Rating the Plausibility of Word Senses in Ambiguous Stories SemEval-2026 Task 5: System Description

Fomin Bogdan Isac Lucian

Abstract

We describe a Transformer-based system for SemEval-2026 Task 5 (Rating Plausibility of Word Senses in Ambiguous Stories through Narrative Understanding). Our submission fine-tunes **microsoft/deberta-v3-large** on paired inputs consisting of (i) the full story context with the ambiguous target explicitly marked and (ii) a structured sense description. We train with a mixture of hard-label and soft-label cross-entropy to reflect annotator uncertainty, and decode predictions using an expected-value strategy. The resulting system achieves about **0.77** accuracy on the development set and **0.72** on the test set. We additionally provide a comparison against a constant-rating baseline (about **0.52** accuracy) and report that smaller Transformer backbones such as DistilBERT or RoBERTa-base underperform (about **60–65%**).

1 Introduction

Lexical ambiguity in narratives is often resolved by global coherence, discourse cues, and event knowledge rather than local word context. SemEval-2026 Task 5 targets this challenge by asking systems to assign plausibility ratings to word senses in ambiguous stories. We approach the task as supervised 5-way classification with a strong pretrained Transformer backbone and an input formulation designed to explicitly connect story-level context to a candidate sense.

2 Task

Given a story containing an ambiguous homonym and a candidate sense, the goal is to predict a plausibility rating in $\{1, 2, 3, 4, 5\}$. The official evaluation for this task is based on classification performance (reported here as accuracy).

3 Baseline

We compare our system against a simple constant-rating baseline (provided by the organizers):

- **Dev baseline:** predict label **3** for all development examples.
- **Test baseline:** predict label **4** for all test examples.

This baseline achieves approximately **0.52** accuracy and serves as a sanity check for label skew and dataset difficulty.

4 Proposed Method

Our implementation trains a paired-sequence classifier with DeBERTa-v3-large.

4.1 Input Construction

Each instance is converted into a pair (A, B):

- **Text A (story):** we concatenate the narrative segments (prefix context, target sentence, suffix context). We then mark the *first* occurrence of the ambiguous homonym using special markers:

  `<t> homonym </t>`

  This encourages the model to attend to the intended ambiguity location.

- **Text B (sense):** we build a structured sense string containing the homonym, the judged meaning, an example sentence, and an annotation-derived nonsensicality rate (fraction of "nonsensical" votes). Concretely, the notebook constructs:

  `homonym = meaning. Example: ... Nonsense_votes: r`

We tokenize (A, B) as a standard sentence-pair input. To preserve the sense specification, we apply truncation to *only the first* sequence (the story) while keeping the sense string intact.

4.2 Model

We fine-tune **microsoft/deberta-v3-large** as a 5-class sequence classifier (softmax over 5 logits). Since we introduce the new markers `<t>` and `</t>`, the tokenizer vocabulary is extended and the model embeddings are resized accordingly.

Given tokenized input (A, B), the model outputs logits $\mathbf{s} \in R^5$ and probabilities

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{s}).$$

4.3 Soft Labels and Training Loss

The dataset provides multiple human choices per example. We convert these into a **soft label distribution** $\mathbf{y} \in R^5$ by normalizing vote counts.

The notebook optimizes a **mixture** of:

- **Hard-label cross-entropy:** using the argmax of the soft distribution as the hard label.
- **Soft-label cross-entropy:** cross-entropy between $\hat{\mathbf{y}}$ and $\mathbf{y}$.

Let $y^\star = \arg\max_i y_i$. The loss used in the implementation is:

$$\mathcal{L} = (1-p) \cdot \mathrm{CE}(\mathbf{s}, y^\star) \; + \; p \cdot \Big(- \sum_{i=1}^{5} y_i \log \hat{y}_i \Big),$$

with a tuned mixing coefficient p (in the notebook $p \approx 0.105$). This balances stable hard-label optimization with the uncertainty captured by soft targets.

4.4 Decoding

At inference time, we decode predictions using **expected-value (EV) decoding**:

$$\mathrm{EV} = \sum_{i=1}^{5} i \cdot \hat{y}_i,$$

then map EV to an integer rating in $\{1, \ldots, 5\}$ by thresholding at 1.5, 2.5, 3.5, and 4.5.

4.5 Training Details

Key hyperparameters (as set in the notebook) include:

- Max sequence length: 320 (with truncation on story only)
- Learning rate: 6.6053×10^{-6}
- Epochs: 8
- Batch size: 8
- Weight decay: 0.0387
- Warmup ratio: 0.0795
- Scheduler: cosine
- Optimizer: AdamW
- Mixed precision: FP16 when CUDA is available

5 Results

Our system performance (as reported from development experiments and post-submission test feedback) is:

- **Development accuracy:** ≈ 0.77
- **Test accuracy:** ≈ 0.72

6 Baseline vs. Proposed System

The table compares the constant-rating baseline to our full system.

System	Dev Accuracy	Test Accuracy
Constant-rating baseline (Dev=3, Test=4)	0.52	0.52
Proposed (DeBERTa-v3-large + soft labels + EV decoding)	**0.77**	**0.72**

7 Backbone Size Comparison

In additional experiments, we replaced the main backbone with smaller pretrained models while keeping the same paired-input formulation and training pipeline. We observed that smaller models such as **DistilBERT** or **RoBERTa-base** achieve notably worse performance, around **60–65%** accuracy on the development set, indicating that narrative plausibility assessment benefits from higher-capacity contextual representations.

8 Conclusion

We presented a DeBERTa-v3-large based paired-input classifier for SemEval-2026 Task 5. Key components include explicit target marking in the story, a structured sense description as the paired sequence, soft-label training combined with hard-label supervision, and expected-value decoding. The approach substantially improves over a constant-rating baseline (0.52) and achieves around 0.77 dev / 0.72 test accuracy, while smaller backbones degrade to roughly 60–65%.