PuerAI at SemEval-2026 Task 5: Homograph Appropriateness Assessment via DeBERTa Contrastive Regression and Contextual Grouping

Jiaxu Dao, Zhuoying Li, Hangchao Ma,
Jinli Tong, Xiaoli Lan, Yifan Lu, Zhanji Yang
School of Technology
Pu'er University
Contact:{daojiaxu, lizhuoying}@peu.edu.cn

Abstract

To assess homograph appropriateness in narrative contexts for SemEval-2026 Task 5, we propose a contrastive regression framework. This approach combines candidate sense definitions with full narrative texts to establish an MSE regression baseline, further enhanced by a contextual grouping ranking loss that models relative rationality among senses. Evaluated on the official AmbiStory dataset, our method consistently outperforms the baseline in accuracy and Spearman correlation. These results validate the efficacy of relative order modeling for capturing fine-grained semantic nuances in complex narratives. The code is available at: https://github.com/daojiaxu/Semeval2026_task5.

1 Introduction

Word Sense Disambiguation, as the core task of natural language understanding, has always faced challenges such as fuzzy semantic boundaries and strong context dependence. Traditional tasks assume a "unique correct solution", which differs from the coexistence of polysemy in real language scenarios. SemEval-2026 Task 5 (Gehring et al., 2026)focuses on the assessment of the appropriateness of homographs in narrative contexts, requiring models to provide a 1-5 point rationality judgment for candidate senses, which is more aligned with human language cognition habits.

For this task, this study proposes a modeling method based on contrastive regression: first, the specific meanings of candidate senses are combined with complete narrative texts as model inputs, and the semantic matching relationship is directly modeled through the MSE regression baseline; further, a contextual grouping and contrastive sorting strategy is introduced to construct relative rationality constraints between senses within the same narrative framework, in order to enhance the model's ability to distinguish subtle semantic differences. The experiment is conducted based on the official AmbiStory dataset, aiming to verify the added value of relative order modeling to the task and provide new ideas for semantic evaluation in complex contexts.

2 Related Work

Pre-trained language models (PLMs) such as BERT, RoBERTa, and DeBERTa have become mainstream solutions for semantic understanding and evaluation tasks (Devlin et al., 2019; Liu et al., 2019; He et al., 2020). The design of their training objectives is crucial for task performance. Some studies have achieved dual optimization of numerical fitting accuracy and relative order consistency by introducing margin ranking loss within a regression framework. Word sense disambiguation (WSD), as one of the core tasks of semantic understanding, is considered a long-standing challenge in AI due to the inherent ambiguity of word senses. Traditional methods are divided into three categories: supervised, unsupervised, and knowledge-based, and there is still significant room for improvement in overall accuracy (Navigli, 2009). In recent years, large language models (LLMs) have brought new breakthroughs to the task of word sense disambiguation. Sumanathilaka et al. combined prompt enhancement mechanisms with knowledge bases and evaluated on the FEWS dataset, finding that models such as GPT-4 Turbo performed excellently. Prompt optimization can improve disambiguation accuracy in complex scenarios (Sumanathilaka et al., 2024). Kaskov et al. focused on the issue of homonym duplication in diffusion models and proposed an LLM-based prompt word expansion method, effectively reducing duplication rates and semantic ambiguity (Kaskov et al., 2025).

In terms of evaluating the ambiguity comprehen-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 284–289

sion ability of LLMs, Keluskar et al. found that off-the-shelf LLMs are prone to misunderstandings and hallucinations when dealing with ambiguity issues, but adding context and other untrained disambiguation methods can effectively improve performance (Keluskar et al., 2024). Cross-domain research has also provided new ideas, such as Tang et al. proposing a ranking-based contrastive loss function (RCL) to optimize recommendation systems (Tang et al., 2023), and Setitra et al. integrating deep models with LLMs to enhance visual polysemy and word sense disambiguation performance (Setitra et al., 2025). Large language models are prone to hallucinations due to ambiguous understanding biases. The LI framework proposed by Oxford University can accurately detect such hallucinations and assess the reliability of model outputs by tracking the flow of information at each layer(Kim et al., 2025).

Contrastive learning is a self-supervised representation learning method that trains models by distinguishing between similar and dissimilar samples, and has demonstrated excellent performance across multiple domains(Hu et al., 2024). Zhang Jing and others proposed the LA-UCL framework, which introduces a large language model to enhance contrastive learning for text classification with limited samples. Through self-enhancement and external enhancement modules, it improves discrimination and mitigates overfitting, achieving experimental results superior to the baseline model (Zhang et al., 2024). Xu Xiaodan and others proposed the CodeGPTSensor model, which is based on a contrastive learning framework and UniXcoder semantic encoder, and outperforms existing baselines (Xu et al., 2025).

Overall, pre-trained language models have become a core driving force for semantic understanding and evaluation tasks, demonstrating great potential in scenarios such as word sense disambiguation and ambiguous question answering. Future research can further explore directions such as cross-domain knowledge fusion and efficient prompt learning to enhance the model's understanding and processing capabilities in complex semantic scenarios.

3 Task and Dataset

SemEval 2026 Task 5 focuses on the semantic rationality scoring of homographs in narrative contexts, requiring the model to provide a relevance judgment on a scale of 1-5 for candidate senses

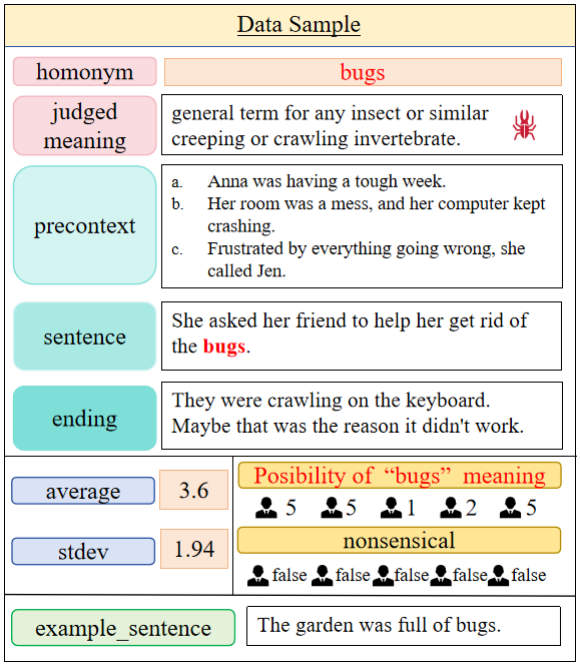

Figure 1: data sample

of the target word. Unlike traditional word sense disambiguation tasks that assume a "unique correct solution", this task is more closely aligned with real-life language scenarios, allowing for multiple reasonable interpretations of word senses within the same context.

The experiment utilizes the official AmbiStory dataset, which comprises five short stories consisting of three background sentences, one ambiguous sentence, and an optional ending. The data sample is shown in Figure 1. The core fields of each sample include:

a) Description of candidate senses (judged_meaning), clarifying the specific meaning of the target word being evaluated.

b) Narrative context fields (precontext, sentence, ending) jointly constitute a complete narrative logic.

c) For annotation data, the training/dev set contains the scoring lists (choices) and their average scores (average) from at least 5 annotators, while the test set does not contain annotation information.

To directly model the matching relationship between candidate senses and contexts, we construct the input as a sentence pair: (judged_meaning,

full_context), where full_context is the concatenated text of precontext+sentence+ending, and the maximum input length is set to 512. This design allows the model to simultaneously capture the semantic features of senses and contextual information, enabling precise judgment of the rationality of senses in a specific narrative.

4 Methods

4.1 MSE regression baseline

The baseline treats SemEval 2026 Task 5 as a regression problem, adopting the classic architecture of attaching a regression head on top of a pretrained encoder. During the training phase, the mean squared error (MSE) is used as the loss function to fit the average of the manually annotated scores for the samples. The input is constructed as a sentence pair combining the specific meaning of the candidate senses with the complete narrative text. After encoding with pretrained models such as DeBERTa-v3-large, a single output regression layer is used to predict continuous scores, while monitoring indicators such as Spearman correlation coefficient and accuracy to evaluate model performance.

In the inference stage, the continuous prediction values output by the model need to be converted into discrete scores that meet the requirements. The specific conversion formula is as follows:

a) Round to the nearest whole number:

$$\hat{y}_{\text{round}} = \text{round}(\hat{y}) \tag{1}$$

b) Truncate to the interval [1,5]

$$\hat{y}_{\text{final}} = \text{clip}(\hat{y}_{\text{round}}, 1, 5) \tag{2}$$

The final result is output in JSONL format, with each line containing the sample ID and a prediction result ranging from 1 to 5, in the format of {"id": "...", "prediction": 1..5 } . This method directly models the degree of matching between senses and contexts, fits the overall trend of human subjective scoring through a regression task, and provides a basic benchmark model for the task.

4.2 Contextual grouping, comparison, and sorting enhancement

In response to the characteristic of multiple candidate senses corresponding to the same narrative context in tasks, we propose a context grouping and comparative sorting enhancement strategy. Samples sharing the same precontext+sentence+ending are divided into the same context group. During training, paired samples are sampled from within the same group to construct sorting constraints, guiding the model to learn the relative rationality relationships between senses.

Loss function design: Let the model outputs of two samples within the same group be denoted as s_i and s_j, and their corresponding average human annotations as y_i and y_j. Define the ranking objective as $t = \text{sign}(y_i - y_j) \in \{+1, -1\}$, where $t = +1$ if $y_i > y_j$ and $t = -1$ if $y_i < y_j$. Note that when $y_i = y_j$, the sample pair is masked out and does not contribute to the ranking loss, which avoids degenerate gradients that would arise from a zero-valued target. The margin-ranking loss is employed:

$$L_{\text{rank}} = \max(0, m - t \cdot (s_i - s_j)) \tag{3}$$

Where m represents the margin hyperparameter. The final loss function is a weighted combination of regression loss and ranking loss:

$$L = L_{\text{mse}} + \lambda \cdot L_{\text{rank}} \tag{4}$$

Where λ represents the contrastive loss weight (contrastive_weight). In implementation, mask processing is applied to paired samples where $y_i = y_j$, since such pairs yield $t = 0$ (i.e., no valid ranking direction) and would produce degenerate zero gradients. Excluding these pairs avoids introducing invalid ranking signals and effectively enhances the model's fine-grained differentiation ability for the rationality of senses.

5 Experiment Setup

5.1 Training configurations

This section reports the key configurations that are consistent with the warehouse script (see the code link for specific implementation). All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. The training configuration is shown in Table 1.

The contrastive learning experiment consistently employs DeBERTa-v3-large as the base model to ensure fairness and comparability with the baseline experiment. During training, intermediate checkpoints are automatically cleaned up to conserve storage space. The final prediction results are saved in JSONL format, and evaluation metrics are calculated using the official scoring.py script.

Configuration item	Parameter value
Maximum input length	512
Optimizer and learning rate	AdamW,$2e-5$,weight decay 0.01
Batch size and gradient accumulation	per-device batch size 4,gradient accumulation 4
Mixed precision	fp16 (automatically enabled when GPU is available)
Training epochs:	8 epochs
Evaluation strategy:	Evaluate every 0.5 epochs and save the optimal checkpoint
Model selection basis	Development set Accuracy (metric_for_best_model)
Logging system	SwanLab (records training process and evaluation metrics)

Table 1: Training configurations

5.2 Contrastive learning model structure

Our model builds on DeBERTa-v3-large as a shared encoder with a dual-input branch structure. At the input layer, samples sharing the same narrative context (precontext, ambiguous sentence, and optional ending) are grouped, and contrastive pairs are formed from different word-sense candidates within each group, creating "sense description + full context" input pairs. These pairs are encoded by the shared DeBERTa encoder, whose disentangled attention captures deep sense–context associations and produces fixed-dimensional representations after pooling. The output layer applies a regression head to predict continuous appropriateness scores, and optimizes via margin-ranking contrastive loss on paired-sample score differences, guiding the model to distinguish fine-grained sense appropriateness under the same context. This shared-encoder design enables efficient cross-sample transfer by unifying contrastive learning with regression. The architecture is illustrated in Figure 2.

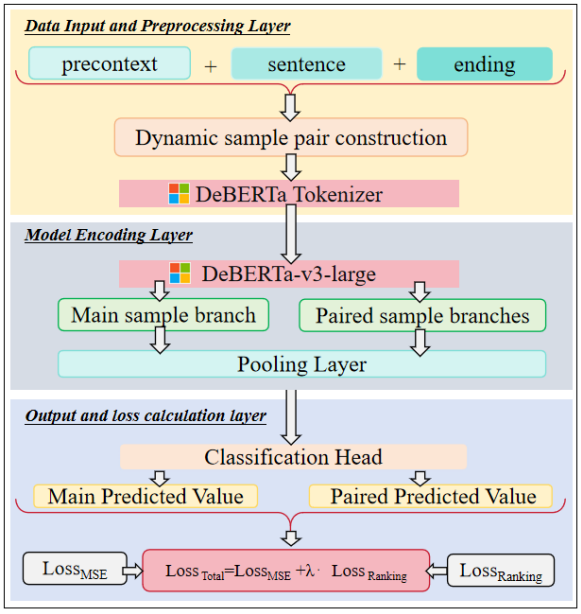

Figure 2: Model Structure

6 Results

Our system achieved 20th place in the official SemEval-2026 Task 5 leaderboard. The following subsections detail the experimental results and analysis.

6.1 Baseline results

The experimental results show that DeBERTa-v3-large significantly outperforms other baseline models in both accuracy (70.75%) and Spearman correlation coefficient (0.5934), demonstrating stronger semantic understanding and ranking capabilities. The results are shown in Table 2. Our project's performance on the test set is as follows: Spearman Correlation 0.533 and Accuracy 0.688.

Model Name	Accuracy	Spearman
deberta-v3-large	0.7075	0.5934
roberta-base	0.6327	0.4285
bert-base-uncased	0.6241	0.3581
deberta-v3-base	0.6224	0.3616
roberta-large	0.5748	0.0254
bert-large-uncased	0.5272	-

Table 2: Performance Comparison of Baseline Models

6.2 Hyperparameter tuning for contrastive sorting

We perform a grid search over two key hyperparameters: margin $\in \{0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 1.5\}$ and contrastive weight $\in \{0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 1.5\}$. Table 3 shows the two best configurations on the development set.

A large margin (1.5) paired with a small con-

Optimization objective	Optimal configuration	Accuracy	Spearman	Baseline improvement
Accuracy	margin=1.5, weight=0.2	0.7228	0.5854	+1.53%
Sorting ability	margin=1.5, weight=0.1	0.7211	0.6144	+3.54%

Table 3: Optimal configuration

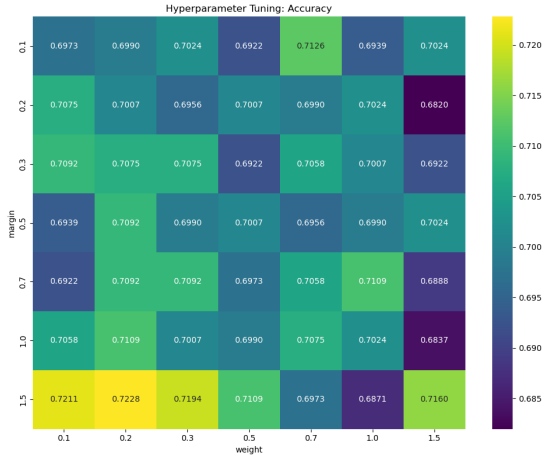

Figure 3: tune heatmap accuracy

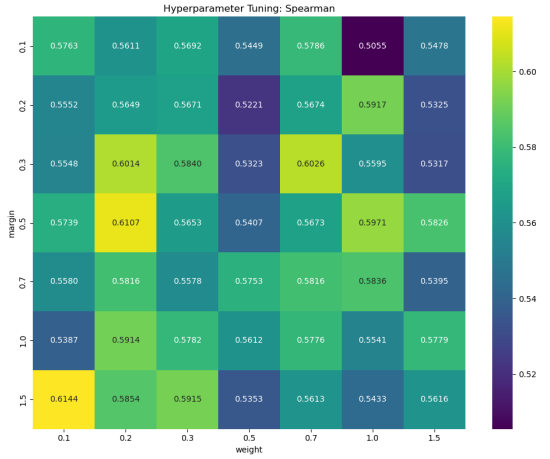

Figure 4: tune heatmap spearman

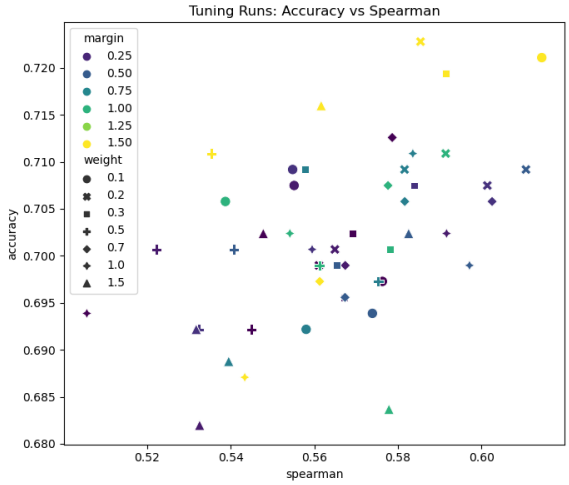

Figure 5: tune scatter accuracy VS spearman

trastive weight (0.1–0.2) yields the best results, improving accuracy by 1.53% and Spearman correlation by 3.54% over the DeBERTa-v3-large baseline. The full tuning landscape is visualized in Figures 3–5.

7 Discussion and Analysis

This study validates the efficacy of modeling semantic matching by pairing sense descriptions with narrative contexts. While DeBERTa-v3-large establishes a robust performance benchmark, our proposed contextual grouping contrastive loss further yields consistent improvements in both accuracy and Spearman correlation. This highlights the value of relative order modeling in mitigating discretization errors within shared narrative frameworks.

The divergence in optimal parameters for accuracy versus Spearman correlation underscores the task's complexity: the former prioritizes absolute point-to-point matching, while the latter emphasizes global ranking consistency. Future research should explore multi-objective optimization to better align loss functions with these dual requirements. Additionally, integrating LLM-based commonsense reasoning and prompt engineering remains a promising avenue for enhancing semantic evaluation in complex narrative scenarios.

Acknowledgments

This work has been supported by the Special Basic Cooperative Research Programs of Yunnan Provincial Undergraduate Universities Association (grant NO. 202401BA070001-049). The authors would like to thank the anonymous reviewers for their constructive comments.

References

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.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. 2024. A comprehensive survey on contrastive learning. *Neurocomputing*, page 128645.

Evgeny Kaskov, Elizaveta Petrova, Petr Surovtsev, Anna Kostikova, Ilya Mistiurin, Alexander Kapitanov, and Alexander Nagaev. 2025. Un-doubling diffusion: Llm-guided disambiguation of homonym duplication. *arXiv preprint arXiv:2509.21262*.

Aryan Keluskar, Amrita Bhattacharjee, and Huan Liu. 2024. Do llms understand ambiguity in text? a case study in open-world question answering. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7485–7490. IEEE.

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. 2025. Detecting llm hallucination through layer-wise information deficiency analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32310–32322. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Insaf Setitra, Praboda Rajapaksha, Aung Kaung Myat, and Noel Crespi. 2025. Leveraging ensemble deep models and llm for visual polysemy and word sense disambiguation. *Multimedia Tools and Applications*, pages 1–33.

Deshan Koshala Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024. Can llms assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 97–108.

Hao Tang, Guoshuai Zhao, Yujiao He, Yuxia Wu, and Xueming Qian. 2023. Ranking-based contrastive loss for recommendation systems. *Knowledge-Based Systems*, 261:110180.

Xiaodan Xu, Chao Ni, Xinrong Guo, Shaoxuan Liu, Xiaoya Wang, Kui Liu, and Xiaohu Yang. 2025. Distinguishing llm-generated from human-written code by contrastive learning. *ACM Transactions on Software Engineering and Methodology*, 34(4):1–31.

Jing Zhang, Hui Gao, Peng Zhang, Boda Feng, Wenmin Deng, and Yuexian Hou. 2024. La-ucl: Llm-augmented unsupervised contrastive learning framework for few-shot text classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10198–10207, Torino, Italia. ELRA and ICCL.