

Evaluating Latent Semantic Pre-training for Fine-grained Word Sense Disambiguation

Soma Simon

University of Szeged
Institute of Informatics
somas@inf.u-szeged.hu

Gábor Berend

University of Szeged
Institute of Informatics
berendg@inf.u-szeged.hu

Abstract

Understanding word meaning in context is a central problem in Natural Language Processing, most directly addressed by word sense disambiguation (WSD). While modern contextual language models have improved WSD performance, it remains unclear how the choice of pre-training objective influences the semantic representations used in different WSD approaches. In this paper, we investigate masked latent semantic modeling (MLSM) as a semantic pre-training objective for WSD, focusing on its behavior across diverse WSD approaches. We contrast the use of MLSM pre-trained models to that of traditionally pre-trained ones for a wide range of WSD modeling scenarios, including centroid and sparse representation-based probing methods, as well as supervised disambiguation systems. Experimental results on standard WSD benchmarks show that relying on MLSM yields consistent improvements in disambiguation performance across models and datasets.

1 Introduction

Word sense disambiguation is the task of identifying the intended meaning of a word in context, given a fixed sense inventory. Although recent transformer-based models and neural WSD approaches have led to substantial improvements on standard benchmarks, WSD remains a challenging problem at the core of language understanding (Maru et al., 2022; Meconi et al., 2025), as subtle semantic differences between word senses often require representations that capture structured semantic information beyond surface lexical patterns.

Standard masked language modeling (MLM), the dominant pre-training objective for modern language models, trains models to predict the identity of masked tokens from context. However, this objective is inherently ambiguous: multiple words may plausibly fit the same context, and predicting exact lexical forms does not necessarily encourage

models to learn representations that reflect structured semantic properties.

Masked Latent Semantic Modeling (MLSM) (Berend, 2023) addresses this limitation by replacing lexical prediction with the prediction of latent semantic properties. Instead of reconstructing exact token forms, models trained with MLSM learn to predict distributions over latent semantic features associated with tokens. In this paper, we carefully investigate whether MLSM pre-trained models develop an improved capability in capturing the semantic distinctions of words.

Although MLSM has previously been evaluated on several supervised fine-tuning tasks, its ability to improve semantic representations has not yet been systematically examined in settings that directly probe semantic discrimination. As WSD requires models to distinguish between closely related meanings, it offers a useful setting for evaluating whether MLSM pre-training leads to semantically more informative representations.

In this work, we conduct a systematic evaluation of MLSM pre-trained models for the task of WSD. That is, our primary goal is to isolate the effect of the pre-training objective through controlled experiments. To this end, we pre-train comparable models using both MLM and MLSM under identical conditions and evaluate them across multiple WSD datasets and modeling paradigms.

Our experiments cover a range of widely used WSD approaches, including models based on contextual embeddings as well as supervised disambiguation systems. In addition to these architectures, our evaluation includes representation-based probing in which semantic structure is examined directly in the learned representations. Rather than training classifiers on top of frozen embeddings, which can lead to misleading conclusions about the information encoded in representations (Belinkov, 2022; Hewitt and Liang, 2019), we analyze representations directly. Specifically, we compute the

sense centroids of contextual embeddings and measure their similarity statistics, a common technique in representation-based semantic analyses (Teglia et al., 2025).

Our contributions are as follows:

- We pre-train and release MLM and MLSM model checkpoints pre-trained under identical setups, enabling the direct and controlled comparison of the effects of their objectives.
- We evaluate these models across multiple WSD datasets and modeling paradigms, including both representation-based and supervised WSD approaches.

Overall, our results provide evidence that replacing MLM with MLSM during pre-training can lead to consistent improvements across a range of WSD model architectures and datasets, suggesting that semantically oriented pre-training objectives can benefit tasks requiring fine-grained meaning distinctions.

2 Related Work

The task of WSD is widely used as a benchmark to evaluate how well contextual representations capture fine-grained semantic distinctions. Unlike broader downstream tasks, WSD requires models to distinguish between closely related meanings of the same word occurrence based solely on contextual cues, making it a sensitive probe of semantic structure in learned representations (Raganato et al., 2017). Recent work further highlights that WSD remains a valuable diagnostic for evaluating the semantic capabilities of modern large language models (Navigli, 2026).

2.1 WSD frameworks

The introduction of contextualized word representations, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), substantially improved WSD performance by modeling word meaning as a function of context rather than a fixed vector. Transformer-based encoders have since become the foundation of most state-of-the-art WSD systems, enabling fine-grained semantic distinctions that were difficult to capture with earlier models (Hadiwinoto et al., 2019). Nevertheless, recent analyses suggest that contextual representations still conflate senses under certain conditions and that their semantic structure is sensitive to modeling

choices and objectives (Maru et al., 2022; Meconi et al., 2025).

Several prominent WSD approaches are built directly on contextual encoders while differing in how sense information is represented and exploited. Early transformer-based models such as GlossBERT (Huang et al., 2019) formulate WSD as a contextual matching task by jointly encoding the sentence and candidate glosses and predicting whether a gloss correctly describes the target word. Bi-encoder models (BEM) perform WSD by comparing contextualized representations of the target word with separately encoded sense definitions, selecting the sense whose embedding is most similar (Blevins and Zettlemoyer, 2020).

In contrast, Barba et al. (2021a) reframed WSD as an extractive sense comprehension task (ESC), where all candidate glosses for a target word are considered jointly within a single model. Their ESCHER model encodes the sentence context jointly with all candidate glosses in a single forward pass, instead of encoding each gloss independently and computing a similarity score with the context representation.

ConSec introduces a sequential disambiguation strategy in which predictions are made following an ordering over the targets in a sentence. Each decision can incorporate the explicit senses assigned to previously disambiguated words, allowing the model to refine the predictions based on previous assignments (Barba et al., 2021b).

EWISER (Bevilacqua and Navigli, 2020) combines transformer-based contextual representations with graph-based propagation over lexical knowledge bases, allowing sense predictions to be informed by the relational structure of WordNet.

More recently, Guzman Olivares et al. (2025) proposed a framework that reframes WSD as cluster-based sense discrimination over a semantic network, incorporating neighboring concepts to improve robustness and performance.

Another emerging direction investigates the use of large language models and decoder-based prompting for WSD (Meconi et al., 2025; Qorib et al., 2024; Yae et al., 2025). These studies evaluate whether generative models can resolve lexical ambiguity directly from context, for example, by predicting sense labels or explaining word meanings.

Although these methods differ in their supervision, architecture, and inference strategy, they all critically depend on the quality and semantic struc-

ture of the underlying contextual representations.

2.2 Sparsity as a proxy to senses

In addition to advances driven by increasingly large contextual language models, prior work has explored structured and interpretable semantic representations. Sparse word embeddings have been shown to encode meaningful semantic dimensions while improving interpretability and performance on lexical semantic tasks (Faruqui et al., 2015). Building on this idea, Berend (2020) introduced sparse contextual word representations derived from dense transformer embeddings, demonstrating consistent gains on fine-grained WSD benchmarks. These results suggest that explicitly structured semantic representations can complement the expressive power of contextual encoders.

Masked latent semantic modeling (MLSM) (Berend, 2023) proposes an alternative pre-training objective to standard masked language modeling by replacing token prediction with the estimation of latent semantic features. By encouraging models to represent words as distributions over semantic properties, MLSM aims to induce representations that better capture sense distinctions and semantic regularities. Although originally motivated by sparse semantic representations, its impact across heterogeneous WSD modeling paradigms and datasets has not yet been systematically studied.

To address this gap, we evaluate MLSM as a semantic pre-training objective across several WSD paradigms, including representation-based probing, PMI-based sparse contextual representations, and supervised disambiguation architectures such as BEM, ESC, and ConSec. Rather than introducing a new WSD model, our goal is to examine how MLSM affects contextual representations under different inductive biases, providing a systematic comparison of latent semantic pre-training across diverse WSD approaches.

3 Masked Latent Semantic Modeling

MLSM is an alternative pre-training objective for language models that replaces lexical identity prediction with the prediction of latent semantic properties. While standard MLM trains models to recover the exact surface form of a masked token, this objective conflates semantic compatibility with lexical choice. However, in many contexts, multiple lexical items may be equally plausible, whereas the underlying semantic category is more stable

and informative. For instance, in the sentence “I ate [MASK] for breakfast.”, many different words could plausibly fill the masked position. While predicting the exact lexical item is therefore inherently ambiguous, identifying the latent semantic category (e.g., food) better reflects the contextual meaning.

To operationalize this intuition, MLSM modifies the training objective so that the model predicts a distribution over latent semantic properties as opposed to vocabulary items. These latent properties are automatically induced using sparse dictionary learning over contextualized representations, yielding interpretable and selective semantic targets.

3.1 Latent semantic targets via sparse coding

Sparse word representations differ from dense embeddings in that only a small number of dimensions are active for any given input, while the remaining dimensions are zero. This property encourages selectivity and interpretability, as individual dimensions tend to correspond to salient semantic features.

We construct sparse contextualized representations by applying dictionary learning to hidden states extracted from a fixed transformer layer l of a pre-trained language model. Throughout this section, l is fixed to the last transformer layer. Consequently, the dictionary D and sparse coefficients α are learned from representations extracted from that layer, and we omit an explicit layer index from them for readability. We learn an overcomplete dictionary $D \in \mathbb{R}^{d \times k}$ and a corresponding sparse coefficient matrix $\alpha \in \mathbb{R}_{\geq 0}^{k \times m}$, where d denotes the dimensionality of the hidden representations, k the number of dictionary atoms, and m the number of token instances. Each contextual embedding is reconstructed using only a small subset of the dictionary atoms by solving the following optimization problem:

$$\min_{\substack{D \in \mathcal{C} \\ \alpha \in \mathbb{R}_{\geq 0}^{k \times m}}} \sum_{i=1}^M \sum_{j=1}^{N_i} \frac{1}{2} \left\| \mathbf{h}_j^{(i)} - D\alpha_j^{(i)} \right\|_2^2 + \lambda \left\| \alpha_j^{(i)} \right\|_1.$$

where $\mathcal{C}$ denotes the set of admissible dictionaries, $i \in \{1, \dots, M\}$ indexes sentences in the corpus of token length N_i , and $j \in \{1, \dots, N_i\}$ denotes token positions within the sentence. The vector $\mathbf{h}_j^{(i)} \in \mathbb{R}^d$ denotes the contextual representation of the j -th token in sentence i , $\alpha_j^{(i)} \in \mathbb{R}_{\geq 0}^k$ is its sparse coefficient vector, D is the dictionary matrix

containing the learned basis atoms, and λ controls the degree of sparsity.

Once the dictionary D has been learned, it is kept fixed. For a target token representation $\mathbf{h}_j^{(i)}$, a sparse contextualized representation is obtained by solving the following LASSO problem:

$$\min_{\alpha_j^{(i)} \in \mathbb{R}_{\geq 0}^k} \frac{1}{2} \left\| \mathbf{h}_j^{(i)} - D\alpha_j^{(i)} \right\|_2^2 + \lambda \left\| \alpha_j^{(i)} \right\|_1.$$

The resulting sparse coefficient vector $\alpha_j^{(i)}$ is then normalized to ℓ_1 , producing a probability distribution over the k latent semantic dimensions. We refer to this distribution as the *semantic profile* of the token. These semantic profiles serve as training targets in MLSM.

3.2 MLSM training objective

Based on the sparse semantic representations described above, MLSM adopts a teacher-student training paradigm (Berend, 2023). A teacher model is first used to produce semantic profiles for masked tokens by applying sparse coding to its hidden states. During training, the student model is required to predict these semantic profiles from the masked input, rather than predicting the original lexical items.

In practice, the k latent semantic dimensions induced by the dictionary are added to the model’s output space. For each masked token, the loss is calculated by comparing the predicted distribution of the student in the semantic dimensions with the target semantic profile produced by the teacher. This objective encourages the student model to recover the latent semantic compatibility of the masked token with its context.

By separating the prediction of meaning from the prediction of a specific word, MLSM biases the model toward learning representations that better reflect the underlying semantic structure. This is especially useful for words whose meaning depends on context, and it also opens the door to applying the model in other tasks that benefit from capturing the core semantic content of language.

4 Experiments

4.1 Experimental setup

We evaluated our models on the popular WSD Hard benchmark (Maru et al., 2022) that provides challenging and fine-grained sense distinctions in English.

The benchmark is organized into several subsets to capture different levels of difficulty and coverage. The ALLamended and S10amended (ALL_{NEW} and $S10_{NEW}$) subsets are cleaned versions of the original ALL and SemEval-2010 datasets, containing 4,917 and 955 polysemous examples, respectively. The 42D subset is a manually annotated multi-domain challenge set with 370 examples that do not appear in SemCor and whose correct senses are not the most frequent sense in WordNet. The softEN subset contains 5,766 examples that at least one of the state-of-the-art WSD systems evaluated in the original study was able to disambiguate correctly. These instances are selected from the union of ALLamended, S10amended, and 42D, making it the largest and most representative subset. Finally, the hardEN subset consists of the remaining 476 particularly challenging instances that current systems fail to solve. Unless stated otherwise, we report micro-F1 as the primary metric for all experiments, following standard WSD evaluation practice.

4.2 Pre-training

To ensure a controlled comparison, we pre-train two models with identical DeBERTa-base architectures. The first model follows the standard masked language modeling (MLM) objective, while the second is trained using MLSM with latent semantic targets. Apart from the pre-training objective, all architectural components and hyperparameters remain identical across the two models.

Both encoders correspond to base-sized DeBERTa configurations (He et al., 2021) with approximately 140 million parameters and 12 transformer layers. For the MLSM pre-trained model, semantic targets are represented in a 3072-dimensional latent semantic space derived from the final-layer representations of BERT-base-cased.

Pre-training is performed from scratch on a cleaned version of the English Wikipedia corpus prepared using the WikiBERT preprocessing pipeline (Pyysalo et al., 2021), resulting in a dataset of roughly three billion tokens. During training, the models process approximately 100 million input sequences, corresponding to about 1.5 passes over the corpus. Optimization is performed using AdamW (Loshchilov and Hutter, 2019) with a linear learning-rate schedule that peaks at $1e-4$ and a warm-up phase covering 1% of the total updates. We used an effective batch size of 1024 sequences, implemented via gradient accumulation with a per-

device batch size of 128. All experiments are implemented using the transformers library (Wolf et al., 2020).

4.3 Centroid-based disambiguation

We first evaluate a simple centroid-based WSD approach to probe the semantic quality of contextual representations, following a setup similar to Teglia et al. (2025).

In our main experiment, for each WordNet sense, we compute a dense centroid by averaging contextualized token representations extracted from the last layer of the model, using its annotated occurrences in SemCor (Miller et al., 1994) and WordNet examples, ensuring coverage even for less frequent senses. In a separate analysis presented later, we also explore the construction of centroids from the WNGT dataset to further extend coverage and evaluate the resulting performance.

At evaluation time, the candidate senses for each target word are obtained from WordNet (Fellbaum, 1998) and filtered by part of speech. We represent the target word using its contextualized token embedding in the test sentence, averaging embeddings when the target spans multiple tokens. For each remaining candidate sense, we compute the cosine similarity between its centroid and the target representation, selecting the sense with the highest similarity as our prediction.

4.4 Sparse contextualized representations

Motivated by the fact that MLSM is explicitly trained to predict sparse latent semantic profiles, we hypothesize that MLSM pre-trained models are particularly well suited for producing structured sparse contextualized representations. Following Berend (2020), we therefore use these representations to evaluate whether the pre-training objective aligns with structured semantic and lexical sense representations.

Building on the sparse semantic profiles introduced in Section 3.1, we next investigate whether these representations can be directly leveraged for word sense disambiguation.

Using the learned dictionary D , we compute sparse coefficient vectors for tokens annotated with sense in the training corpus. Based on these sparse representations, we construct a basis-sense co-occurrence matrix that aggregates how strongly each latent semantic dimension is activated for each WordNet sense. This matrix is then transformed into a basis-sense association matrix using

a positive PMI-based scoring criterion, which emphasizes selective semantic associations between latent dimensions and symbolic sense labels.

At inference time, given a contextualized token representation, we obtain its sparse semantic profile by solving the LASSO problem with the fixed dictionary D . Sense scores are then computed using the precomputed basis-sense association matrix, and the highest scoring sense is selected as the prediction.

In contrast to the centroid-based approach, which relies on similarity in dense embedding space, this method performs disambiguation in the induced latent semantic space, explicitly linking sparse semantic dimensions to lexical sense inventories.¹

4.5 Supervised WSD Models

The previous experiments focused on representation-level and weakly supervised disambiguation settings. We also investigate whether MLSM pre-training benefits fully supervised WSD architectures that are explicitly trained to predict sense labels from annotated corpora.

We evaluated three representative supervised WSD architectures: BEM, ESC, and ConSec. For each architecture, we construct two variants that differ only in encoder initialization, while keeping all other architectural components and training configurations identical within that model.

Unless otherwise stated, we follow the original implementations and training setups, using SemEval-2007 as the development set with early stopping based on validation performance, and apply only minimal modifications to integrate our encoder initialization. To account for variability from random initialization, we repeat training with multiple seeds and report the mean performance.

For BEM, we adopt the standard bi-encoder training and inference procedure, where the target-word representation and each candidate gloss are encoded independently. Disambiguation is performed by comparing the contextual embedding to the gloss embeddings, allowing us to directly assess whether MLSM improves the alignment between contextual and definitional representations.

ESC jointly encodes the sentence and all candidate glosses in a single transformer, allowing direct

¹Our source code for reproducing our experiments is available at <https://github.com/somas-research/combining>.

cross-attention between the context and gloss embeddings. This setup allows us to isolate the effect of the initialization of the MLSM-based encoder on the joint reasoning of the model over contextual and definitional information.

For ConSec, we follow the original training and inference setup, which performs sequential sense prediction over the targets in a sentence. The model incorporates information from previously assigned senses through a feedback mechanism, enabling predictions to be conditioned on earlier disambiguation decisions. Our evaluation examines whether MLSM-based encoder initialization improves the contextual representations used by this sequential reasoning process.

Together, these models cover complementary WSD paradigms: similarity-based retrieval (BEM), joint context–gloss reasoning (ESC), and sequential context-aware disambiguation (ConSec). This selection enables us to systematically evaluate the impact of MLSM pre-training across diverse model architectures.

5 Results

Figure 1 shows the performance of the evaluated architectures on the standard evaluation subsets. Both the Centroid and PMI approaches consistently benefit from MLSM pre-training compared to MLM, with the exception of the S10amended subset, where the PMI variant maintains performance rather than improving. This pattern suggests that MLSM produces representations better suited for structured semantic matching in word sense disambiguation. In particular, the PMI approach achieves a maximum F1 score of 0.7622 on the softEN dataset, compared to 0.7503 for the MLM baseline, which corresponds to an improvement of approximately 0.0119 F1. Despite its simplicity, the PMI method remains competitive with several supervised architectures, outperforming multiple model training seeds across the different evaluation subsets.

Improvements are also consistently observed across the supervised models. On the softEN dataset, ConSec benefits the most, with an average gain of 0.022 F1 across different seeds, while BEM and ESC improve by 0.0055 and 0.011 F1, respectively. These gains are robust across random seeds, demonstrating that the advantage of MLSM persists even when the models are fully trained on the WSD task. The consistent improve-

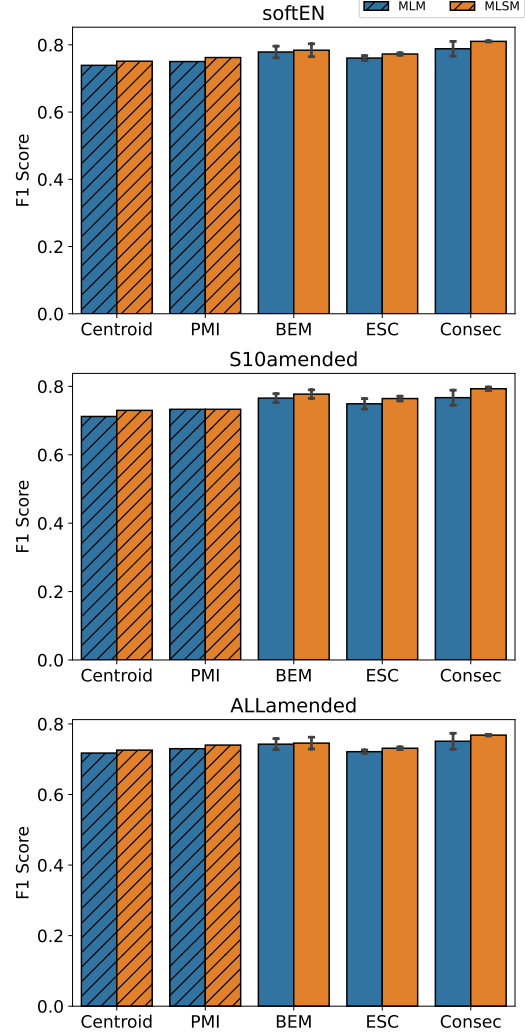


Figure 1: Results on the standard WSD evaluation datasets. Error bars indicate standard deviation across random seeds for supervised models. Centroid and PMI are deterministic representation-based methods and therefore have no error bars.

ments across heterogeneous architectures suggest that MLSM not only enhances the semantic structure of contextual representations, but that these improved representations are also more effective when used in fine-tuned models for downstream tasks.

Figure 2 shows the results on the challenge datasets. On hardEN, the effect of MLSM pre-training is mixed: while the average performance across seeds decreases for some architectures, individual seeds sometimes show improvements, reflecting the high variability inherent to this small and difficult dataset. In contrast, on 42D, all architectures benefit from MLSM, although performance still varies between seeds more than on the larger standard benchmarks.

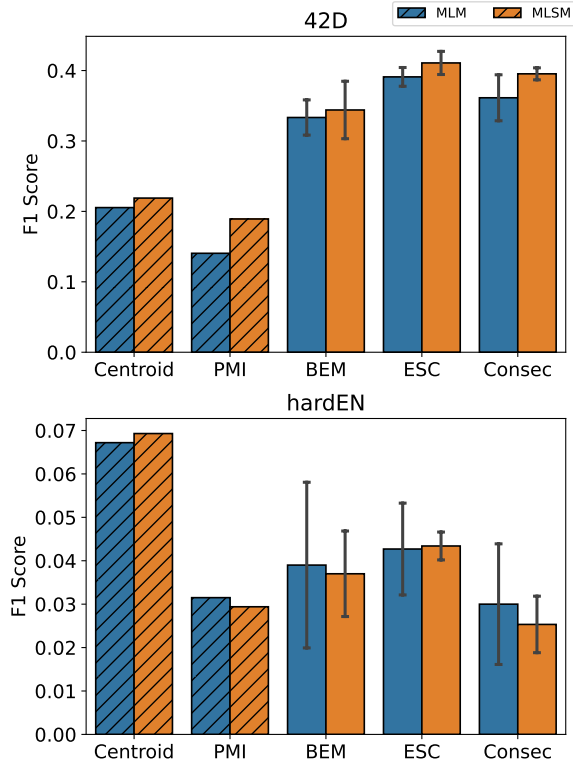


Figure 2: F1 scores on the challenging subsets. Centroid and PMI are non-parametric evaluation methods based on fixed representations and therefore do not involve multiple training seeds. Error bars are shown for the supervised methods and indicate the standard deviation across random seeds.

For the unsupervised methods, the PMI approach achieves an F1 score of 0.1892 with MLSM, compared to 0.1405 with MLM, corresponding to an improvement of approximately 0.05 F1 on 42D. These results suggest that, while MLSM primarily improves performance on typical or representative instances, it can also provide gains on more challenging cases.

5.1 Layer-wise analysis

To better understand the effects of MLSM pre-training, we conducted a series of analyses that go beyond the overall F1 performance. In particular, we examine how the improvements associated with MLSM manifest across the different layers of the transformer encoder. This analysis is performed for both the centroid-based and PMI-based approaches, allowing us to assess whether the benefits of MLSM are concentrated in specific layers or more evenly distributed throughout the model.

For this experiment, we used the ALLamended subset, as it is one of the largest evaluation subsets and exhibits relatively low bias with respect to the

types of examples it contains.

Figure 3 presents the layer-wise performance of the centroid-based probing method and the sparse PMI variant. Consistent with previous findings (Berend, 2020), the PMI approach consistently outperforms the centroid-based variant across all layers, regardless of the pre-training objective. In both methods, peak performance typically occurs in the final or penultimate layers of the encoder.

Regarding the effect of pre-training, the PMI approach benefits from MLSM in nearly every layer, with the exception of the third, and the gains are more prominent in the upper layers of the model.

For the centroid-based method, improvements appear from the sixth layer onward, where MLSM either increases or at least maintains performance relative to the MLM baseline. This pattern suggests that MLSM improves the semantic quality of contextual representations across the network, with the effects becoming particularly useful for WSD in the middle and upper layers.

5.2 Effect of training data coverage

In addition to the layer-wise evaluation, we further investigated the robustness of the observed improvements by repeating the centroid and PMI experiments under different training data conditions. Specifically, we varied the amount of available sense-annotated data used to construct the representations in order to determine whether the performance advantage of MLSM pre-training persists across different training corpora or whether it might be specific to a particular dataset configuration.

Figure 4 presents the results for the PMI-based method. As expected, increasing the coverage of the senses in the training data generally leads to improved performance in the evaluation datasets. The greatest improvements are observed on the more challenging benchmarks. For example, the 42D dataset was specifically designed to contain instances that cannot be easily resolved using only SemCor or the most frequent sense heuristic.

Consequently, when relying solely on SemCor data, the PMI method achieves an F1 score of below 0.01 on this dataset. However, when WordNet examples and the WNGT dataset are used, performance increases substantially, reaching 0.3703 F1, which is competitive with the results of several supervised methods across different training seeds.

Comparing the two pre-training objectives, representations extracted from models pre-trained with MLSM consistently lead to better performance

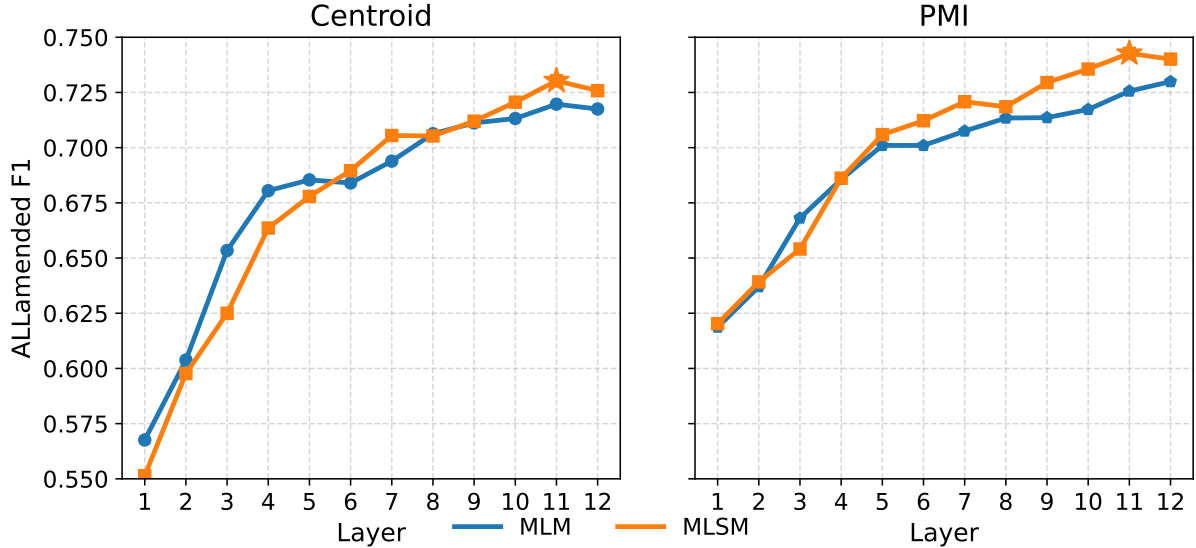


Figure 3: Layer-wise comparison of the centroid and PMI methods using MLM and MLSM pre-training on the ALLamended subset of the WSD Hard benchmark. Each point corresponds to the F1 score obtained from representations extracted from a specific transformer layer.

than those from MLM pre-trained models across nearly all evaluation datasets, with the exception of hardEN. This pattern is also consistent with the results reported in the previous section.

These findings indicate that the advantages of MLSM are not limited to a particular training corpus such as SemCor, but instead persist as additional training resources are used. This suggests that MLSM produces representations that scale effectively with increasing amounts of sense-annotated data. For completeness, we report the corresponding centroid-based experiments in Appendix A; the same general trend holds as additional sense-annotated resources are introduced.

5.3 Macro-F1 Results

Standard WSD evaluation typically reports micro-F1, which weights all instances equally and is therefore influenced by the distribution of frequent and infrequent senses. Since prior work has shown that standard WSD benchmarks can be affected by sense-frequency biases, we additionally report macro-F1 scores as a complementary metric. Macro-F1 gives equal weight to each sense class and therefore provides a less frequency-dominated view of model behavior.

Table 1 reports macro-F1 scores across the evaluation datasets. For the centroid and PMI methods, the values correspond to the same runs reported in the main evaluation.

The macro-F1 results are broadly consistent with

the micro-F1 results reported in the main paper. MLSM improves over MLM in most cases, although the gains are generally smaller under macro-F1. As in the main evaluation, hardEN remains the main exception and shows less consistent behavior. In addition, PMI and BEM show slight decreases on S10amended, suggesting that the MLSM advantage is weaker on this subset under macro-F1. Overall, these results indicate that the benefits of MLSM are not solely driven by frequent senses, but remain visible under a less frequency-dominated evaluation metric.

6 Conclusion

We presented an empirical study of Masked Latent Semantic Modeling as a semantic pre-training objective for word sense disambiguation. Rather than proposing a new architecture, we analyzed how MLSM influences semantic representation quality across multiple WSD paradigms. Our results show that MLSM consistently improves or matches masked language modeling across representation analysis, sparse semantic matching, and fully supervised WSD models.

In general, the findings suggest that MLSM promotes more structured and semantically coherent contextual representations, improving compatibility with downstream disambiguation mechanisms. These results highlight the potential of latent semantic prediction objectives as an alternative pre-training signal for fine-grained semantic tasks.

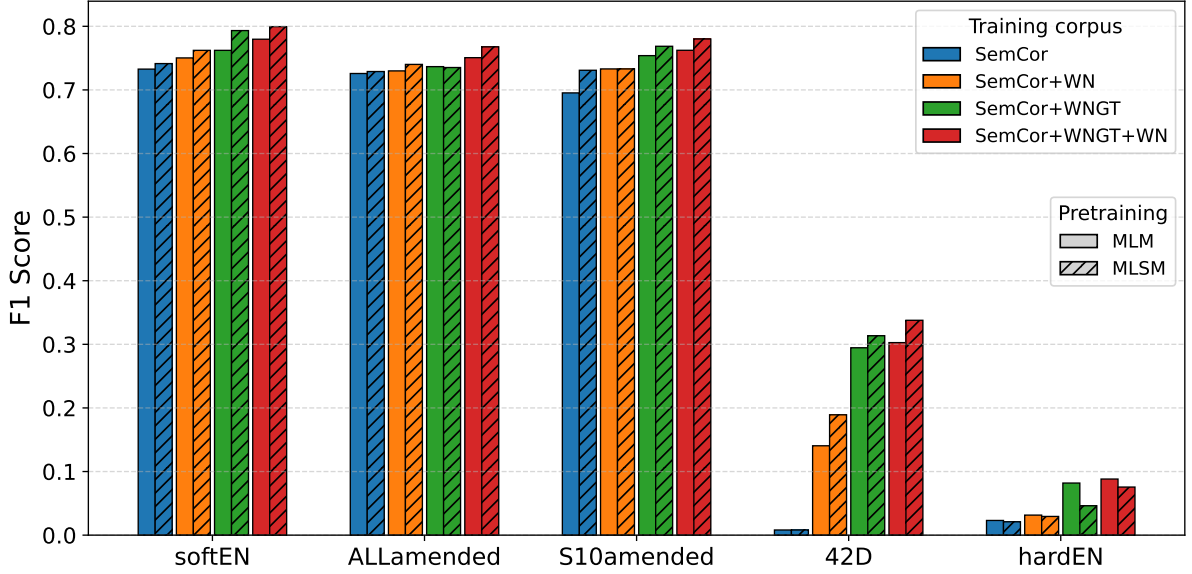


Figure 4: Performance of the PMI-based method across evaluation datasets under different training-data configurations. WN denotes the example sentences available in WordNet, whereas WNGT denotes the WordNet Gloss Corpus.

Method	Pre-training	softEN	hardEN	42D	ALLamended	S10amended	Avg.
Centroid	MLM	0.7039	0.1941	0.3475	0.6983	0.6908	0.5269
	MLSM	0.7300	0.1792	0.3785	0.7143	0.6980	0.5400
PMI	MLM	0.7116	0.1110	0.3049	0.7025	0.7147	0.5089
	MLSM	0.7365	0.0964	0.3312	0.7254	0.7105	0.5200
ESC	MLM	0.6886	0.0576	0.4228	0.6570	0.7028	0.5058
	MLSM	0.7027	0.0541	0.4393	0.6696	0.7062	0.5144
ConSec	MLM	0.7402	0.0348	0.4090	0.7090	0.7426	0.5271
	MLSM	0.7532	0.0263	0.4463	0.7165	0.7517	0.5388
BEM	MLM	0.7198	0.0577	0.3678	0.6948	0.7183	0.5117
	MLSM	0.7254	0.0529	0.3685	0.7038	0.7157	0.5133

Table 1: Macro-F1 scores across the evaluation datasets. Bold values indicate the better pre-training objective for each method and dataset. Avg. denotes the unweighted average across datasets.

Future work may investigate whether using model output logits rather than hidden representations affects WSD performance. Another promising direction is to explore decoder-based architectures, which were outside the scope of the present study.

Acknowledgments

This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The research received additional support from project no. 2024-1.2.3-HU-RIZONT-2024-00017, which has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, fi-

nanced under the 2024-1.2.3-HU-RIZONT funding scheme. Additionally, we are grateful for the possibility to use ELKH Cloud (see Héder et al., 2022; <https://science-cloud.hu/>) which helped us achieve the results published in this paper.

Limitations

This study focuses on English WSD benchmarks and the effectiveness of MLSM in multilingual settings remains to be explored, however, it would have required pre-training additional MLSM models on multilingual pre-training data, which would have incurred excessive additional computational cost. Additionally, our experiments compare only MLM and MLSM pre-training objectives. Including additional objectives would enable a broader

comparison, however, MLM is arguably the most predominantly applied pre-training approach for encoder-based neural language models.

References

- Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. [ESC: Redesigning WSD with extractive sense comprehension](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.
- Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. [ConSeC: Word sense disambiguation as continuous sense comprehension](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Gábor Berend. 2020. [Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8498–8508, Online. Association for Computational Linguistics.
- Gábor Berend. 2023. [Masked latent semantic modeling: an efficient pre-training alternative to masked language modeling](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13949–13962, Toronto, Canada. Association for Computational Linguistics.
- Michele Bevilacqua and Roberto Navigli. 2020. [Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.
- Terra Blevins and Luke Zettlemoyer. 2020. [Moving down the long tail of word sense disambiguation with gloss informed bi-encoders](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.
- 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.
- Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. [Sparse overcomplete word vector representations](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.
- C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. Mit Press.
- Daniel Guzman Olivares, Lara Quijano, and Federico Liberatore. 2025. [SANDWiCH: Semantical analysis of neighbours for disambiguating words in context ad hoc](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7019–7033, Albuquerque, New Mexico. Association for Computational Linguistics.
- Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. [Improved word sense disambiguation using pre-trained contextualized word representations](#). 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 5297–5306, Hong Kong, China. Association for Computational Linguistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.
- Mihály Héder, Ernő Rigó, Dorottya Medgyesi, Róbert Lovas, Szabolcs Tenczer, Ferenc Török, Attila Farkas, Márk Emődi, József Kadlecsek, György Mező, Ádám Pintér, and Péter Kacsuk. 2022. [The past, present and future of the ELKH cloud](#). *Információs Társadalom*, 22(2):128.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. [GlossBERT: BERT for word sense disambiguation with gloss knowledge](#). 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.

- Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. [Nibbling at the hard core of Word Sense Disambiguation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.
- Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavallo, and Roberto Navigli. 2025. [Do large language models understand word senses?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33897–33916, Suzhou, China. Association for Computational Linguistics.
- George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. [Using a semantic concordance for sense identification](#). In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Roberto Navigli. 2026. [Is word sense disambiguation dead in the llm era?](#) In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- 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.
- Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. [WikiBERT models: Deep transfer learning for many languages](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–10, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Muhammad Reza Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. [Are decoder-only language models better than encoder-only language models in understanding word meaning?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand. Association for Computational Linguistics.
- Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. [Word sense disambiguation: A unified evaluation framework and empirical comparison](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.
- Simone Teglia, Simone Tedeschi, and Roberto Navigli. 2025. [How much do encoder models know about word senses?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2266–2277, Vienna, Austria. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jung H Yae, Nolan C Skelly, Neil C Ranly, and Phillip M LaCasse. 2025. Leveraging large language models for word sense disambiguation. *Neural Computing and Applications*, 37(6):4093–4110.

A Centroid-based experiments under different training data conditions

Figure 5 presents the corresponding experiments for the centroid-based approach under different training data configurations. In general, we observe trends similar to those reported for the PMI-based method in Section 5.2. Increasing the amount of sense-annotated data generally improves performance across the evaluation datasets, particularly on the more challenging benchmarks.

Compared to the PMI-based approach, the centroid method typically achieves lower overall performance on most evaluation datasets. However, on 42D, centroid-based representations remain competitive and in some configurations outperform the PMI variant.

Across nearly all dataset configurations, MLSM pre-trained representations consistently outperform or match their MLM counterparts. These results reinforce the observations from the PMI-based experiments in Section 5.2 and indicate that the advantages of MLSM persist even as additional sense-annotated training resources are introduced.

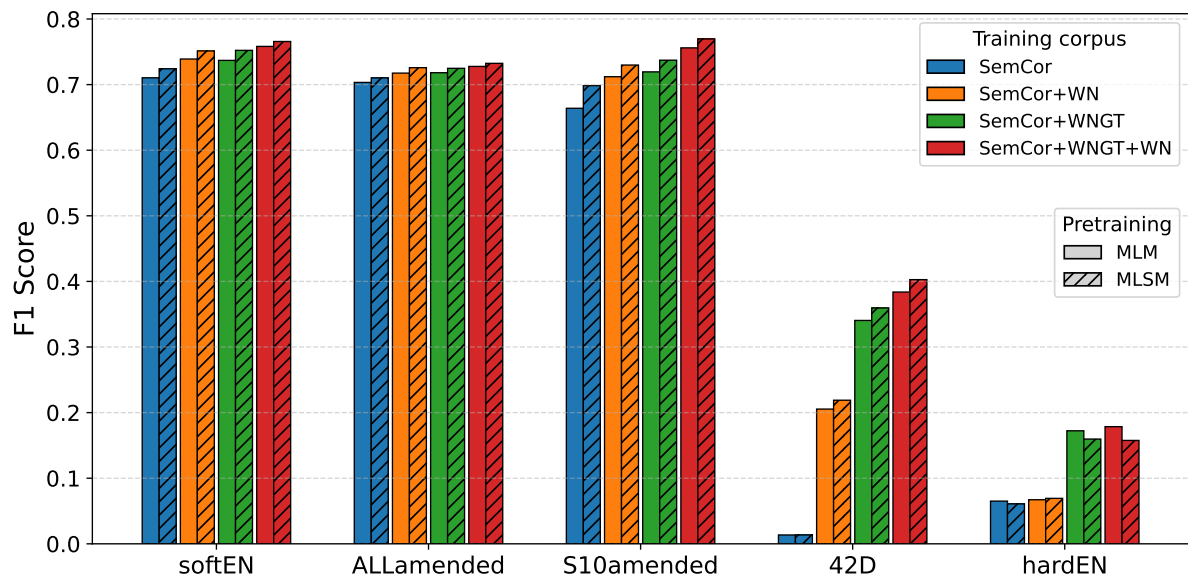


Figure 5: Performance of the centroid-based method across evaluation datasets under different training data configurations.