Cross-Lingual Embedding Similarity Persists Across Languages but Weakly Reflects Polysemy

Anonymous Authors
Affiliation
email@example.com

1 Abstract

Multilingual encoders are widely used for cross-lingual retrieval and transfer, yet it remains unclear whether cosine similarity in these spaces meaningfully distinguishes shared senses when words are polysemous. We examine this question with XLM-R by comparing (i) static similarity for translation pairs from MUSE dictionaries and (ii) contextual similarity for sense-labeled cross-lingual pairs from MCL-WIC. Our analysis shows that translation pairs are much more similar than mismatched pairs (mean 0.513 vs. 0.375; Cohen's $d \approx 1.03$), confirming strong lexical alignment. For contextual embeddings, same-sense pairs are only slightly more similar than different-sense pairs (0.9847 vs. 0.9822; $d \approx 0.46$), despite highly significant tests. A simple threshold trained on development data yields 0.53 accuracy on cross-lingual MCL-WIC test sets, only marginally above chance. These results suggest that multilingual contextual embeddings preserve shared meaning at a coarse level but that raw cosine similarity is too blunt for robust sense discrimination without additional modeling. Our findings provide practical guidance for cross-lingual semantic search and lexicon induction: embedding similarity is reliable for translation-level alignment but only weakly sensitive to sense mismatches.

2 Introduction

Cross-lingual retrieval, bilingual lexicon induction, and multilingual search all rely on the assumption that multilingual embeddings place words with similar meanings close together. This assumption is often taken for granted because multilingual encoders perform well on downstream transfer tasks, yet it is less clear whether raw cosine similarity actually distinguishes shared senses when words are polysemous.

why does polysemy matter? Polysemy is the norm in natural language: many words express multiple senses that may or may not align across languages. If a single embedding averages over these senses, then cross-lingual similarity could remain high even when the intended meanings differ. This ambiguity matters for sense-level matching, dictionary induction, and semantic search where sense mismatches can degrade precision.

what is missing in existing work? Prior work shows strong cross-lingual alignment for static embeddings and multilingual contextual modelsConneau et al. [2018], Artetxe et al. [2018], Conneau et al. [2020]. However, the effect of polysemy on cross-lingual similarity has been less directly quantified under controlled sense-labeled conditions. We aim to fill this gap by comparing static translation similarity to contextual sense similarity within the same multilingual model.

Our approach contrasts static similarities from MUSE translation dictionaries with contextual similarities from MCL-WIC cross-lingual word-in-context pairs using XLM-R. We quantify how much same-sense (T) pairs differ from different-sense (F) pairs, and we report statistical tests and effect sizes (see figure 2 and table 2).

Quantitatively, translation pairs are much more similar than mismatches (mean 0.513 vs. 0.375; Cohen's $d \approx 1.03$), while same-sense contextual pairs are only slightly more similar than different-sense pairs (0.9847 vs. 0.9822; $d \approx 0.46$). A simple cosine threshold yields 0.53 accuracy on cross-lingual MCL-WIC test sets, only marginally above chance. These results suggest strong alignment at the translation level but weak sense discrimination with raw cosine similarity.

Our main contributions are:

- We conduct a controlled comparison of static and contextual cross-lingual similarity using MUSE and MCL-WIC within a single multilingual encoder.
- We quantify the impact of sense agreement on contextual similarity across four language pairs and report effect sizes and significance tests.
- We provide an empirical assessment of how far cosine similarity can go for sense-level matching and identify where it fails.

Paper organization. section 3 reviews prior work, section 4 describes datasets and extraction, section 5 presents results and analyses, and section 6 discusses implications and limitations.

3 Related Work

Static cross-lingual alignment. Early work on bilingual lexicon induction aligns monolingual spaces with supervised or unsupervised mappings. MUSE introduces adversarial initialization with Procrustes refinement and CSLS retrievalConneau et al. [2018], while VecMap proposes a robust self-learning framework for unsupervised mappingArtetxe et al. [2018]. These methods show strong translation retrieval but do not directly measure sense-level effects. Our static analysis uses MUSE dictionaries to establish a comparable alignment baseline within a multilingual encoder.

Contextual alignment across languages. Several studies align contextual representations with parallel data or context-aware mappingsSchuster et al. [2019], Aldarmaki and Diab [2019]. These works show that contextual embeddings can be aligned, but they do not focus on polysemy-driven similarity gaps. We instead analyze similarity directly in a multilingual model without additional alignment to isolate sense effects.

Multilingual pretraining. XLM and XLM-R demonstrate that multilingual masked language modeling yields strong cross-lingual transferLample and Conneau [2019], Conneau et al. [2020]. These models underpin many multilingual applications, making them a natural target for probing cross-lingual similarity and polysemy.

Sense-aware evaluation. WiC-style benchmarks test whether context pairs share the same sense. XL-WIC expands WiC to multiple languagesRaganato et al. [2020], and MCL-WIC defines multilingual and cross-lingual variants for SemEvalMartelli et al. [2021]. Recent work on multi-sense alignment explicitly targets polysemy in contextual embeddingsLiu et al. [2022]. Our study complements this line by providing a direct measurement of similarity shifts under sense agreement across languages.

4 Methodology

Problem formulation. We test two hypotheses: (i) translation pairs are more similar than mismatches in static embeddings, and (ii) same-sense cross-lingual contexts are more similar than different-sense contexts in contextual embeddings. For each pair, we compute cosine similarity between token or span representations and compare distributions across conditions.

Datasets. We use MCL-WIC cross-lingual word-in-context data for en–ar, en–fr, en–ru, and en–zh, each with 1000 test examples and gold T/F labels indicating whether the target words share a sense. We also use MUSE bilingual dictionaries (8–11k pairs per language) for static translation similarity. A polysemy lexicon is derived from the MCL-WIC en–en training split by selecting lemmas with both T and F labels.

Preprocessing and span handling. Each example provides target spans (possibly multiple ranges). We convert ranges to a single contiguous span by taking the minimum start and maximum end, then align wordpiece offsets to the span. For static similarities, we average XLM-R input embeddings

Pair Type	Count	Mean Cosine	Std
Translation (MUSE)	8000	**0.5131**	0.1646
Mismatch	8000	0.3747	0.0937

Table 1: Static similarity with XLM-R input embeddings for MUSE dictionaries. Translation pairs are substantially more similar than mismatches.

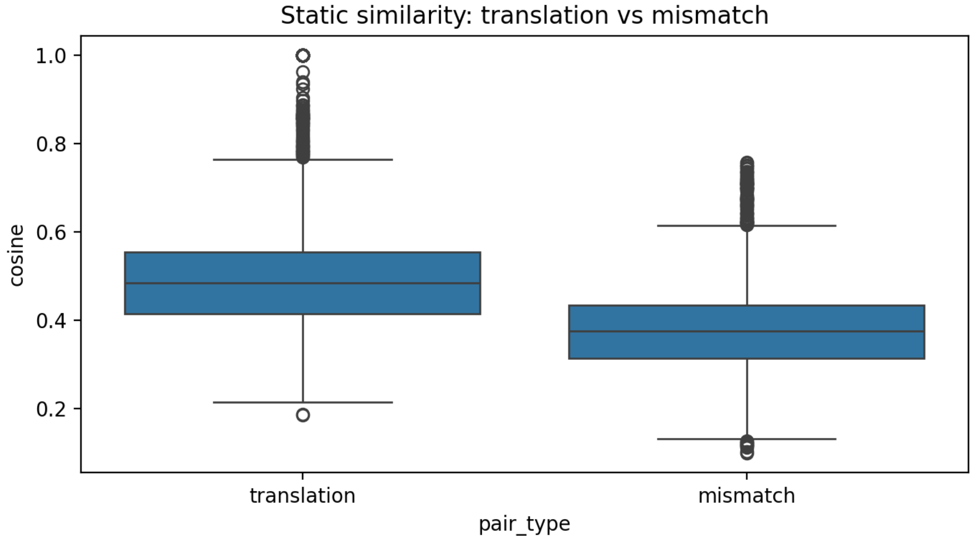

Figure 1: Static similarity distribution for MUSE translation pairs versus mismatches using XLM-R input embeddings. Translation pairs are consistently more similar.

across the wordpieces of each word. For contextual similarities, we average the last-layer hidden states across the target span.

Model and implementation. We use XLM-R (`xlm-roberta-base`) with max sequence length 128 and batch size 64. Inference is deterministic with seed 42 on an NVIDIA RTX 3090 (24GB). We do not fine-tune.

Evaluation metrics. We report cosine similarity distributions, Welch's t-test and Mann–Whitney U for T vs. F comparisons, and Cohen's d effect sizes. For an interpretable baseline, we fit a cosine threshold on the MCL-WiC en–en development set and report accuracy on cross-lingual test sets.

Baselines and comparisons. The static comparison (translation vs. mismatch) acts as a strong alignment baseline for MUSE. The contextual comparison (T vs. F) tests whether sense agreement is reflected in XLM-R representations without additional alignment.

5 Results

Static alignment baseline. table 1 shows a clear separation between translation and mismatched pairs in MUSE dictionaries. Translation pairs have much higher cosine similarity (0.5131 vs. 0.3747), corresponding to a large effect size (Cohen's $d \approx 1.03$). Figure 1 visualizes the separation across languages and confirms robust lexical alignment in XLM-R input embeddings.

Contextual similarity and sense agreement. table 2 reports cross-lingual MCL-WiC similarities. Same-sense (T) pairs are only slightly more similar than different-sense (F) pairs (0.9847 vs. 0.9822). Welch's t-test (t=14.57, p=7.56 $\times$ 10^{-47}) and Mann–Whitney U (U=2,541,088.5,

Label	Count	Mean Cosine	Std
Same sense (T)	2000	**0.9847**	0.0051
Different sense (F)	2000	0.9822	0.0059

Table 2: Contextual similarity on cross-lingual MCL-WiC test sets using XLM-R span embeddings. Same-sense pairs are slightly more similar.

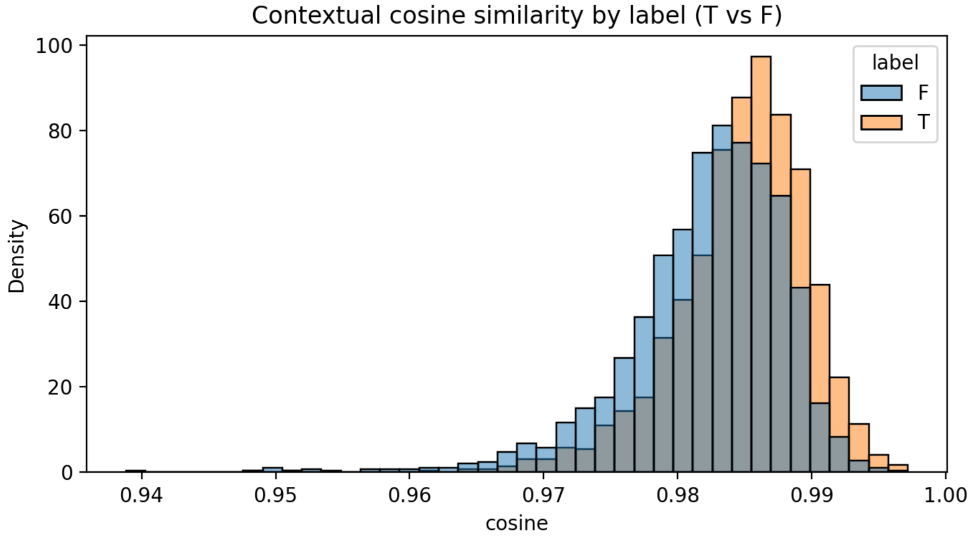

Figure 2: Histogram of contextual cosine similarities on cross-lingual MCL-WiC test sets. The T and F distributions largely overlap.

p=1.15×10^{-49}) are significant, but the effect size is modest ($d \approx 0.46$). Figure 2 and Figure 3 highlight the strong overlap between distributions.

Polysemy analysis. Using the training-derived polysemy lexicon, we observe minimal separation in static similarity and a small increase for polysemous words in contextual similarity (Figure 4). This suggests that polysemy does not reduce similarity in this setting and may even slightly increase similarity due to context averaging effects.

Thresholded WiC accuracy. A cosine threshold tuned on the MCL-WiC en–en development set achieves 0.53 accuracy on cross-lingual test sets, only slightly above chance. This confirms that raw cosine similarity alone is not sufficient for reliable sense discrimination.

6 Discussion

Interpretation. The static baseline shows that XLM-R input embeddings strongly align translation pairs, consistent with prior alignment results. In contrast, contextual similarities for MCL-WiC pairs are compressed near 1.0, which blurs the gap between same-sense and different-sense contexts. This suggests that span-averaged contextual representations are dominated by shared context and high-level semantics rather than fine-grained sense distinctions.

Why cosine similarity saturates. The extremely high cosine values for both T and F pairs indicate that raw similarity in XLM-R's last-layer space may be too coarse. The span averaging step further reduces variance, especially when the target span includes multiple wordpieces or when the sentence context dominates the representation.

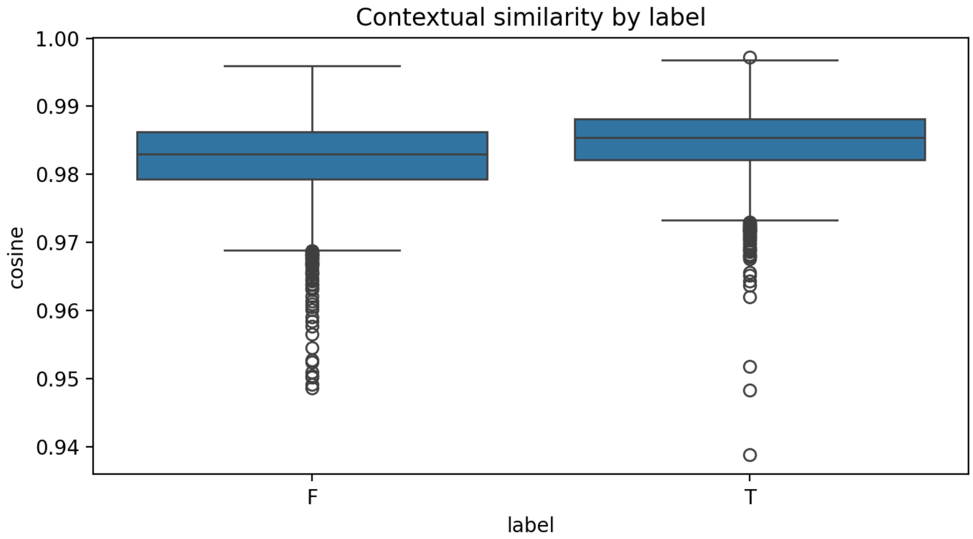

Figure 3: Box plot of contextual similarities for same-sense (T) and different-sense (F) pairs. The median difference is small despite statistical significance.

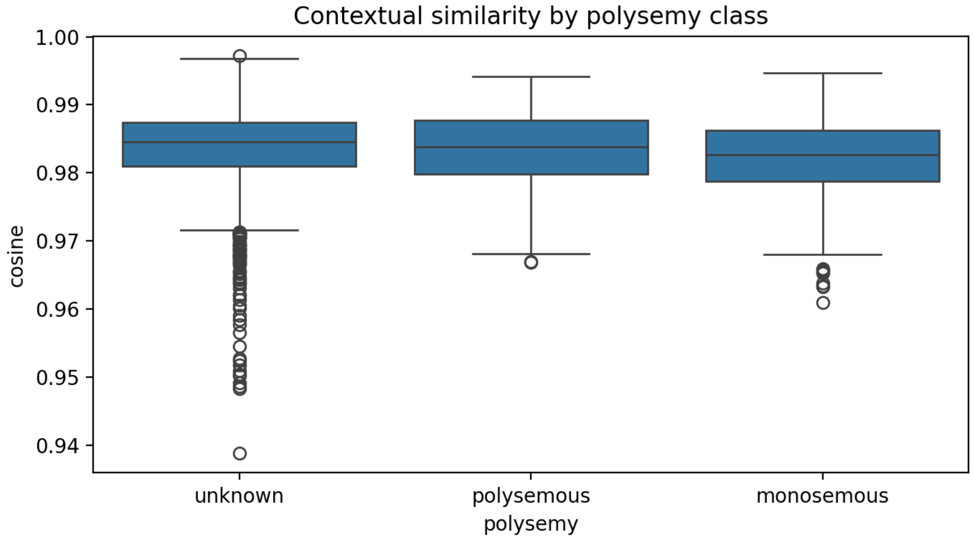

Figure 4: Contextual similarity by polysemy group, based on whether lemmas exhibit both T and F labels in training. Differences are small.

Limitations. First, our static baseline uses XLM-R input embeddings rather than standalone monolingual vectors (e.g., fastText), which may inflate alignment. Second, polysemy is approximated by training labels and may be incomplete. Third, we merge multi-span targets into a single span, which can blur multiword expressions. Fourth, we evaluate only one multilingual model without fine-tuning or alignment-specific training.

Implications. For cross-lingual retrieval, cosine similarity in multilingual encoders is reliable for translation-level alignment but not sufficient for sense-level discrimination. Sense-aware objectives, contrastive probing, or alignment fine-tuning may be necessary when precise meaning matching is required.

7 Conclusion

We tested whether multilingual embeddings align shared meaning across languages and whether polysemy reduces cross-lingual similarity. Using MUSE dictionaries and MCL-WIC sense-labeled pairs with XLM-R, we found strong static alignment for translation pairs but only a small contextual gap between same-sense and different-sense contexts. The main takeaway is that cosine similarity in multilingual encoders captures broad semantic alignment but weakly reflects sense distinctions.

Future work should replace the static baseline with aligned fastText or MUSE vectors to isolate model effects, evaluate additional benchmarks such as XL-WIC, and test contrastive or sense-aware probes that may recover finer-grained sense differences.

References

Hanan Aldarmaki and Mona Diab. Context-aware cross-lingual mapping. *arXiv preprint arXiv:1903.03243*, 2019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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*, 2018.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *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*, 2020.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, 2019.

Hongwei Liu, Ying Zhang, Yang Liu, et al. Towards multi-sense cross-lingual alignment of contextual embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2022.

Federico Martelli, Roberto Navigli, et al. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation*, 2021.

Alessandro Raganato, Yves Scherrer, and J"org Tiedemann. XL-WiC: A multilingual benchmark for word-in-context disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Sebastian Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of contextual word embeddings. *arXiv preprint arXiv:1902.09492*, 2019.