Graph Alignment Metric

CodeScientist (and Human Domain Expert)

February 26, 2025

Abstract

Domain Expert Note: The report generator consistently failed on this experiment, we beleive due to the truly large number of figures generated (one figure for each graph that it examined). Instead, we provide the automated summary generated by CodeScientist, and manually include several of the figures that it generated.

1 Automatically Generated Summary

The automatically generated summary of results produced by CodeScientist is shown in Table 1.

Field	Value
summary	This experiment tested different similarity metrics for aligning text descriptions with graph representations in TextWorldExpress cooking games. Three metrics were compared: a baseline word overlap ratio, Jaccard similarity, and a custom graph-text similarity measure that weighted nodes, spatial relations, and actions. The experiment was run in PILOT mode with 15 games and 10 episodes each, collecting 30 text-graph pairs for evaluation. The custom similarity metric (mean=0.317) significantly outperformed the baseline word overlap metric (mean=0.101) with p¡0.001 in bootstrap resampling tests. Interestingly, the Jaccard similarity performed identically to the baseline (mean=0.101, p=1.0). The experiment successfully implemented the core comparison of similarity metrics, though with some deviations from the original specification in terms of the number of games used (15 instead of 3). The results suggest that incorporating graph structure and relationship information through the custom metric provides better text-graph alignment than simple word overlap approaches.
summary (short)	Custom graph-aware similarity metric outperforms word overlap baselines for text-graph alignment in cooking games.
summary (medium)	A comparison of text-graph similarity metrics in TextWorldExpress cooking games found that a custom metric incorporating graph structure (mean=0.317) significantly outperformed both word overlap and Jaccard similarity baselines (mean=0.101) with p¡0.001. The experiment used 30 text-graph pairs from 15 games with 10 episodes each, demonstrating the value of considering graph relationships in text-graph alignment.
hypothesis	Graph-aware similarity metrics that incorporate structural relationships will perform better at text-graph alignment than simple word overlap methods.
hypothesis (operationalized)	A custom similarity metric that weights node matches (0.5), spatial relations (0.3), and action relations (0.2) will achieve higher similarity scores than word overlap ratio and Jaccard similarity when matching game state descriptions to their corresponding graph representations.
hypothesis (inference)	The results strongly support the hypothesis. The custom graph-aware metric achieved a significantly higher mean similarity score (0.317) compared to both baseline metrics (0.101) with p¡0.001 in bootstrap resampling tests. The consistent performance across 30 pairs and clear statistical significance suggests this is a reliable finding.
hypothesis (category)	support
faithfullness (details)	The experiment implemented the core comparison of similarity metrics as requested, but had some notable deviations: 1) Used 15 games instead of the specified 3 games in PILOT mode, 2) Did not implement all requested visualizations (confusion matrices, similarity distributions, progress correlation plots), 3) Successfully implemented the three similarity metrics and bootstrap statistical testing as specified. The deviations affect the scale of the experiment but not the fundamental validity of the comparison.
faithfullness (category)	deviations
interesting results	true

Table 1: Summary of Experimental Results

2 Selected Automatically Generated Figures

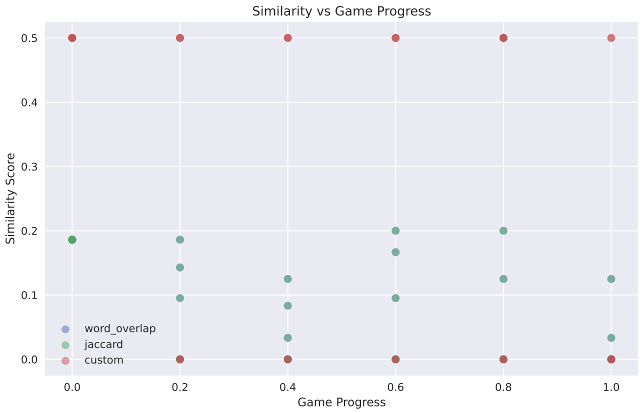

Figure 1: Progress Correlation

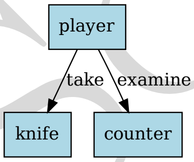

Figure 2: An example of one of the simpler (non-empty) graphs it appears to consider (Game 13, Episode 8, Step 15)

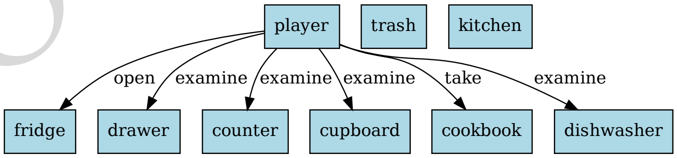

Figure 3: An example of the most complex graphs it appears to consider (Game 10, Episode 0, Step 0)