LANE: Lexical Adversarial Negative Examples for Word Sense Disambiguation

Anonymous ACL submission

Abstract

Fine-grained word meaning resolution remains a critical challenge for neural language models (NLMs) as they often overfit to global sentence representations, failing to capture local semantic details. We propose a novel adversarial training strategy, called LANE, to address this limitation by deliberately shifting the model's learning focus to the target word. This method generates challenging negative training examples through the selective marking of alternate words in the training set. The goal is to force the model to create a greater separability between same sentences with different marked words. Experimental results on lexical semantic change detection and word sense disambiguation benchmarks demonstrate that our approach yields more discriminative word representations, improving performance over standard contrastive learning baselines. We further provide qualitative analyses showing that the proposed negatives lead to representations that better capture subtle meaning differences even in challenging environments. Our method is model-agnostic and can be integrated into existing representation learning frameworks.

1 Introduction

Word Sense Disambiguation (WSD), the task of identifying the precise meaning of a word in context, remains central to deep semantic understanding. It underpins applications such as Lexical Semantic Change Detection (LSCD) (de Sá et al., 2024) and word similarity tasks (Armendariz et al., 2020). For example, WSD systems aim to determine whether the word "crazy" denotes insanity (as in *That's a crazy man*) or excitement (as in *That's crazy, man*), and whether "bank" refers to a financial institution or a river edge. While these words can appear in similar contexts, its particular usage can modulate completely different meanings.

Modern neural language models like XLM-Roberta achieve impressive results in this task, but their final word representation often overfit to global context rather than encoding a word's specific sense (Liu et al., 2021). They solve the task by capturing topical cues, for instance, inferring that "bank," "loan," and "interest" signal finance, but can misrepresent a word's fine-grained contribution (Xu et al., 2025) (McCoy et al., 2019). This limitation surfaces in cases like heavy rain vs. heavy traffic: both imply "a lot," yet with different nuances, intensity/volume versus density/severity. Such failures to capture precise contextual meaning hinder performance in downstream tasks requiring genuine semantic nuance, like neologism identification (McCrae, 2019).

A widely adopted strategy to improve lexical sensitivity in WSD is target-word highlighting, where the word of interest is marked (e.g., with special tokens) before being encoded by a language model (Cassotti et al., 2023). The underlying assumption is that explicit marking encourages the model to attend more directly to the lexical semantics of the word. Yet, as our analysis reveals (Figure 1), this assumption is fragile: embeddings of the same sentence remain nearly identical regardless of which word is highlighted, indicating that the model often resolves the task using only the surrounding context (Liu et al., 2021).

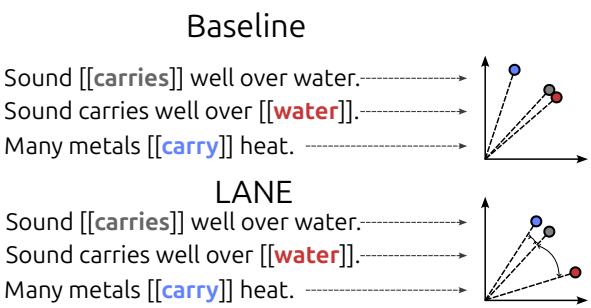

Figure 1: Comparison of traditional (Baseline) and Lexical Adversarial Negative Example (LANE) learned representations. Colors represent anchor (gray), positive (blue), and adversarial negative (red).

In practice, this results in nearly indistinguishable representations for semantically distinct words, while instances of the same sense across different contexts may be mapped to more distant vectors. Such behavior illustrates the persistence of shortcut learning (Robinson et al., 2021), where models exploit superficial contextual regularities rather than grounding their predictions in the lexical meaning of the marked token.

This imbalance produces a collapsed representation space, where embeddings concentrate in a narrow region irrespective of the highlighted word. The consequence is reduced lexical separability and weak generalization on tasks requiring sensitivity to word-level sense, especially in contexts that are semantically similar. Taking this into consideration, in this paper, we investigate *How new approaches can ensure that the representation space meaningfully reflects the contribution of the target word?*

To address this, we introduce the LANE framework, which biases representations toward lexical relevance. LANE generates adversarial negatives by substituting the highlighted word while keeping the sentence context unchanged. These hard negatives share the full contextual frame of the anchor sentence but differ in the target token, forcing the model to focus on the word's semantics rather than relying solely on surrounding context.

By presenting nearly identical sentences that differ only in the marked word, LANE sharpens semantic boundaries and produces more discriminative representations. This encourages the model to attend precisely to the lexical contribution of the target word, resulting in embeddings that better capture fine-grained word meaning.

We evaluate LANE across both English and multilingual datasets, spanning a diverse set of architectures. Results show consistent improvements in lexical representation quality and cross-domain generalization, demonstrating that targeted lexical supervision can enhance robustness even in high-resource, context-rich settings. Furthermore, the method's agnostic nature allows it to be incorporated into existing training pipelines with minimal computational overhead or implementation complexity, making it broadly applicable for multilingual lexical tasks.

2 Related Work

Representing fine-grained word meaning has long been a longlasting challenge in natural language processing. Early approaches tackled this problem through lexical resources such as WordNet (Miller, 1995) and the creation of sense-annotated corpora to support supervised WSD methods (Raganato et al., 2020) (Bevilacqua et al., 2021) (Huang et al., 2024a). With the advent of deep learning, contextualized word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) became standard, enabling substantial gains in WSD and related tasks by dynamically adjusting representations according to context.

Yet despite these advances, contextualized models often fail to distinguish subtle word-level meaning differences, tending instead to overfit to sentence-level semantics (Ethayarajh, 2019). This weakness is particularly visible in LSCD, which requires tracking the semantic shifts of individual words across time and domains. While static embedding approaches offered initial baselines (Schlechtweg et al., 2019) (Martinc et al., 2020), more recent work has leveraged contextualized embeddings to capture dynamic variation (Schlechtweg et al., 2020) (Kutuzov and Giulianelli, 2020) (Giulianelli et al., 2020). Nevertheless, isolating the semantics of the target word from the broader discourse remains an open challenge.

To address these limitations, recent research has turned to contrastive learning as a way to induce more discriminative semantic representations. Inspired by advances in computer vision (Chen et al., 2020), contrastive objectives have been adapted to NLP, yielding stronger sentence and word embeddings for tasks such as semantic similarity and clustering (Gao et al., 2021).

Within lexical semantics, several directions have emerged. One line of work leverages multilingual pretraining on large-scale, diverse corpora to improve contextual coverage of word senses (Cassotti et al., 2023) (Yadav and Schlechtweg, 2025). Another incorporates auxiliary supervision, such as in-context sense induction, to encourage sense separation (Giulianelli et al., 2023) (Mosolova et al., 2024) (Liétard et al., 2024) (Li'etard and Loiseau, 2025). A third explores data augmentation, for example by altering input structure: Martelli et al. proposed swapping sentence order in a cross-encoder (Martelli et al., 2021), though this strategy does not extend to bi-encoder architectures where contrastive losses are typically computed.

Complementary to these efforts, adversarial strategies aim to construct more challenging negatives to prevent representational collapse. Prior

work has explored lightweight perturbations such as case alternation (Wang et al., 2022), synonym and antonym substitution (Wang et al., 2021), and token replacements with masked language model predictions (Chuang et al., 2022). Multilingual adversarial signals have further been obtained through cross-lingual links in Wikipedia, which align English sentences with low-resource counterparts while introducing cross-lingual semantic contrasts (Liu et al., 2021). While these methods increase robustness and sharpen decision boundaries, they typically operate at the sentence level, focusing on global semantic differences rather than directly enforcing separability of meanings for a specific word in context.

Our work addresses this gap by combining the strengths of contrastive learning and adversarial augmentation while targeting their shared limitation: the absence of adversarial signals that operate at the word level. Instead of relying on heuristic perturbations or external lexical resources, we generate adversarial negatives by marking different words within the same sentence, producing pairs that are contextually identical yet lexically distinct. This strategy is model-agnostic and integrates seamlessly into existing frameworks for representation learning.

In doing so, it complements multilingual pretraining and auxiliary-task-based sense induction, while providing adversarial challenges directly tied to the phenomenon of interest: the fine-grained disambiguation of word meaning. Our experiments on both WSD and LSCD detection confirm that this approach yields more discriminative word representations than standard contrastive and adversarial baselines, underscoring the importance of aligning adversarial objectives with lexical semantics.

3 Datasets

Evaluating language models on a single dataset has long been standard practice in Natural Language Processing (NLP), yet this evaluations often provide an incomplete view of generalization and robustness (Lones, 2024). Even within a single task, datasets may differ substantially in domain, linguistic complexity, annotation conventions, and underlying assumptions. Consenquently, relying on one benchmark risks overfitting to dataset-specific idiosyncrasies rather than assessing a model's true capacity for semantic discrimination in WSD.

To obtain a more comprehensive evaluation, we curated four datasets covering diverse sources and temporal spans: SEMCOR (Miller et al., 1993), MASC[1], FEWS (Blevins et al., 2021), and DWUG (Schlechtweg et al., 2024). These datasets differ primarily along two dimensions: (i) time of creation (i.e. 1993 vs. 2020) and (ii) source type (i.e. dictionaries, books, blogs). This diversity enables controlled variation in language register and data quality (ranging from the standardized style of dictionaries to the more informal, conversational tone of blogs) thus exposing models to a broad spectrum of lexical and contextual phenomena.

Additional to the created datasets we evaluate our models in WiC data (Pilehvar and Camacho-Collados, 2019) and DWUG (Schlechtweg et al., 2021). While WiC and DWUG are formatted for WSD formulations, SEMCOR, MASC, and FEWS required preprocessing as they just present usages and sense keys. We generated contrastive pairs by merging instances with identical lemma and POS tags, then pairing them across distinct context. This process yielded pairs of the same lexical item in different environments, supporting fine-grained semantic discrimination. Ground-truth labels were automatically assigned using sense keys: pairs sharing the same key were labeled as positive (1), while those differing in sense were labeled as negative (0). This alignment to sense inventories ensures that the resulting pairs genuine semantic distinctions rather than superficial contextual variation.

For SEMCOR, MASC, and FEWS, we partitioned data lexicographically such that test sets contain words beginning with letters "P" or later. This reduces lexical overlap between training and testing splits, forcing models to generalize to unseen lexical items. Unlike DWUG and WiC, our resources include adjectives and adverbs in addition to nouns and verbs, thereby broadening the coverage of syntactic categories and semantic phenomena.

For multilingual evaluation, we adopt XL-WiC (Raganato et al., 2020) which extends WiC's word in context formulation to multiple languages (see Table 4). XL-WiC draws on both WordNet and Wiktionary to maintain cross-lingual consistency. Additionally, we train with XL-LEXEME, a composite resource integrating XL-WiC, MCL-WiC, and AM2iCO, to further enhance cross-lingual and cross-categorical representativeness.

The Word-in-Context task requires models to

[1] https://anc.org/data/masc/

Table 1: Examples of sentences pairs in the WiC dataset and corresponding hard_negative examples generated by LANE

Dataset	Label	Sentence 1	Sentence 2
SEMCOR (1993)	positive	The heavens opened, pelting them with *hail* the size of walnuts.	Drought, *hail*, disease, and insects take their toll of crops.
MASC (2003)	negative	The report wasn't hard, I had already read the book, and so I just jotted down a few *notes* for him.	Passing *notes* when the teacher isn't paying attention.
FEWS (2021)	positive	Virchow stated that premature fusion of this suture results in *pachycephalic* deformity.	[...] that a judge may be learned in the law, but woefully *pachycephalic* in matters scientific.

distinguish whether two target word usages convey the same meaning. A common learning paradigm for this task is contrastive learning (Cassotti et al., 2023), where a sentence (x) serves as an anchor, paired with positive (x, x^+) or negative (x, x^-) examples. In WSD, the anchor contains a target word appearing in two contexts: in positive pairs, the meanings coincide; in negative pairs, they differ. When encoding these pairs with Transformer models, target words are marked explicitly, either via prefix notation (e.g., `word<s>context`) or inline markers (e.g., `<t>word</t>` in the sentence). However, such marking schemes can cause the model to rely on sentence-level cues rather than the target word itself (Liu et al., 2021). To address this, our mehtod (LANE) generates hard-negative examples by varying the marked token within identical sentence contexts, thereby enforcing a stronger focus on the target word (see Table 2).

Table 2: Examples of positive, negative, and hard negative pairs in WiC. Hard negatives are generated by LANE.

Label	Sentence 1	Sentence 2
positive	Sound *carries* well over water.	Many metals *carry* heat.
negative	Sound *carries* well over water.	You must *carry* your camping gear.
hard negative	Sound *carries* well over water.	Sound carries well over *water*.

4 Methodology

Ideally, an embedding space for word senses should map all occurrences of the same sense to a single, consistent vector regardless of context. In practice, this entails positioning anchor–positive pairs (same sense, different contexts) close together in the embedding space, while pushing anchor–negative pairs (different senses) apart, a principle often formalized through the ball-packing problem (Robinson et al., 2020).

We propose a revised objective that goes beyond distinguishing positive and negative pairs by enforcing self-differentiation. In this setting, the surrounding sentence remains identical, but the marked target word differs. To operationalize this, we introduce adversarial negative examples generated through a rule-based procedure that randomly replaces the highlighted word in the sentence with another lexical candidate drawn from the training data (Algorithm 1). These adversarial negatives prevent the model from relying solely on contextual cues and instead compel it to attend to the lexical identity of the target word. During training, adversarial negatives are integrated alongside standard positive and negative pairs, enhancing the model's ability to learn fine-grained word sense representations.

In the following section, we describe how these adversarial negatives are generated dynamically and progressively replace a small proportion of the training data over epochs.

4.1 Adversarial Negative Examples

Bi-encoders independently represent vectors in the embedding space without explicit awareness of which target word the model should attend to. Although marking tokens are intended to signal the target, they have minimal effect in practice (Section 5). To improve word-sense learning, we aim to reduce the model's reliance on contextual cues during training. To this end, we generate adversarial negative examples with two desirable properties:

- **Property 1:** A contextualized word representation should produce a distinct vector whenever the marked lexical item changes, even if the surrounding context remains identical.

- **Property 2:** For words X in sentence A and Y in sentence B that share the same sense, their embeddings should remain similar only when X and Y are the marked items; the embeddings should diverge whenever a different lexical item is marked in A or B.

The vector representation of a word should change whenever a different lexical item in the same sentence is marked as the target (Property 1), following the assumption of avoidance of repetition (Walter et al., 2007). When two words share the same sense across different sentences, their embeddings should diverge if another word is marked (Property 2).

Building on these properties, we generate adversarial examples that explicitly encode this learning constraint. During training of each dataset, we construct: (i) pairs of identical sentences that differ only in which word is marked as the target; and (ii) pairs in which sentences A and B contain the same sense for a given word, but a different word in A is marked. This construction ensures that even identical or highly similar contexts yield distinct representations when the lexical target changes.

In Algorithm 1, we illustrate our method for dynamically generating negative examples in-batch. The input consists of a dictionary containing two target words (word1, word2), two corresponding sentences (sentence1, sentence2), and a similarity label (label):

- First, the function extracts all tokenized words from 'sentence1' and filters out any occurrences of 'word1'.

- From the remaining tokens, one word is randomly selected as newword.

- If the original 'label = 0.0' (indicating dissimilarity between the pair), the output replaces 'word2' with 'newword' while making both sentences identical to 'sentence1'.

- Otherwise ('label $\neq$ 0.0'), the function replaces 'word1' with 'newword' while keeping 'sentence2' unchanged.

In both cases, the newly generated example is assigned a label of '0.0', ensuring that it serves as a challenging negative instance. Because the contexts remain nearly identical, the model is forced to focus on the marked lexical item rather than relying on the surrounding context. We note that the likelihood of 'newword' sharing the original meaning is low, consistent with the *avoidance of repetition* assumption.

Input: $w_1, s_1, w_2, s_2, label$
Output: $w_1, s_1, w_2, s_2, label$
$words \leftarrow$ split-into-words(s_1);
$candidates \leftarrow \{\, w \mid w \in words,\ w \neq w_1 \,\}$;
$newword \leftarrow$ random-choice$(candidates)$;
if $label = 0$ **then**
  $w_2 \leftarrow newword$;
  $s_2 \leftarrow s_1$;
end
else
  $w_1 \leftarrow newword$;
  $label \leftarrow 0$;
end
return $w_1, s_1, w_2, s_2, label$;

Algorithm 1: Pseudo-code for creating adversarial negative examples (`lexical-negative`)

We apply a multilingual word splitter to accommodate languages such as Farsi and Japanese, where whitespace is not a reliable delimiter. Finally, to prevent the model from overfitting to these challenging negatives or collapsing into a local optimum, we introduce the adversarial examples gradually through a scheduled insertion strategy.

4.2 Scheduler

Introducing adversarial negatives too early in training risks convergence to suboptimal local minima (Section 8). In such cases, the model may overfit to a narrow subset of challenging comparisons before developing robust and generalizable representations (Xuan et al., 2020).

To mitigate this issue, we employ a linear scheduling strategy, scaled by the training epoch, to gradually introduce adversarial negatives. During the initial warm-up phase, the contrastive language model is trained without any adversarial examples, allowing it to form stable base representations. After this stage, the probability of sampling adversarial examples increases linearly with each epoch, ensuring a smooth and controlled transition from easy to hard comparisons. This progressive introduction balances representational stability with in-

creasing task difficulty. All negatives are computed dynamically from in-batch data rather than precomputed, enabling the model to adapt continuously as training evolves.

Our ablation studies corroborate the effectiveness of this strategy (Section 8): the linear scheduler prevents early collapse, enhances training stability, and achieves superior overall performance compared to settings where adversarial examples are introduced prematurely or at a fixed rate.

4.3 Optimization Objective

To differentiate senses, we optimize our models with a contrastive learning objective. Following recent literature in lexical differentiation, we employ an in-batch, cosine–based loss (Yadav and Schlechtweg, 2025). This formulation leverages implicit negatives within each batch, thereby promoting semantically coherent clustering via cosine similarity. Our choice aligns with prior state-of-the-art methods and provides a strong, empirically validated baseline. All models are optimized using the CoSENT loss (Huang et al., 2024b), as formalized in the equation below:

$$\mathcal{L} = \log\left(1 + \sum_{i,j \;\; y_i < y_j} e^{\lambda(s_i - s_j)}\right) \qquad (1)$$

Here, s_k denotes the cosine similarity score for the k-th embedding pairs such that the expected similarity of s_i is greater than s_j. The summation extends over all ordered pairs in the batch such that $y_i < y_j$. If the embeddings are u_k and v_k, then $s_k = \cos(u_k, v_k) = \frac{u_k \dot{v}_k}{|u_k||v_k|}$. y_k represents the ground-truth similarity label for the k-th pair, and λ is a trainable scaling factor that controls the sharpness of the distribution.

If the model ranks correctly ($s_j > s_i$), the difference $(s_i - s_j)$ is negative, and the exponential term $e^{\lambda(s_i - s_j)}$ becomes small, contributing minimally to the overall loss. Conversely, when the model ranks incorrectly ($s_i > s_j$), the difference $(s_i - s_j)$ is positive, causing $e^{\lambda(s_i - s_j)}$ to grow large and yield a higher loss value. This formulation thus penalizes misranked pairs more heavily while suppressing contributions from correctly ranked ones. In all experiments, we use the standard scaling parameter $\lambda = 20$.

5 Evaluation

We assess the effectiveness of the proposed LANE method on the datasets described in Section 3. Each dataset is split into train, development, and test partitions. During training, negative mining is applied exclusively to encourage more discriminative word representation, while the development split is used for model selection, and the test split is reserved for final evaluation.

To quantify the added value of LANE, we conduct controlled comparisons using two modern transformer architectures as backbone models: DeBERTa-v3 (DV3) and ModernBERT (MBERT), and the established architecture XLM-RoBERTa (XLM-R). For each architecture, we evaluate both the baseline model trained with standard contrastive objectives and the same model augmented with LANE. This design isolates the impact of LANE on learning fine-grained word-sense distinctions, independent of the underlying architecture.

All models are trained using AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 1e-5, 500 warm-up steps, a weight decay of 0.01, and an effective batch size of 64. Models with LANE are trained for 20 epochs, while baseline models without LANE are trained for 10 epochs, given its early convergence, with model selection based on performance on the development split. This evaluation framework ensures a fair and direct comparison, allowing us to highlight improvements in word-sense representation and downstream performance attributable specifically to LANE.

Table 3: Comparison of methods across datasets in terms of accuracy for test data.

Model	WiC	DWUG	SEMCOR	MASC	FEWS
MBERT	0.567	**0.735**	0.722	0.733	0.509
MBERT+**LANE**	**0.589**	0.734	**0.756**	**0.759**	**0.523**
DV3	0.655	0.735	**0.759**	0.754	0.629
DV3+**LANE**	**0.660**	**0.737**	0.756	0.754	0.629
XLM-R	0.705	0.739	**0.759**	0.752	0.627
XLM-R+**LANE**	**0.721**	**0.742**	**0.759**	**0.773**	**0.647**

As shown in Table 3, incorporating LANE adversarial negatives consistently improves, or at minimum maintains, accuracy across all evaluated datasets. For instance, MBERT sees notable gains on WiC, SEMCOR, MASC and FEWS when combined with LANE, highlighting its effectiveness in enhancing word-sense discriminability. Similarly, DeBERTa-v3 exhibits improvements with LANE, particularly on the WiC and DWUG datasets, demonstrating that even strong monolin-

gual baselines benefit from the adversarial word-focused training signal.

Interestingly, XLM-RoBERTa, despite being older than DeBERTa-v3 and MBERT, achieves the highest baseline accuracy across several datasets, reflecting its strong pretraining on large-scale multilingual data. When augmented with LANE, XLM-R not only improves further, most noticeably on WiC, MASC, and FEWS, but also attains the overall highest scores across the board, confirming the robustness of LANE across architectures.

These results indicate that LANE consistently enhances word-level representations, improving performance on diverse datasets regardless of the underlying model architecture. While DeBERTa-v3 achieves performance comparable to XLM-R on monolingual datasets, its monolingual design restricts its applicability in multilingual settings, as further illustrated in Table 4. Overall, the evaluation demonstrates that LANE provides a substantial and architecture-agnostic boost to contrastive learning of fine-grained word senses.

Building on the monolingual experiments, we further evaluate our approach in a multilingual word sense disambiguation setting using the XL-LEXEME datasets (XL-WiC (Raganato et al., 2020), MCL-WiC (Martelli et al., 2021), AM2ICO (Liu et al., 2021)).

We denote our model as XL-WiC+LANE (Discriminative Enhanced Lexical Training with Adversarial negatives) and have released it on HuggingFace[2]. Table 4 reports our results on the XL-WiC test set, comparing XL-WiC+LANE with the base XL-LEXEME model and prior state-of-the-art methods. As shown in the table, incorporating LANE consistently improves performance across most languages. XL-WiC+LANE achieves notable gains in English, French, German, Bulgarian, Chinese, and Dutch, demonstrating that adversarial word-level negatives enhance cross-lingual word-sense discrimination. These results confirm that LANE provides a measurable benefit even when applied to strong multilingual baselines. We further assess our approach in an unsupervised cross-lingual generalization scenario using XL-WiC across nine target languages: Bulgarian (bg), Chinese (zh), Croatian (hr), Danish (da), Dutch (nl), Estonian (et), Farsi (fa), Japanese (ja), and Korean (ko). The model is trained solely on XL-LEXEME data and compared to XL-DUREL, which leverages substantially more training resources. Despite this difference, XL-WiC+LANE achieves comparable performance, highlighting its data efficiency and robust cross-lingual generalization. Overall, these experiments extend the findings from the monolingual evaluation: LANE consistently enhances word-level representations, providing performance gains across languages, and supporting effective sense disambiguation in both supervised and cross-lingual settings.

[2]Omitted for review.

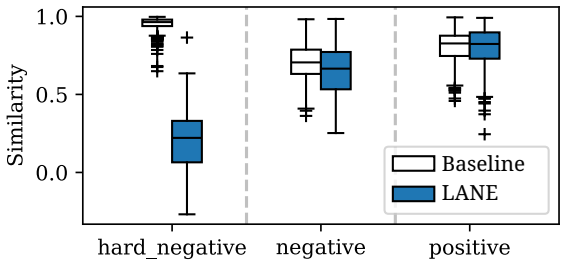

Figure 2: Similarity for representations learned under a traditional regime and under LANE.

6 Discussion

Our results highlight a persistent limitation of contextual language models: even highly capable encoders often rely on global sentence semantics rather than representing the fine-grained meaning of the target word. Consistent with prior work (Liu et al., 2021), we observe that embeddings of the same sentence remain nearly identical regardless of which word is marked, indicating that lexical information contributes little to the learned representation. This reliance on coarse contextual cues undermines interpretability and restricts generalization to new domains or languages where such cues differ.

The proposed LANE framework mitigates this issue by introducing lexically controlled adversarial negatives—sentences that differ only in the marked token. This formulation constrains the learning process to capture distinctions that are attributable to the target word itself. As shown in Figure 2, this results in more structured embedding spaces: sentences containing the same sense are drawn closer together, while those differing in meaning are pushed farther apart. The improved intra-sense cohesion and inter-sense separation explain the consistent accuracy gains reported in the evaluation.

Figure 3 further illustrates that models trained with LANE attend more strongly to the marked token, suggesting that adversarial negatives shift

Table 4: Comparison of classifiers with LANE for supervised multi-lingual (XL-WiC).

Model	en	fr	de	it	bg	zh	hr	da	nl	et	fa	ja	ko	MEAN
XL-LEXEME	0.722	0.785	0.848	**0.756**	0.827	0.802	0.727	**0.766**	0.782	0.664	0.666	0.682	0.798	0.755
XL-DUREL**	0.732	0.778	0.850	0.729	0.754	0.777	**0.752**	0.756	0.792	0.676	**0.706**	**0.697**	**0.801**	0.753
XL-WiC+**LANE**	**0.734**	**0.804**	**0.871**	0.746	**0.847**	**0.805**	0.735	**0.766**	**0.800**	**0.684**	0.642	0.683	0.799	**0.762**

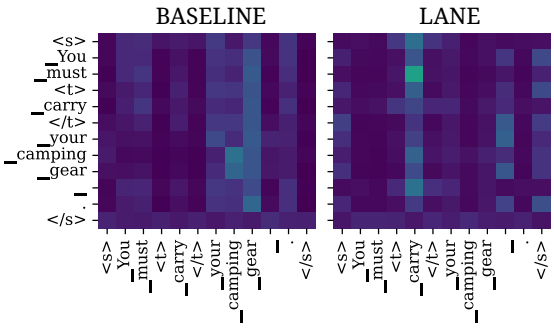

Figure 3: Last attention head heatmap for the baseline and LANE.

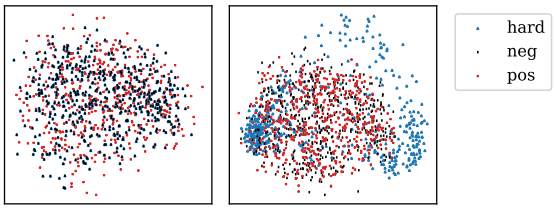

Figure 4: PCA for representations learned under a traditional regime (left) and under LANE (right). Adversarial negative examples are placed further apart in the embedding space.

attention away from irrelevant contextual cues and toward the lexical element being disambiguated. This demonstrates that the benefits of LANE are not merely representational but also functional, influencing how the model allocates its focus during inference.

Finally, Figure 4 shows that LANE yields more isotropic and semantically organized vector spaces. By evenly distributing representations and increasing separation among adversarial examples, the method promotes a more meaningful use of the embedding space.

In summary, LANE enhances both lexical sensitivity and representational structure in contextual models. Beyond improving WSD and cross-lingual generalization, it offers a principled mechanism for aligning neural optimization with linguistic distinctions—an essential step toward models that genuinely encode meaning rather than memorizing context.

7 Conclusion

We presented LANE, a simple and computationally efficient method for generating adversarial negatives in lexical representation tasks. By focusing on the target word rather than the surrounding context, LANE encourages models to encode fine-grained lexical distinctions, resulting in more robust and discriminative representations. Our experiments demonstrate that LANE consistently improves performance across monolingual and multilingual word sense disambiguation benchmarks, including out-of-distribution settings, without requiring complex hyperparameter tuning.

Importantly, we show that traditional approaches using marked words alone often fail to induce representations that prioritize lexical meaning, instead relying heavily on sentence-level contextual cues. LANE addresses this limitation by explicitly enforcing word-level separability, producing embeddings that better reflect true semantic distinctions.

Overall, LANE provides a simple, model-agnostic, and effective strategy for enhancing lexical semantic differentiation, improving generalization, and strengthening interpretability. Its ease of integration into existing frameworks makes it a practical tool for a wide range of lexical representation tasks, including cross-lingual and low-resource scenarios.

Limitations

A potential limitation of LANE lies in its word-substitution mechanism. If a randomly selected substitute is a synonym or near-synonym of the target word (e.g., replacing “buy” with “purchase” in the same sentence), the resulting pair may constitute a false negative, encouraging the model to separate semantically identical contexts. Although such occurrences are rare, the current formulation does not explicitly prevent them, representing an area for future refinement. Additionally, LANE assumes simple tokenization by spaces, which may be insufficient for languages with complex word formation or rich morphology, potentially affecting its ability to accurately distinguish lexemes in such languages.

References

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465, Online. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in characterization of semantic change. *ArXiv*, abs/2402.19088.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Dandan Huang, Lu Cao, Zhenting Li, and Yue Zhang. 2024a. Which sense dominates multisensory semantic understanding? a brain decoding study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17557–17563, Torino, Italia. ELRA and ICCL.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S. Yu. 2024b. Cosent: Consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2800–2813.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Bastien Liétard, Pascal Denis, and Mikaela Keller. 2024. To word senses and beyond: Inducing concepts with contextualized language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2684–2696, Miami, Florida, USA. Association for Computational Linguistics.

Bastien Li'etard and Gabriel Loiseau. 2025. Cale : Concept-aligned embeddings for both within-lemma and inter-lemma sense differentiation.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael A. Lones. 2024. Avoiding common machine learning pitfalls. *Patterns*, 5(10):101046.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. 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, Online. Association for Computational Linguistics.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Discovery team at SemEval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 67–73, Barcelona (online). International Committee for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

John Philip McCrae. 2019. Identification of adjective-noun neologisms using pretrained language models. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Anna Mosolova, Marie Candito, and Carlos Ramisch. 2024. Injecting Wiktionary to improve token-level contextual representations using contrastive learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 34–41, St. Julian's, Malta. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *ArXiv*, abs/2010.04592.

Joshua Robinson, Li Sun, Ke Yu, K. Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. 2024. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara

McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mary Ann Walter et al. 2007. *Repetition avoidance in human language*. Ph.D. thesis, Massachusetts Institute of Technology.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: Contrastive learning with semantic negative examples for natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2332–2342, Online. Association for Computational Linguistics.

Wei Wang, Liangzhu Ge, Jingqiao Zhang, and Cheng Yang. 2022. Improving contrastive learning of sentence embeddings with case-augmented positives and retrieved negatives. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Lei Xu, Sarah Alnegheimish, Laure Berti-Equille, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2025. Single word change is all you need: Using llms to create synthetic training examples for text classifiers. *Expert Systems*, 42(8).

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. 2020. Hard negative examples are hard, but useful. In *European Conference on Computer Vision*.

Sachin Yadav and Dominik Schlechtweg. 2025. Xl-durel: Finetuning sentence transformers for ordinal word-in-context classification. *ArXiv*, abs/2507.14578.

8 Scheduling Adversarial Negatives

The early introduction of adversarial examples traps the model in a local optima, as we show in the Figure below.

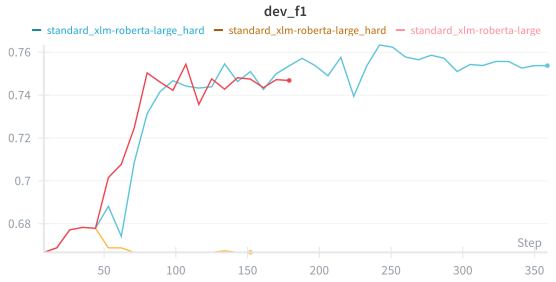

Figure 5: F1 score in the development data with different training settings.

The red line represent a training without adversarial examples, the yellow line the training with an early insertion of adversarial examples (first epoch), and the blue line the scheduled insertion of adversarial examples.

9 Datasets Licensing and Usage

The datasets we compiled from existing resources (MASC, SEMCOR, FEWS), are under CC BY-SA 4.0. WIC and XL-WIC belongs to the original licensing CC BY-NC 4.0.

This data is intended to use for word in context differentiation, similar to the word-in-context tasks. In the table below we list the statistics of each dataset.

Dataset	Split	Instances
SEMCOR	Train	33,313
SEMCOR	Dev	7,000
SEMCOR	Test	9,674
MASC	Train	7,280
MASC	Dev	7,000
MASC	Test	2,968
FEWS	Train	132,237
FEWS	Dev	7,000
FEWS	Test	29,708
DWUG	Train	32,424
DWUG	Dev	7,000
DWUG	Test	6,993

Table 5: Dataset split sizes for SemCor, MASC, FEWS, and DWUG.

10 Compute Costs

To train the models on SEMCOR, MASC, DWUG, WIC, and FEWS we used a NVIDIA V100 32GB for less then 4 hours. The XL-WIC takes approximately 35 hours in the same GPU.

11 Use of AI

We use AI as a code assistant and as a writing assistant, for improving grammar and readability.