Narrative Team at SemEval-2026 Task 5: Rating Plausibility of Word Senses in Ambiguous Sentences through Narrative Understanding

Valentin Istrate, Mocanu Octavian
Faculty of Computer Science
Alexandru Ioan Cuza University
(valentin.istrate, andrei.mocanu)
@student.uaic.ro

Tatiana Khaidukova
Information Technologies and Programming Faculty
ITMO University
467904@niuitmo.ru

Abstract

This paper describes our system for SemEval-2026 Task 5, which focuses on predicting the plausibility of word senses in ambiguous narrative contexts. The task requires assigning a real-valued plausibility score to short narrative texts based on aggregated human judgments. Our approach compares two modeling paradigms: (i) a pretrained transformer-based regression model using DistilBERT fine-tuned on the task data, and (ii) a lightweight neural baseline based on a bidirectional LSTM trained either from scratch or initialized with GloVe embeddings. Input representations combine a candidate sense definition with narrative context and the target sentence, separated by a special token. On the official test set, the DistilBERT model achieves the strongest Acc@SD score of 0.54, while the best BiLSTM configuration reaches 0.52. Although transformer-based models remain the strongest option in our experiments, the recurrent baseline remains competitive under the tolerance-based metric. We discuss model variants, reproducibility details, and limitations of our analysis.

1 Introduction

Narrative understanding requires more than identifying a single correct interpretation: human readers often tolerate ambiguity and assign graded plausibility to alternative word senses depending on context, underspecification, and personal expectations. SemEval-2026 Task 5 targets this phenomenon by framing word-sense plausibility as a scalar prediction problem: systems must estimate how plausible a candidate sense of an ambiguous homonym is within a short story context (Gehring and Roth, 2026).

Pretrained transformer encoders fine-tuned for regression provide a strong starting point for semantic scoring tasks because they produce context-sensitive representations of the full input sequence (Devlin et al., 2019). Distilled transformer variants offer a favorable trade-off between compute and performance, often retaining much of the representational power of larger encoders at lower cost (Sanh et al., 2019). In addition to a transformer solution, we explore a simpler recurrent baseline based on a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to better understand how much of the task can be captured by sequence modeling without large-scale contextual pretraining.

Our paper makes three practical contributions: (1) we use a compact input representation that places the candidate sense definition next to the story context; (2) we compare a DistilBERT regressor and BiLSTM regressors under the shared-task evaluation setting; and (3) we report recurrent-model ablations for embedding initialization and model size, while explicitly separating confirmed official test results from validation-only observations.

2 Background

2.1 Task Description

SemEval-2026 Task 5 introduces AmbiStory, a dataset of five-sentence short stories designed to probe narrative-based word sense plausibility (Gehring and Roth, 2026). Each story setup contains: (1) a precontext of three sentences grounding the scenario; (2) an ambiguous sentence containing a target homonym with two distinct candidate senses; and (3) an optional ending that may bias the interpretation toward one sense.

For each story and each candidate sense, human annotators rate plausibility on a 1–5 scale. Training examples include aggregated human ratings, while evaluation requires systems to predict plausibility without access to the gold ratings.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1089–1093

2.2 Evaluation Metrics

The shared task uses two primary evaluation measures:

- **Spearman correlation (ρ)**: measures rank correlation between predicted scores and average human ratings.
- **Accuracy within standard deviation (Acc@SD)**: the proportion of predictions whose absolute error is within the annotator-rating standard deviation for that sample, reflecting agreement sensitivity.

These metrics reward complementary behaviors: Spearman correlation prioritizes correct relative ordering, while Acc@SD rewards predictions consistent with annotator consensus and allows wider tolerance when human judgments vary.

2.3 Related Work

Word-sense disambiguation is often formulated as selecting one correct sense, but SemEval-2026 Task 5 instead evaluates graded plausibility judgments for candidate senses in context (Gehring and Roth, 2026). Contextual encoders such as BERT have become common backbones for semantic classification and regression because self-attention allows each token representation to condition on the full sequence (Devlin et al., 2019). DistilBERT compresses BERT through knowledge distillation, making it attractive when computational cost matters (Sanh et al., 2019).

Recurrent neural networks, including LSTMs, remain useful baselines for sequential text modeling (Hochreiter and Schmidhuber, 1997). Static pretrained word vectors such as GloVe provide lexical prior knowledge (Pennington et al., 2014), but unlike transformer embeddings they do not change with sentential context. This distinction is important for homonym plausibility, where the same surface form may support different interpretations depending on the surrounding narrative.

3 System Overview

Our system architecture is shown in Figure 1. Both models consume the same input text representation and output a single plausibility score.

3.1 Input Representation

Each labeled example provides (a) a candidate sense definition and (b) a narrative context comprising the precontext, ambiguous sentence, and optionally an ending. We convert each example to a single sequence:

```
Definition: judged_meaning [SEP] Story:
precontext + sentence + ending
```

For example, for a story containing the ambiguous word *track*, the definition segment may describe either a railway-related sense or an evidence-related sense, and the story segment provides the narrative context used to judge which sense is plausible.

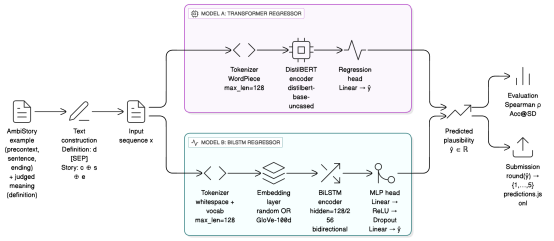

Figure 1: Architecture for plausibility regression.

3.2 Transformer regressor: DistilBERT

Our primary submission fine-tunes `distilbert-base-uncased` for regression using the Hugging Face Transformers library (Wolf et al., 2020). We attach a sequence-classification head with `num_labels = 1`. WordPiece tokenization, truncation, and padding are applied with a maximum sequence length of 128 tokens. The model produces a single scalar prediction from the regression head.

3.3 Recurrent baseline: BiLSTM regressor

To compare against a pretrained contextual encoder, we implement a word-level BiLSTM model in PyTorch (Paszke et al., 2019). The model maps tokens to embeddings, applies a bidirectional LSTM, and uses a feed-forward head to predict a scalar score. We experiment with two embedding initializations: randomly initialized trainable embeddings and GloVe-initialized embeddings fine-tuned during training (Pennington et al., 2014).

The BiLSTM produces final forward and backward hidden states, which are concatenated into h:

$$h = [h_T^{\rightarrow}; h_1^{\leftarrow}]. \quad (1)$$

We then apply an MLP with ReLU and dropout and map to a scalar plausibility score $\hat{y}$:

$$\hat{y} = W_2 \,\mathrm{Dropout}(\mathrm{ReLU}(W_1 h + b_1)) + b_2. \quad (2)$$

4 Experimental Setup

We follow the task-provided training and development data. For model selection, we concatenate the provided train and development sets into a single pool and perform an 85/15 train–validation split with random seed 42. We use this internal split to compare model variants under a consistent validation protocol before producing official test predictions. This choice increases the amount of data available for training while still preserving a held-out portion for local diagnostics; however, it also means our validation numbers are not directly comparable to the original development-set split.

Each training instance includes:

- **label** (float): the average plausibility rating;
- **stdev** (float): the standard deviation across annotators, used for Acc@SD computation.

4.1 Training Details

We fine-tune DistilBERT using mean squared error loss through the regression objective in Transformers. The hyperparameters are listed in Table 1.

Parameter	Value
Model	`distilbert-base-uncased`
Loss	MSE/regression loss
Optimizer	AdamW
Epochs	3
Batch size	16
Learning rate	2×10^{-5}
Weight decay	0.01
Max sequence length	128
Random seed	42

Table 1: DistilBERT hyperparameters.

During evaluation, we compute Spearman (ρ) between predicted scores and gold averages, and Acc@SD using the stored per-sample standard deviation.

Table 2 shows the hyperparameters of BiLSTM training. We use a simple word tokenizer: lowercase text, remove the literal `[SEP]` marker, and split on whitespace. We build a vocabulary from the training corpus and keep tokens with frequency greater than 1; remaining tokens map to `<UNK>`. Inputs are padded or truncated to 128 tokens. This preprocessing may discard rare lexical cues, which is a limitation for word-sense plausibility tasks.

The DistilBERT and BiLSTM epoch counts differ because they serve different optimization regimes: DistilBERT starts from a large pretrained encoder and can overfit quickly on small shared-task data, while the BiLSTM models require more epochs to learn useful representations from randomly initialized or static embeddings.

Parameter	Value
Loss	MSE
Optimizer	Adam
Learning rate	1×10^{-3}
Epochs	50
Dropout	0.3
Embedding options	Random, GloVe
Max sequence length	128
Random seed	42

Table 2: BiLSTM hyperparameters.

For GloVe variants, we initialize embedding rows with pretrained vectors where available and sample random vectors for missing tokens. For submission, model outputs are rounded to the nearest integer rating, following the official scoring format used in our submitted runs.

Code Availability. Our implementation is available at: https://github.com/t-v-khaidukova/NarrativeTeam5.

5 Results

Table 3 reports the official testing-phase results for our submitted runs.

System	Acc@SD	Spearman ρ
DistilBERT regressor	**0.54**	**0.17**
BiLSTM, random embeddings	0.52	0.02
BiLSTM, GloVe-initialized	0.48	0.02

Table 3: Official testing-phase results.

DistilBERT achieved the strongest official result among our submissions, with 0.54 Acc@SD, 0.17 Spearman correlation, and a combined average score of 0.36. The random-embedding BiLSTM reached 0.52 Acc@SD and 0.02 Spearman, while the GloVe-initialized BiLSTM reached 0.48 Acc@SD and 0.02 Spearman.

Although DistilBERT obtained the best Acc@SD and Spearman score among our submissions, the low Spearman values indicate that the systems were better at producing tolerance-compatible plausibility scores than at preserving the global ranking of examples by human-rated plausibility.

5.1 Ablation and Validation Analysis

Table 4 summarizes the recurrent-model ablations.

Emb.	Hid.	Dense	Acc@SD	ρ
Random	128	0	0.6163	0.5385
Random	128	1	0.6070	0.5373
Random	128	2	0.5930	0.5400
Random	256	0	0.6000	0.5421
Random	256	1	0.6116	**0.5601**
GloVe	128	0	**0.6256**	0.4714
GloVe	128	1	0.5837	0.4722
GloVe	256	0	0.6140	0.5124
GloVe	256	1	0.6209	0.4792

Table 4: BiLSTM validation ablations on the internal 85/15 split.

The validation ablations show that increasing hidden size from 128 to 256 was more useful than adding extra dense layers in the random-embedding setting. GloVe initialization produced the highest validation Acc@SD in one configuration, but it did not improve the official test Acc@SD. This pattern may reflect a mismatch between static lexical embeddings and context-sensitive plausibility judgments.

5.2 Error Analysis

We observed three recurring sources of difficulty. First, examples with an ending sentence can shift plausibility toward one sense, and models may underweight this late contextual cue when the precontext supports a different interpretation. Second, cases with high annotator disagreement are inherently difficult: Acc@SD is more permissive for such items, but Spearman still penalizes poor ranking. Third, the BiLSTM preprocessing maps rare words to <UNK>, which can remove important clues for a target homonym or its surrounding context.

Because this analysis is based on internal inspection rather than a separately annotated error taxonomy, we treat it as a diagnostic discussion rather than a complete quantitative error taxonomy.

5.3 Compute and Reproducibility

DistilBERT requires more computation per update than the BiLSTM baseline, but it converges in fewer epochs and uses contextual representations that are better suited to ambiguity. The BiLSTM is cheaper and easier to train, making it useful as a baseline and as a sanity check for whether the input representation alone carries useful signal.

6 Conclusion

We present systems for SemEval-2026 Task 5, comparing a DistilBERT regressor against BiLSTM baselines with different embedding initializations and model sizes. DistilBERT achieved the best official testing-phase Acc@SD (0.54), while our strongest BiLSTM variant reached 0.52. These results suggest that pretrained contextual encoders are helpful for graded word-sense plausibility, but simpler recurrent models can remain competitive under tolerance-based evaluation.

Future work includes pairwise modeling across candidate senses, improved handling of ending-conditioned shifts in plausibility, stronger pretrained encoders, and more detailed error analysis with validation curves.

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 NAACL-HLT*, pages 4171–4186.

Janosch Gehring and Michael Roth. 2026. SemEval-2026 Task 5: Rating plausibility of word senses in ambiguous sentences through narrative understanding. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Shared task description.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas K"opf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

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 EMNLP: System Demonstrations*, pages 38–45.