Polysemy Degrades Cross-Lingual Embedding Alignment in Multilingual Transformers

Anonymous Author(s)

Abstract

Multilingual transformer models such as MBERT and XLM-R learn shared representations across languages, but how polysemy affects cross-lingual embedding alignment remains poorly understood. We present a systematic study measuring how word sense ambiguity degrades cross-lingual similarity in these models across five typologically diverse language pairs. We find that translation equivalents have dramatically higher cosine similarity than random word pairs (Cohen's $d = 0.90$–3.01), confirming that meaning alignment emerges without explicit cross-lingual supervision. Critically, polysemy significantly weakens this alignment: monosemous translation pairs show 20–70% higher similarity than highly polysemous ones (Cohen's $d = 0.14$–0.57), with a consistent negative correlation between sense count and cross-lingual similarity (Spearman $\rho = -0.07$ to -0.37). However, when polysemous words appear in sense-disambiguating context, same-sense cross-lingual pairs recover substantially higher similarity than different-sense pairs (Cohen's $d = 1.0$–1.6), and this sense discrimination peaks at layer 10 in both models. Our results reveal polysemy as a systematic confound in cross-lingual evaluation and demonstrate that contextual embeddings from upper-middle layers can largely overcome this limitation.

1 Introduction

Multilingual transformer models serve billions of users across more than 100 languages, powering machine translation, cross-lingual search, and multilingual question answering. A central premise of these models is that words with similar meanings in different languages occupy similar regions of embedding space—even without explicit cross-lingual supervision [Wu et al., 2020, Pires et al., 2019]. But how robust is this alignment when words carry multiple meanings?

Consider the English word "bank," which has 18 senses in WORDNET [Miller, 1995]. Its French translation "banque" captures only the financial sense. When a multilingual model maps both words to a single embedding, the English representation must accommodate all 18 senses while the French one concentrates on just one or two. Intuitively, this mismatch should reduce their similarity—but by how much, and can context resolve it?

Prior work has established that multilingual models create cross-lingual representations [Wu et al., 2020, Pires et al., 2019, Conneau et al., 2020] and that polysemy hurts static embedding alignment [Zhang et al., 2019]. However, no systematic study has measured how sense count affects cross-lingual embedding similarity in modern contextual models across multiple language pairs, nor compared type-level and token-level similarity for polysemous words.

We address this gap with a large-scale empirical study across two models (MBERT and XLM-R), five language pairs (EN–FR, EN–DE, EN–ES, EN–RU, EN–ZH), and three complementary evaluation settings. We classify English words by polysemy level using WORDNET synset counts

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

and measure how cross-lingual similarity varies with sense count at both the type level (isolated words) and the token level (words in context).

Our main findings are:

- We show that translation equivalents have dramatically higher cosine similarity than random pairs (Cohen's d up to 3.01), confirming cross-lingual meaning alignment in both MBERT and XLM-R (section 4.1).
- We demonstrate that polysemy systematically degrades this alignment: monosemous pairs show 20–70% higher similarity than highly polysemous pairs, with a consistent negative correlation between sense count and similarity (section 4.2).
- We find that contextual embeddings recover sense-level cross-lingual alignment (Cohen's $d =$ 1.0–1.6 between same-sense and different-sense pairs), with layer 10 providing the best sense discrimination (section 4.3).
- We validate our approach against human similarity judgments on SEMEVAL-2017 Task 2, where MBERT achieves Spearman correlations of 0.36–0.49 after language-specific centering (section 4.4).

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes our methodology. Section 4 presents our experimental results. Section 5 discusses implications and limitations, and section 6 concludes.

2 Related Work

Cross-lingual word embeddings. The observation that word embedding spaces across languages share approximate structural similarity dates to Mikolov et al. [2013], who showed that a simple linear mapping can align monolingual word embeddings across languages. Subsequent work developed both supervised [Conneau et al., 2018] and unsupervised alignment methods, surveyed comprehensively by Ruder et al. [2019]. These approaches rely on the isomorphism assumption: that embedding spaces in different languages are approximately rotations of each other. Our work complements this line of research by measuring alignment quality as a function of polysemy, revealing a systematic source of noise in cross-lingual mappings.

Multilingual pretrained models. Modern multilingual transformers learn cross-lingual representations through shared parameters and multilingual pretraining. MBERT [Devlin et al., 2019] trains a single BERT model on 104 languages using Wikipedia, while XLM-R [Conneau et al., 2020] scales to 100 languages using CommonCrawl data. Pires et al. [2019] showed that MBERT transfers across languages even without shared vocabulary, and Wu et al. [2020] demonstrated that cross-lingual structure emerges from parameter sharing alone. Dufter and Schütze [2021] further found that shared position embeddings and special tokens suffice for cross-lingual alignment. We build on these findings by directly measuring embedding similarity for translation equivalents and testing how polysemy modulates this alignment.

Language neutrality and centering. Libovický et al. [2020] showed that MBERT representations contain a strong language-specific component that can be removed by subtracting the per-language mean embedding (centering). This simple technique improves cross-lingual similarity and is essential for fair comparison. We adopt centering throughout our experiments and quantify its impact on both type-level and token-level similarity.

Polysemy in cross-lingual embeddings. Polysemy poses a well-known challenge for cross-lingual alignment. Zhang et al. [2019] showed that polysemous words act as noise during supervised alignment of contextual embeddings and proposed filtering them during training. Upadhyay et al. [2017] introduced multi-sense multilingual embeddings, demonstrating that different languages can disambiguate senses that are conflated in any single translation pair. Scarlini et al. [2020] created sense embeddings from BERT and extended them multilingually via BabelNet. Unlike these approaches, we do not propose a new method for handling polysemy; instead, we systematically quantify how polysemy affects cross-lingual similarity in existing models, providing evidence for the magnitude of the problem.

Word-in-context disambiguation. The word-in-context (WiC) task [Raganato et al., 2020] tests whether two occurrences of a word share the same sense. Martelli et al. [2021] extended this to a cross-lingual setting with the MCL-WIC dataset, which we use to evaluate sense-level cross-lingual alignment. Our use of this dataset differs from standard WiC evaluation: rather than training a classifier, we directly compare embedding similarities for same-sense and different-sense pairs, revealing how well the raw representation space captures sense distinctions across languages.

Layer-wise analysis of transformers. Tenney et al. [2019] showed that different BERT layers encode different types of linguistic information, with semantic features concentrated in upper layers. Wu et al. [2020] found that early layers are more similar across languages in multilingual models. We extend this analysis to cross-lingual sense discrimination, showing that layer 10 (out of 12) provides the best sense-discriminative alignment—consistent with the view that upper-middle layers encode the richest semantic information.

3 Methodology

We design four experiments to test how polysemy affects cross-lingual embedding alignment. Section 3.1 describes the models, section 3.2 the datasets, section 3.3 the embedding extraction procedure, and section 3.4 the experimental protocol.

3.1 Models

We evaluate two widely used multilingual transformers:

MBERT (`bert-base-multilingual-cased`) [Devlin et al., 2019]: A 12-layer, 178M-parameter transformer trained on Wikipedia text from 104 languages with a shared WordPiece vocabulary of 110K tokens.

XLM-R (`xlm-roberta-base`) [Conneau et al., 2020]: A 12-layer, 278M-parameter transformer trained on CommonCrawl data from 100 languages with a SentencePiece vocabulary of 250K tokens. XLM-R uses more training data and a larger vocabulary than MBERT, and generally achieves stronger downstream performance.

Both models use subword tokenization, so multi-token words require aggregation (see section 3.3).

3.2 Datasets

MUSE Bilingual Dictionaries [Conneau et al., 2018] provide translation pairs for 110 language pairs. We use EN–FR, EN–DE, EN–ES, EN–RU, and EN–ZH, sampling up to 5,000 single-word, alphabetic pairs per language pair after deduplication. These pairs serve as the foundation for Experiments 1 and 2.

SEMEVAL-2017 Task 2 [Camacho-Collados et al., 2017] provides 500 manually curated word pairs per language with graded similarity scores (inter-annotator agreement $\approx$ 0.9). We use the cross-lingual subsets EN–DE, EN–ES, EN–IT, and EN–FA (914–978 pairs per pair after filtering). This dataset validates our similarity measurements against human judgments (Experiment 4).

MCL-WIC [Martelli et al., 2021] provides 1,000 sentence pairs per cross-lingual setting (EN–FR, EN–ZH, EN–RU), each labeled as same-sense (T) or different-sense (F) for a shared target word. The dataset is balanced with 500 examples per label. We use this for Experiment 3, which tests sense-level cross-lingual alignment.

Polysemy classification. We classify English words by the number of synsets in WORDNET [Miller, 1995]: *monosemous* (1 synset; 767–1,010 words per language pair), *polysemous* (2–4 synsets; 954–1,347 words), and *highly polysemous* (5+ synsets; 617–1,001 words).

3.3 Embedding Extraction

Subword aggregation. Both models tokenize words into subword units. We compute word-level embeddings by mean-pooling subword token representations, excluding special tokens (`[CLS]`, `[SEP]` for MBERT; `<s>`, `</s>` for XLM-R).

Language-specific centering. Following Libovický et al. [2020], we remove the language-specific bias by subtracting the per-language mean embedding:

$$\tilde{\boldsymbol{x}}_i^{(\ell)} = \boldsymbol{x}_i^{(\ell)} - \frac{1}{|\mathcal{D}_\ell|} \sum_{j \in \mathcal{D}_\ell} \boldsymbol{x}_j^{(\ell)}, \tag{1}$$

where $\boldsymbol{x}_i^{(\ell)}$ is the embedding of word i in language ℓ and $\mathcal{D}_\ell$ is the set of all words in language ℓ. This removes the language-identity component that inflates within-language similarity and deflates cross-lingual similarity.

Similarity metric. We use cosine similarity between centered embeddings:

$$\cos(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \frac{\tilde{\boldsymbol{x}} \cdot \tilde{\boldsymbol{y}}}{\|\tilde{\boldsymbol{x}}\| \cdot \|\tilde{\boldsymbol{y}}\|}. \tag{2}$$

For type-level experiments (Experiments 1, 2, 4), we feed isolated words to the model. For token-level experiments (Experiment 3), we feed full sentences and extract the contextualized embedding of the target word.

3.4 Experimental Protocol

Experiment 1: Translation pair similarity. For each language pair, we compute cosine similarity for all translation pairs and for shuffled (random) control pairs, then test whether translation similarity exceeds random similarity using the Mann–Whitney U test. We report Cohen's d as the effect size.

Experiment 2: Polysemy effect on similarity. We split translation pairs by WORDNET polysemy category and compare similarity distributions across categories using Mann–Whitney U tests. We also compute the Spearman rank correlation between sense count and cosine similarity.

Experiment 3: Contextualized sense-level similarity. Using MCL-WIC, we extract contextualized embeddings for target words in their sentence contexts and compare cosine similarity for same-sense (T) vs. different-sense (F) pairs. We analyze this across layers 1, 4, 7, 10, and 12 to identify which layers best discriminate senses cross-lingually.

Experiment 4: Correlation with human judgments. We compute cosine similarity for SEMEVAL-2017 word pairs and correlate with human gold scores using Spearman and Pearson correlations. This validates that our embedding similarities reflect human notions of semantic similarity.

Statistical testing. All comparisons use the Mann–Whitney U test (non-parametric, appropriate for cosine similarity distributions). We report Cohen's d for effect sizes and use $\alpha = 0.001$ as the significance threshold, which is conservative given our large sample sizes.

4 Results

4.1 Experiment 1: Translation Equivalents Have High Cross-Lingual Similarity

Table 1 reports cosine similarity for translation pairs vs. random pairs in both models (last layer, centered embeddings). Translation equivalents consistently show high similarity (0.22–0.77), while random pairs hover near zero. The effect sizes are massive (Cohen's d = 0.90–3.01), with all p-values below 10^{-300}.

The ranking of language pairs is consistent across models: EN–FR > EN–ES > EN–DE > EN–ZH > EN–RU. This ordering reflects typological and orthographic proximity—French and Spanish share Latin-derived vocabulary and script with English, while Russian uses Cyrillic, reducing subword overlap. Figure 1 visualizes these distributions.

4.2 Experiment 2: Polysemy Degrades Cross-Lingual Alignment

Table 2 reports similarity broken down by polysemy category. Across all language pairs and both models, monosemous words have consistently higher similarity than polysemous words. The effect

Model	Lang. Pair	Trans. Sim	Random Sim	Cohen's d
mBERT	EN–FR	0.765 ± 0.35	-0.000 ± 0.10	**3.01**
	EN–ES	0.697 ± 0.36	-0.001 ± 0.10	2.63
	EN–DE	0.591 ± 0.41	-0.002 ± 0.10	2.01
	EN–ZH	0.546 ± 0.40	-0.002 ± 0.10	1.91
	EN–RU	0.306 ± 0.25	0.002 ± 0.09	1.63
XLM-R	EN–FR	0.715 ± 0.42	-0.002 ± 0.18	2.24
	EN–ES	0.646 ± 0.43	0.003 ± 0.17	1.95
	EN–DE	0.547 ± 0.46	0.001 ± 0.17	1.58
	EN–ZH	0.480 ± 0.46	0.000 ± 0.18	1.38
	EN–RU	0.219 ± 0.30	-0.001 ± 0.17	0.90

Table 1: Cosine similarity of translation pairs vs. random pairs (last layer, centered). All differences are significant ($p < 10^{-300}$, Mann–Whitney U test). Best Cohen's d in **bold**.

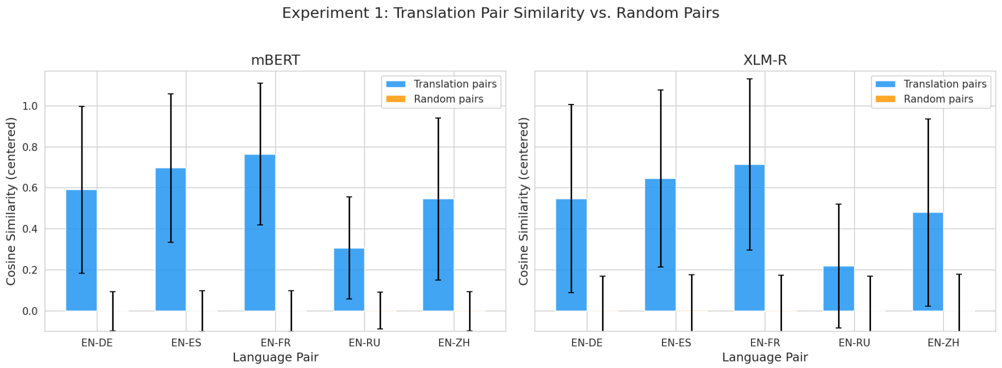


Figure 1: Cosine similarity distributions for translation pairs (colored) vs. random pairs (gray) across five language pairs. Translation pairs are clearly separated from random pairs in both models, with the gap largest for typologically similar languages.

is medium-to-large (Cohen's $d = 0.14$–0.57), and the Spearman correlation between sense count and similarity is negative and significant in every case.

Figure 2 shows that the polysemy effect is monotonic: as the number of senses increases, cross-lingual similarity decreases. This dose-response pattern is consistent across all language pairs, though the magnitude varies. The effect is weakest for EN–RU, where overall alignment is already low.

4.3 Experiment 3: Context Recovers Sense-Level Alignment

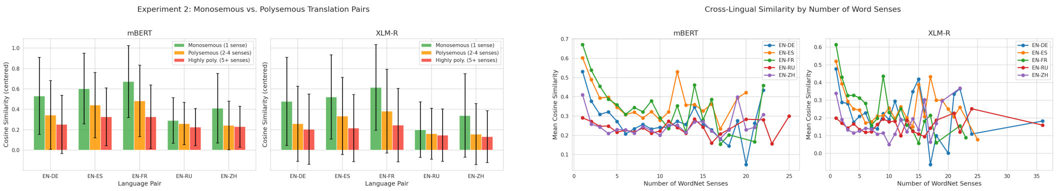
Table 3 compares cosine similarity for same-sense vs. different-sense pairs from MCL-WiC, using centered embeddings at layer 10 (the best-performing layer; see below). Same-sense pairs show substantially higher similarity in both models, with Cohen's d ranging from 1.18 to 1.58.

Centering is essential for XLM-R. Without centering, XLM-R produces near-saturated cosine similarities ($\sim$0.98) that mask sense discrimination (Cohen's $d = 0.48$–0.58). With centering, XLM-R achieves comparable effect sizes to mBERT (Cohen's $d = 1.18$–1.43). mBERT is less affected by centering because its representations already exhibit greater language separation.

Layer analysis. Figure 3 shows how sense discrimination varies across layers. Cohen's d increases monotonically from layer 1 ($d \approx 0.1$–0.2) to layer 10 ($d \approx 1.2$–1.6), then decreases slightly at layer 12. This pattern holds across all language pairs and both models, supporting the finding that upper-middle layers encode the richest semantic information [Tenney et al., 2019].

Model	Lang. Pair	Mono (1)	Poly (2–4)	High (5+)	Cohen's d	ρ
mBERT	EN–FR	**0.671**	0.482	0.327	0.54	-0.37^{***}
	EN–ES	**0.603**	0.441	0.327	0.49	-0.31^{***}
	EN–DE	**0.531**	0.344	0.252	0.52	-0.28^{***}
	EN–ZH	**0.410**	0.242	0.230	0.57	-0.16^{***}
	EN–RU	**0.291**	0.260	0.225	0.14	-0.12^{***}
XLM-R	EN–FR	**0.614**	0.381	0.244	0.56	-0.33^{***}
	EN–ES	**0.520**	0.335	0.217	0.47	-0.30^{***}
	EN–DE	**0.478**	0.258	0.205	0.55	-0.24^{***}
	EN–ZH	**0.339**	0.157	0.132	0.51	-0.15^{***}
	EN–RU	**0.200**	0.161	0.146	0.15	-0.07^{***}

Table 2: Mean cosine similarity by polysemy category (last layer, centered). Cohen's d compares monosemous vs. highly polysemous words. ρ: Spearman correlation between sense count and similarity (***: $p < 0.001$). Best similarity in **bold**.



(a) Similarity by polysemy category

(b) Similarity vs. sense count

Figure 2: *(Left)*Cosine similarity distributions by polysemy category. Monosemous words (1 sense) consistently show higher similarity than polysemous words. *(Right)*Negative correlation between WORDNET sense count and cross-lingual cosine similarity.

4.4 Experiment 4: Correlation with Human Judgments

Table 4 reports Spearman correlations between model-based cosine similarity and human gold scores on SEMEVAL-2017 Task 2. mBERT achieves moderate correlations (0.36–0.49) after centering, which provides a 10–13 percentage point boost over raw similarities. XLM-R performs lower on this task (0.12–0.25).

5 Discussion

5.1 Interpretation of Results

Why does polysemy degrade alignment? When a multilingual model represents a polysemous word like "bank" (18 WORDNET senses) as a single type-level embedding, it must distribute representational capacity across all senses. The French translation "banque" concentrates on just one or two senses. The resulting embeddings overlap only partially, reducing cosine similarity. This effect scales with the number of senses: words with 5+ senses show 20–70% lower similarity than monosemous words.

Why does context help? Contextual embeddings resolve the polysemy problem by specializing to the relevant sense. When "bank" appears in "I deposited money at the bank," its contextualized representation shifts toward the financial sense, better matching "banque" in a corresponding French context. Our Experiment 3 results (Cohen's $d = 1.0$–1.6) confirm that this mechanism works cross-lingually.

Why does mBERT outperform XLM-R on type-level similarity? This finding is initially surprising, since XLM-R consistently outperforms mBERT on downstream tasks [Conneau et al., 2020]. We hypothesize that XLM-R's larger vocabulary (250K vs. 110K tokens) and training on noisier CommonCrawl data (vs. mBERT's Wikipedia) produce more distributed representations.

Model	Lang. Pair	Same-Sense	Diff-Sense	Cohen's d
mBERT	EN–FR	0.324	0.154	1.23
	EN–ZH	0.291	0.114	**1.58**
	EN–RU	0.274	0.119	1.25
XLM-R	EN–FR	0.402	0.211	1.18
	EN–ZH	0.339	0.161	1.43
	EN–RU	0.352	0.175	1.31

Table 3: Same-sense vs. different-sense cosine similarity from MCL-WiC (layer 10, centered). All differences are significant ($p < 10^{-30}$). Best Cohen's d in **bold**.

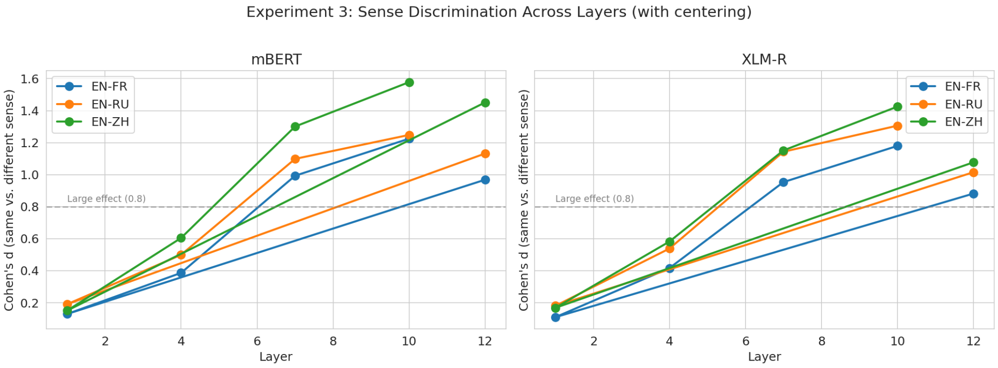


Figure 3: Cohen's d for same-sense vs. different-sense discrimination across transformer layers (centered embeddings). Sense discrimination increases monotonically from early to upper-middle layers, peaking at layer 10 in both models.

These representations are powerful for contextual tasks but yield weaker signal when queried with isolated words. Supporting this interpretation, the gap between models narrows substantially in the contextualized setting (Experiment 3), where XLM-R achieves comparable sense discrimination after centering.

The importance of centering. Centering provides a 10–13 percentage point boost in Spearman correlations on SemEval-2017 (table 4) and is essential for XLM-R in the MCL-WiC setting, where uncorrected similarities saturate near 0.98. This confirms that language-specific components dominate raw cosine similarity in multilingual models [Libovický et al., 2020], and any cross-lingual similarity measurement should control for this bias.

5.2 Implications

For practitioners. Applications that rely on cross-lingual word-level similarity—bilingual dictionary induction, cross-lingual information retrieval, word-level translation quality estimation—should account for polysemy. Using contextualized embeddings from layer 10 rather than the last layer or type-level embeddings provides better sense-discriminative alignment.

For evaluation. Polysemy is a systematic confound in cross-lingual evaluation benchmarks. When comparing models on bilingual lexicon induction or cross-lingual word similarity, controlling for polysemy level would provide a more nuanced picture of model quality. Our results suggest that performance differences between models may partly reflect differences in how they handle polysemous words.

For model development. The consistent negative correlation between sense count and cross-lingual similarity suggests room for improvement. Sense-aware training objectives, such as those incorporating word sense disambiguation during pretraining, could strengthen cross-lingual alignment for polysemous words.

Model	Lang. Pair	ρ (centered)	ρ (raw)	r (centered)
mBERT	EN–DE	**0.493**	0.376	0.498
	EN–ES	0.475	0.386	0.486
	EN–IT	0.455	0.364	0.462
	EN–FA	0.359	0.253	0.368
XLM-R	EN–DE	0.250	0.096	0.270
	EN–ES	0.203	0.082	0.208
	EN–IT	0.228	0.109	0.217
	EN–FA	0.119	0.047	0.109

Table 4: Correlation with human similarity judgments on SemEval-2017 Task 2 (last layer). ρ: Spearman; r: Pearson. Centering consistently improves correlations. Best ρ in **bold**.

5.3 Limitations

WordNet coverage. Our polysemy classification relies on English WordNet, which may not perfectly reflect a word's semantic complexity. Some words with one synset may have multiple pragmatic uses, while WordNet's fine-grained sense distinctions may overcount senses for others. Additionally, polysemy is classified only on the English side; the target-language word may itself be polysemous in ways we do not measure.

Type-level embeddings from contextual models. Feeding isolated words to models designed for sentences may produce suboptimal representations. While this is standard practice [Wu et al., 2020, Libovický et al., 2020], the models were never trained on isolated word inputs, and the resulting embeddings may not represent the models' full capabilities.

Language pair selection. We test only English-centric pairs. Non-English pairs (e.g., FR–DE, ZH–RU) may show different patterns, especially for polysemy effects that depend on sense overlap between the specific languages involved.

Model selection. We test only base-sized models (178M and 278M parameters). Larger models such as `xlm-roberta-large` (550M parameters) may show different polysemy effects, potentially with stronger alignment that is more robust to sense ambiguity.

6 Conclusion

We present a systematic study of how polysemy affects cross-lingual embedding alignment in multilingual transformers. Our experiments across two models, five language pairs, and three evaluation settings yield three main findings.

First, translation equivalents occupy similar regions of embedding space, with cosine similarities of 0.22–0.77 after centering (Cohen's d up to 3.01), confirming that cross-lingual meaning alignment emerges in both mBERT and XLM-R.

Second, polysemy systematically degrades this alignment. Monosemous translation pairs show 20–70% higher similarity than highly polysemous ones, with a consistent negative dose-response relationship between sense count and similarity across all language pairs.

Third, contextual embeddings largely overcome the polysemy problem. When polysemous words appear in sense-disambiguating context, same-sense cross-lingual pairs show substantially higher similarity than different-sense pairs (Cohen's $d = 1.0$–1.6), with layer 10 providing the best sense discrimination.

These findings have direct implications for cross-lingual NLP applications: practitioners should prefer contextualized embeddings from upper-middle layers when working with polysemous words, and evaluation benchmarks should control for polysemy to provide fairer model comparisons. Future work should extend this analysis to non-English language pairs, larger models, and sense-aware training objectives that could further strengthen cross-lingual alignment.

References

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, 2017.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, 2019.

Philipp Dufter and Hinrich Schütze. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2214–2231, 2021.

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

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, 2021.

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

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7193–7206, 2020.

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

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8758–8765, 2020.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019.

Shyam Upadhyay, Kai-Wei Chang, Matt Taddy, Adam Kalai, and James Zou. Beyond bilingual: Multi-sense word embeddings using multilingual context. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 101–110, 2017.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, 2020.

Mingsi Zhang, Yinglu Yin, Xiaodan Zhu, and Pierre Zweigenbaum. Cross-lingual contextual word embeddings mapping with multi-sense words in mind. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3554–3564, 2019.

A Additional Results

A.1 Experiment 1: Layer Analysis

Figure 4 shows how translation pair similarity varies across layers. In both models, similarity generally increases from early to later layers, with the last layer providing the highest type-level alignment. This contrasts with the sense-discrimination analysis in Experiment 3, where layer 10 outperforms layer 12, suggesting that the last layer optimizes for token prediction rather than sense-level semantics.

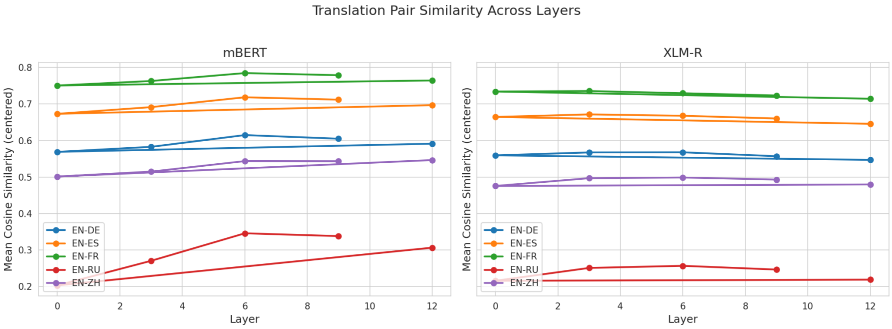


Figure 4: Translation pair cosine similarity across layers for both models. Similarity generally increases with layer depth, with the last layer providing the highest type-level alignment.

A.2 Experiment 3: Raw Similarity Without Centering

Table 5 reports MCL-WiC results without centering. mBERT shows moderate discrimination (Cohen's $d = 1.02$–1.51), while XLM-R produces near-saturated similarities ($\sim$0.98) with much weaker discrimination (Cohen's $d = 0.48$–0.58). This highlights the importance of centering, especially for XLM-R.

Model	Lang. Pair	Same-Sense	Diff-Sense	Cohen's d	Acc.
mBERT	EN–FR	0.444	0.342	1.02	0.719
	EN–ZH	0.416	0.303	1.51	0.786
	EN–RU	0.421	0.327	1.18	0.712
XLM-R	EN–FR	0.986	0.984	0.48	0.564
	EN–ZH	0.983	0.979	0.58	0.604
	EN–RU	0.986	0.983	0.52	0.565

Table 5: Same-sense vs. different-sense similarity *without centering* (last layer). XLM-R produces near-saturated similarities that mask meaningful sense discrimination.

A.3 Experiment 3: Similarity Across Layers

Figure 5 shows the raw cosine similarity values for same-sense and different-sense pairs across layers. The gap between the two conditions widens progressively from early to upper-middle layers.

A.4 Experiment 4: Visualization

Figure 6 visualizes the Spearman correlations from Experiment 4. mBERT consistently outperforms XLM-R on type-level cross-lingual word similarity, and centering improves both models.

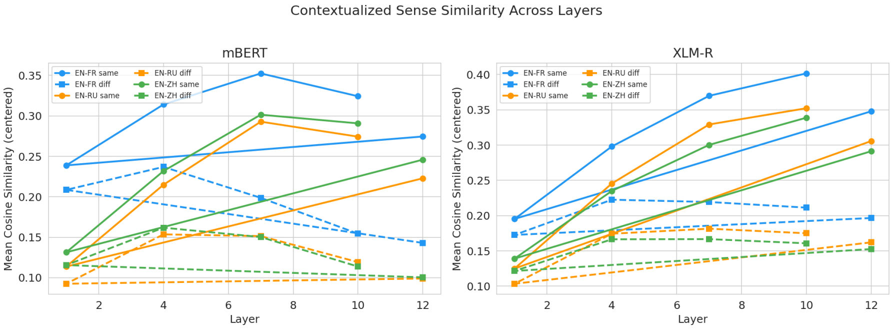


Figure 5: Same-sense and different-sense cosine similarity across layers (centered). The gap between conditions widens from early to upper-middle layers, with both converging slightly at the last layer.

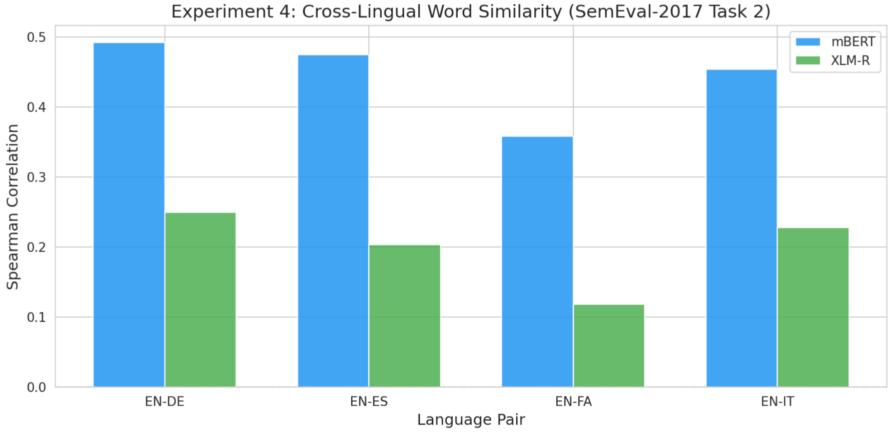


Figure 6: Spearman correlations with human similarity judgments on SEMEVAL-2017 Task 2. MBERT achieves higher correlations than XLM-R, and centering provides a consistent boost for both models.