* (2013), arXiv-1309.
- [66] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. 3111–3119.
- [67] David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2873–2878.
- [68] Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 555–565.
- [69] Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41, 2 (2009), 1–69.
- [70] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29, 1 (2003), 19–51.
- [71] 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*. 4990–4995.
- [72] Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics* 106 (2016), 125–146.
- [73] Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
- [74] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.
- [75] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2227–2237.
- [76] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv:1802.05365. Retrieved from <https://arxiv.org/abs/1802.05365>
- [77] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4996–5001.

- [78] Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outliers dimensions that disrupt transformers are driven by frequency. In *Findings of EMNLP 2022*. Association for Computational Linguistics.
- [79] Sara Rajae and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*. 1309–1316.
- [80] 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)*. Association for Computational Linguistics.
- [81] Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation?. In *Proceedings of the First NLP Workshop on Deep Learning for Natural Language Processing*. 29–36.
- [82] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research* 65 (2019), 569–631.
- [83] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *EMNLP 2020*. 1627–1643.
- [84] Dominik Schlechtweg, Anna Hätyy, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. arXiv:1906.02979. Retrieved from <https://arxiv.org/abs/1906.02979>
- [85] Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of NAACL-HLT*. 1599–1613.
- [86] Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8854–8861.
- [87] 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)*. 778–788.
- [88] Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 2375–2401. DOI : <https://doi.org/10.18653/v1/2025.acl-long.118>
- [89] 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*. 1477–1490.
- [90] 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)*. 8696–8706.
- [91] Matej Ulčar and Marko Robnik-Šikonja. 2022. Cross-lingual alignments of ELMo contextual embeddings. *Neural Computing and Applications* 34, 15 (2022), 13043–13061.
- [92] Oleg Vasilyev, Fumika Isono, and John Bohannon. 2024. Linear cross-lingual mapping of sentence embeddings. In *Findings of the Association for Computational Linguistics: ACL 2024*. 8163–8171.
- [93] Chao Wang, Hengshu Zhu, Peng Wang, Chen Zhu, Xi Zhang, Enhong Chen, and Hui Xiong. 2021. Personalized and explainable employee training course recommendations: A bayesian variational approach. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2021), 1–32.
- [94] Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. 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)*. 5721–5727.
- [95] Zirui Wang, Jiateng Xie, Ruo Chen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.
- [96] John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4602–4608.
- [97] Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. arXiv:1911.01464. Retrieved from <https://arxiv.org/abs/1911.01464>
- [98] Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 833–844.

- [99] 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)*. 4471–4482.
- [100] Haoran Xu and Philipp Koehn. 2021. Cross-lingual bert contextual embedding space mapping with isotropic and isometric conditions. arXiv:2107.09186. Retrieved from <https://arxiv.org/abs/2107.09186>
- [101] Jingwei Yi, Fangzhao Wu, Chuhan Wu, Xiaolong Huang, Binxing Jiao, Guangzhong Sun, and Xing Xie. 2022. Effective and efficient query-aware snippet extraction for web search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 3035–3046.
- [102] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1959–1970.
- [103] Zheng Zhang, Ruiqing Yin, Jun Zhu, and Pierre Zweigenbaum. 2019. Cross-lingual contextual word embeddings mapping with multi-sense words in mind. arXiv:1909.08681. Retrieved from <https://arxiv.org/abs/1909.08681>
- [104] 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*. 229–240.
- [105] Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 430–440.

Received 31 October 2023; revised 23 May 2025; accepted 11 August 2025