SemTechLab at SemEval-2026 Task 5: Context-Aware Homonym Disambiguation via Span-Specific Interaction Features

Karlo Babić[1,2], **Ana Meštrović**[1,2], **Slobodan Beliga**[1,2]
[1]University of Rijeka, Faculty of Informatics and Digital Technologies, Rijeka, Croatia
[2]University of Rijeka, Center for Artificial Intelligence and Cybersecurity, Rijeka, Croatia
Correspondence: sbeliga@inf.uniri.hr

Abstract

This paper presents the SemTechLab system submitted to SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding. The task involves predicting the plausibility of a specific word sense given a short story context. Our approach (HINTS) utilizes a hybrid Transformer architecture based on `nli-mpnet-base-v2`. Unlike standard Cross-Encoders that rely solely on the `[CLS]` token, HINTS extracts span-specific embeddings for the target homonym from both the narrative context and the sense definition. We compute interaction features (concatenation, difference, and element-wise product) between these spans to explicitly model the semantic alignment between the story and the proposed sense. The model is trained using Kullback-Leibler Divergence to predict the full distribution of human ratings. For the official submission phase, scores were rounded to integers (1–5). However, subsequent analysis and ablation studies detailed in this paper utilize continuous (float) scores derived from the expected value for improved metric sensitivity. On the test set, our best configuration, which relies exclusively on local homonym features, achieved a Spearman correlation of **0.603** and an accuracy of **75.8%**.

1 Introduction

Word Sense Disambiguation (WSD) is an important task in natural language processing (NLP) concerned with identifying the intended meaning of an ambiguous word in context. Recent advances in representation learning, driven by contextualized Transformer encoders and large language models (LLMs), have substantially improved performance on standard WSD benchmarks (Bevilacqua et al., 2021). These models capture fine-grained contextual semantics and discourse-relevant dependencies required for tasks that rely on context-sensitive interpretation (Chang et al., 2024; Zhao et al., 2023).

Traditionally framed as a single-label classification task, WSD assumes that one correct sense can be selected from a predefined inventory. However, this formulation abstracts away from the graded nature of sense applicability. Some studies show that contextual meaning may support multiple interpretations to varying degrees (Erk and McCarthy, 2009; Jurgens, 2012). Following this idea, AmbiStory formulates homonym interpretation as graded plausibility estimation in short narratives, thereby explicitly capturing partial compatibility and annotator uncertainty (Gehring and Roth, 2025). Motivated by AmbiStory, SemEval-2026 Task 5 asks systems to rate, for five-sentence English short stories containing a lexically ambiguous sentence, how plausible each of two candidate word senses is in the given narrative context on a 1–5 scale (Gehring et al., 2026).

In this paper, we introduce **HINTS** (**H**ybrid transformer architecture leveraging Homonym **INT**eraction **S**pan-specific features) to address the challenge of graded plausibility estimation. HINTS is motivated by the observation that, in narrative settings, plausibility judgments hinge on whether a candidate sense is supported by concrete evidence around the ambiguous mention, not just by a general impression of topical relatedness. We therefore model each instance as a paired (story, sense) input so the encoder can directly compare the narrative context with the provided sense information, leveraging MPNet-based representations (Song et al., 2020), while explicitly steering the representation toward the target homonym. Concretely, the model isolates span-specific representations of the homonym in both the story and the sense text and uses simple interaction signals to make their semantic agreement (or mismatch) more explicit. In addition, rather than collapsing annotator responses into a single score, HINTS is trained to predict the full distribution of human ratings, which allows it to better reflect uncertainty that

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2245–2250

is inherent to graded plausibility. Our experimental results and ablations show that emphasizing homonym-centered evidence is critical for performance and provides a transparent handle for analyzing typical failure cases under the official metrics.

2 Background

2.1 Task Description

The AmbiStory dataset consists of short stories containing a target homonym. For each sample, the inputs are:

1. **Story:** A precontext (3 sentences), the target sentence, and an ending.

2. **Sense:** A target meaning (gloss) and an example sentence.

The system must output a scalar prediction representing the plausibility (1–5). Evaluation metrics are Spearman correlation and Accuracy (percentage of predictions within one standard deviation of the human mean).

2.2 Related Work

A wide range of NLP tasks requires interpreting lexical meaning within broader contextual and narrative structures. Phenomena such as polysemy, synonymy, and contextual ambiguity require models to evaluate how alternative interpretations align with discourse-level information (Garcia, 2021). Related challenges arise in lexical ambiguity resolution (Haber and Poesio, 2024), idiom interpretation (Beliga and Filipović Petrović, 2025; Filipović Petrović and Beliga, 2025), and narrative reasoning tasks that involve selecting plausible continuations or alternatives based on causal and discourse coherence (Perak et al., 2024). These problems share a common requirement: modeling how local lexical meaning interacts with wider contextual structure. Among them, WSD represents widely studied NLP task focusing on assigning appropriate meanings to ambiguous words given their usage context.

Depending on the formulation, WSD is often treated as a single-label classification task (Conia and Navigli, 2021). However, empirical evidence suggests that sense applicability can be graded. Erk and McCarthy (2009) show that annotators assign intermediate ratings and may consider multiple senses simultaneously applicable, motivating evaluation beyond categorical decisions. Jurgens (2012) further formalizes graded word sense assignment by relaxing the single-label assumption and distinguishing between detecting applicable senses and modeling their relative strength within context. Based on these ideas, AmbiStory provides a narrative-based benchmark in which candidate senses for a target homonym are rated for plausibility within short stories, explicitly capturing partial compatibility and annotator uncertainty (Gehring and Roth, 2025).

Previous work in WSD has largely utilized BERT-based architectures (Devlin et al., 2019). Some approaches formulate WSD as modeling the interaction between a context and a candidate gloss using sentence-pair encoders. For example, GlossBERT treats WSD as a sentence-pair classification problem by constructing context–gloss pairs for each candidate sense and fine-tuning BERT to score their compatibility (Huang et al., 2019). More generally, these methods compare the usage context with each candidate sense definition and assign a compatibility score. Sentence-BERT proposed by Reimers and Gurevych (2019) popularized siamese encoders for semantic similarity, while NLI-trained encoders are particularly well suited for capturing entailment-like relations between paired texts. Building on this paradigm, our system adopts an NLI-MPNet-based architecture (Song et al., 2020) to explicitly model semantic alignment between the narrative context and a proposed word sense.

3 System Overview

The proposed **HINTS** employs a hybrid architecture that combines the contextualization benefits of a Cross-Encoder with the feature engineering typically found in Bi-Encoders.

3.1 Input Representation

The core idea behind our input design is to enable the model to perform direct semantic comparison between the narrative usage of a word and its formal definition. Our approach is motivated by recent findings that providing even small, expert-curated lexicographic evidence (such as sense definitions) at inference time can significantly mitigate model biases in challenging literal–idiomatic disambiguation tasks (Beliga et al., 2026). To address this, we construct two input strings for the tokenizer:

1. **Story Sequence (S):** Concatenation of the precontext, the sentence containing the homonym, and the ending.

2. **Sense Sequence (M):** We inject the judged meaning into the example sentence using parentheses immediately following the target word (e.g., "The detective found a crucial track (evidence pointing to a solution).").

These sequences are packed into a single input pair `[CLS] S [SEP] M [SEP]` to allow full self-attention across the story and meaning.

3.2 Span Identification

To extract focused representations, we identify the token spans corresponding to the homonym in both sequences. We employ a three-stage fallback strategy:

1. Strict word-boundary regex match.
2. Loose substring match.
3. Fuzzy match using Levenshtein distance to handle potential morphological variations.

This results in two binary masks, m_{story} and m_{sense}, indicating the positions of the homonym in the story and the sense definition, respectively.

3.3 Hybrid Architecture

We feed the tokenized input into `nli-mpnet-base-v2`. Let $H \in \mathbb{R}^{L \times d}$ be the last hidden state of the transformer, where L is sequence length and d is hidden dimension.

We compute the global context embedding $h_{CLS} = H_0$. Additionally, we compute span-specific embeddings u (story homonym) and v (sense homonym) via masked average pooling:

$$u = \frac{\sum_i H_i \cdot m_{story,i}}{\sum_i m_{story,i}}, \quad v = \frac{\sum_i H_i \cdot m_{sense,i}}{\sum_i m_{sense,i}} \tag{1}$$

To capture the relationship between the usage of the word in the story and the proposed meaning, we adopt the feature interaction strategy proposed by Conneau et al. (2017) and widely used in NLI tasks (Reimers and Gurevych, 2019). We construct a feature vector F combining the embeddings and their heuristic matching information:

$$F = [u; v; |u - v|; u * v; h_{CLS}] \tag{2}$$

where $[;]$ denotes concatenation, $|\cdot|$ is absolute difference, and $*$ is element-wise multiplication. By including both the span-specific interactions and the global h_{CLS}, proposed model (HINTS) captures both local alignment and global context consistency.

This vector $F \in \mathbb{R}^{5d}$ is passed through a feed-forward network (Linear $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear) to produce logits $z \in \mathbb{R}^5$.

These specific heuristic interaction features (element-wise difference $|u - v|$ and product $u * v$) were chosen because they have been empirically proven to be highly effective and computationally efficient representations of semantic divergence and similarity in Natural Language Inference (NLI) tasks (Conneau et al., 2017; Reimers and Gurevych, 2019). While more complex mechanisms, such as late cross-attention between spans, could theoretically be employed, these parameter-free operations provide a robust baseline for explicit semantic alignment without the risk of overfitting the relatively small training set.

3.4 Training Objective

The dataset provides individual annotator votes. Instead of training on the mean score (regression), we treat this as a distribution learning problem. We convert the raw votes into a probability distribution P_{target}.

The model outputs logits z, which are converted to probabilities $P_{pred} = \text{softmax}(z)$. We minimize the Kullback-Leibler (KL) Divergence loss:

$$\mathcal{L} = \mathcal{D}_{KL}(P_{target} || P_{pred}) \tag{3}$$

We opted for KL Divergence over a simpler regression objective (such as Mean Squared Error on the average score) because MSE inherently assumes a unimodal, Gaussian distribution of human judgments. Empirically, treating this as a distribution learning problem allows the model to better internalize annotator uncertainty. Rather than forcing the network to predict an artificial "middle" score when human opinions are highly polarized, KL divergence trains the model to recognize and output flattened or bimodal distributions when multiple interpretations are viable.

3.5 Inference

During inference, we calculate the expected value of the predicted distribution to obtain a continuous score:

$$\text{Score} = \sum_{k=1}^{5} k \cdot P_{pred,k} \tag{4}$$

For the official competition submission, this score was rounded to the nearest integer (1–5). However, for post-submission analysis and ablation

studies presented here, we utilize the continuous float score, as this provides a fairer comparison metric, especially for the Accuracy within Standard Deviation score.

4 Experimental Setup

4.1 Data and Preprocessing

We utilized the official train and development datasets[1] provided by the organizers. No external datasets were used. The maximum sequence length was set to 256 tokens.

4.2 Hyperparameters

The model was implemented using PyTorch and HuggingFace Transformers, the code is publicly available on GitHub [2]. We trained for 10 epochs with a batch size of 4. We employed a differential learning rate strategy:

- **Backbone (MPNet):** 5×10^{-6}
- **Classifier Head:** 5×10^{-5}

This allows the task-specific head to learn rapidly while preserving the pre-trained knowledge in the backbone. The dropout rate in the projection head was set to 0.1.

4.3 Evaluation

We evaluated the model on the development set using the official scoring script[3]. The primary metrics are Spearman Correlation and Accuracy (Correctness within standard deviation).

5 Results

5.1 Development Set Results

Table 1 presents the initial performance evaluation on the development set. The continuous float scoring achieved a significant gain in accuracy (77.0% vs 73.5%) compared to the integer-rounded submission score.

5.2 Ablation Studies on Test Set

We performed the final evaluation of all architectural configurations on the held-out test set using continuous float predictions, averaged over 8 seeds for robust statistical analysis. Table 2 presents these results. The test set results demonstrated strong consistency with the development phase.

The configuration relying only on Local Features ("w/o Global Context") proved marginally superior, achieving the highest performance metrics: Spearman $\rho = 0.603$ and Accuracy 75.8%. Given the similarity in performance between this configuration and the full baseline (Spearman 0.599), we conclude that the global `[CLS]` embedding does not contribute unique predictive power beyond what is already captured by the specific span features and their interactions. This redundancy suggests that the local-only strategy is the most efficient architectural choice.

Conversely, removing the local homonym features entirely ("w/o Local Context") resulted in the largest performance drop (Spearman ρ dropped to 0.468), solidifying the view that focusing on the target word spans is the most critical element of the architecture.

The interaction features ($|u - v|, u * v$) and the explicit Meaning Injection were highly critical. Removing the interactions caused Spearman correlation to drop from 0.603 to 0.534, while removing the explicit gloss caused a drop to 0.538. Both components are essential for precise semantic alignment. Finally, the specialized `nli-mpnet-base-v2` backbone maintained its clear advantage over the generic `bert-base-uncased` backbone (0.603 vs 0.534 Spearman).

To contextualize our performance within the shared task leaderboard, we compare our system's combined score (the average of Accuracy and Spearman ρ) against the organizer-provided baselines. Our best HINTS configuration achieves a combined score of 0.681 (averaging 0.758 Accuracy and 0.603 Spearman). This significantly outperforms the provided Llama-3.1 8B baseline (0.563) and remains competitive with the massive, proprietary GPT-4o model (0.756). This demonstrates that focused, span-specific feature engineering applied to smaller, computationally efficient NLI-encoders remains a highly viable and transparent approach for graded narrative WSD, without requiring the vast parameters of generalized LLMs.

6 Conclusion

We presented HINTS, a hybrid transformer approach for the SemEval-2026 AmbiStory task. By combining span-specific interaction features and training on the full distribution of human ratings, our model successfully captured the nuance of

[1] https://github.com/Janosch-Gehring/ambistory
[2] https://github.com/karlo-babic/semeval26-05
[3] https://github.com/Janosch-Gehring/semeval26-05-scripts

Model	Spearman (ρ)	Accuracy
Majority Baseline (Est.)	0.05	0.45
Random Baseline (Est.)	0.12	0.40
HINTS (Submission)	0.581	0.735
HINTS (Dev - Continuous)	0.581	0.770

Table 1: Performance on the Development Set. The Submission row reflects integer-rounded scores, while the Continuous row reflects float scores (Averages over 8 seeds).

Configuration	Avg. Spearman (ρ)	Avg. Accuracy
Full Model (Baseline)	0.599 ± 0.012	0.752 ± 0.011
w/o Interaction Features	0.534 ± 0.014	0.725 ± 0.014
w/o Global Context (h_{CLS})	$\mathbf{0.603 \pm 0.012}$	$\mathbf{0.758 \pm 0.008}$
w/o Local Context (u, v)	0.468 ± 0.021	0.696 ± 0.011
w/o Meaning Injection	0.538 ± 0.006	0.723 ± 0.010
BERT Backbone (BERT-Base)	0.534 ± 0.010	0.712 ± 0.008

Table 2: Final ablation results on the Test set using continuous predictions (Averages over 8 seeds).

word sense plausibility. Our ablation studies, conducted using multi-seed averages on the final test set, yielded two key findings:

1. The final best configuration relied exclusively on local homonym features and their interactions, demonstrating that the most relevant information for this task resides in the immediate semantic alignment of the target word usage and its definition, rather than the general contextual embedding (h_{CLS}).

2. The use of a backbone pre-trained for semantic similarity (`nli-mpnet-base-v2`) provided a significant and consistent performance boost over a standard BERT model.

Our best performing configuration of HINTS, utilizing continuous score predictions, achieved a Spearman correlation of 0.603 and an accuracy of 75.8% on the test set.

Limitations

We conducted a qualitative error analysis of the model's predictions on the development set to identify recurring limitations in our approach. The analysis highlights three primary sources of error:

1. Over-reliance on Local Context While local interaction features drive the model's overall performance, they cause errors when long-range narrative cues contradict the immediate syntax. For example (ID 572), the target sentence *"She watched the trailer carefully"* strongly aligns locally with the "movie trailer" sense. The model predicts a high plausibility (4.61), failing to integrate the global narrative ending (*"As it passed by, she saw what was inside"*) which indicates the vehicle sense (gold human average: 1.20).

2. Bimodal Human Disagreement The task features graded annotations that occasionally result in highly divided human judgments. For instance, in a story evaluating the word "rare" (ID 447), annotator votes were distinctly bimodal (`[1, 1, 5, 4, 5]`). While our KL Divergence objective naturally models this uncertainty by predicting bimodal distributions, collapsing these distributions into a single expected value for evaluation metrics (Accuracy and Spearman) inherently penalizes the model in highly ambiguous cases.

3. Sensitivity to Span Detection Because HINTS relies on extracting local features (u, v), it is sensitive to token alignment failures. If the target homonym undergoes extreme morphological changes not caught by the Levenshtein fallback, or if it is pushed beyond the 256-token sequence limit, span extraction fails. In such cases, the extraction mask defaults to the `[CLS]` token. This safely prevents runtime errors but degrades the model's capability to that of a standard cross-encoder, neutralizing the benefits of the interaction features.

Acknowledgments

This research was supported by the project Hybrid AI Approaches to Natural Language Processing and Knowledge Generation – HyAI

(uniri-iz-25-215), funded by the European Union – NextGenerationEU.

References

Slobodan Beliga and Ivana Filipović Petrović. 2025. Ai- and corpus-based strategies for identifying phraseme constructions: A pilot study on croatian repetitive constructions. *Electronic lexicography in the 21st century (eLex 2025) Intelligent Lexicography*, pages 95–115.

Slobodan Beliga, Ivana Filipović Petrović, and Ana Meštrović. 2026. Injecting structured lexicographic knowledge into LLMs for non-literal expression disambiguation: A controlled study on Croatian. In *Proceedings of the LREC 2026 Workshop: Learning Non-Literal Expressions with Small Data*, Palma de Mallorca, Spain. ELRA. Forthcoming.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, and 1 others. 2021. Recent trends in word sense disambiguation: A survey. In *IJCAI*, pages 4330–4338.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 440–449.

Ivana Filipović Petrović and Slobodan Beliga. 2025. Can AI understand Croatian idioms? Assessing large language models in lexicographic tasks. *Prispevki za novejšo zgodovino*, 65(3):218–242.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. 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 3625–3640.

Janosch Gehring, Selina Meyer, and Michael Roth. 2026. SemEval-2026 task 5: Rating plausibility of word senses in ambiguous stories through narrative understanding. In *Proceedings of the 20th International Workshop on Semantic Evaluation*, San Diego, California. Association for Computational Linguistics.

Janosch Gehring and Michael Roth. 2025. Ambistory: A challenging dataset of lexically ambiguous short stories. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (* SEM 2025)*, pages 152–171.

Janosch Haber and Massimo Poesio. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing 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.

David Jurgens. 2012. An evaluation of graded sense disambiguation using word sense induction. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 189–198.

Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating dialect understanding into llm using rag and prompt engineering techniques for causal commonsense reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 220–229.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124.