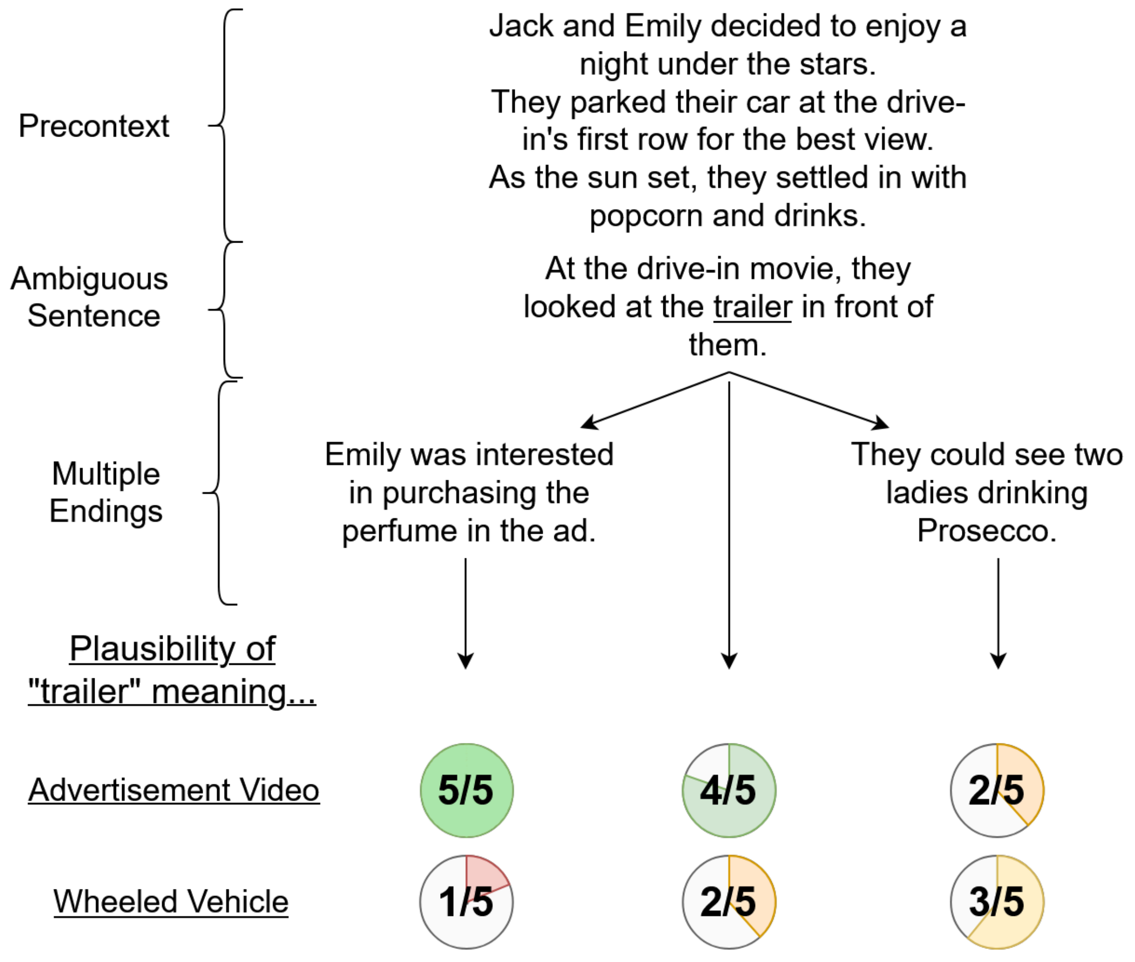
UTN
University of Technology Nuremberg

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

Overview


Precontext
Jack and Emily decided to enjoy a night under the stars. They parked their car at the drive-in's first row for the best view. As the sun set, they settled in with popcorn and drinks.
Ambiguous Sentence
At the drive-in movie, they looked at the trailer in front of them.
Multiple Endings
Emily was interested in purchasing the perfume in the ad.
They could see two ladies drinking Prosecco.
Plausibility of "trailer" meaning...
Advertisement Video
5/5
4/5
2/5
Wheeled Vehicle
1/5
2/5
3/5

Word Sense Disambiguation tasks commonly assume one word sense to be the 'correct' one, but that is not necessarily reflective of reality. Ambiguities, underspecification and personal opinions can influence which word senses one finds plausible in a given context, and there is a difference between the intuition of humans and the predictions of computational models. To study and benchmark this phenomenon, we introduce the **AmbiStory** dataset, a dataset of 5-sentence short stories. The task is to disambiguate a target homonym in the fourth sentence through contextual clues in surrounding sentences.

Our stories consist of three parts: A *precontext*, consisting of three sentences that ground the story, an *ambiguous sentence*, containing a homonym that causes it to have two widely different plausible interpretations, and optionally one of two *endings*, which often imply a specific word sense of the homonym.

We asked Prolific participants to rate the plausibility of a given word sense in the context of a story on a scale from 1 to 5. We collect at least five ratings for each sense/story sample. Since each story setup has either one of two endings or none at all, and we focus on two word senses per homonym, we obtain six annotation samples per setup.

Task and Metrics

In essence, the task is to predict the human-perceived plausibility of a word sense by selecting a score between 1 and 5. We evaluate this using two primary metrics:

- **Spearman Correlation**: The predicted plausibility score should ideally correlate to the average of scores assigned by humans.
- **Accuracy Within Standard Deviation**: Some samples have a clearer consensus than others, so we calculate accuracy as the proportion of model predictions that are within standard deviation (at least 1) from the average judgment by annotators.

Sample Data

We provide some sample data to give you an idea of the task. The data consists of 5 stories/30 samples.

Download the sample data here!

Training Data

The training and validation sets are now available. Download them here!

The test set is not yet publically available. SemEval 2026 task evaluation period is scheduled for Jan 10th 2026 - Jan 31st 2026.