Explicit Semantic Modeling for Solving Word Association Puzzles

Aayush Pandey 2025201058

Srushti Pekamwar 2025201066

Suparshwa Patil 2025201090

Anurag Kacholiya 2025202025

Abstract

Word connection puzzles such as **New York Times Connections** require grouping a fixed set of words into subsets based on shared semantic, contextual, or structural relations. While large language models and embedding-based clustering methods show moderate success, prior analyses demonstrate that they fail systematically on puzzles involving polysemy, distractors, and form-based categories, largely due to their reliance on surface-level similarity and lack of explicit relational reasoning. Cognitive and psycholinguistic research suggests that human word grouping relies on structured semantic networks and constraint satisfaction rather than independent similarity judgments. This project investigates whether explicit semantic knowledge, combined with neural scoring and global optimization, can better model the reasoning required for such puzzles.

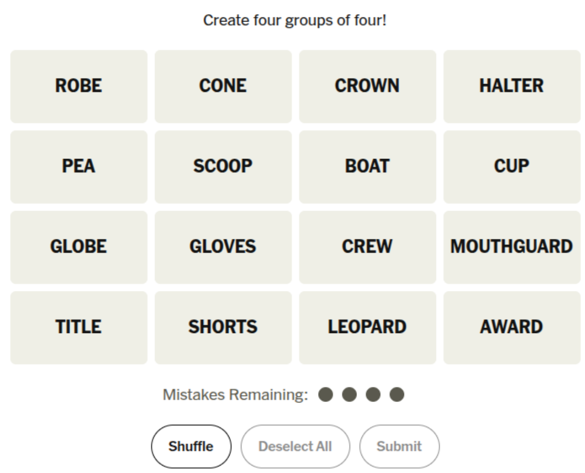
We propose a knowledge-guided neural framework that treats puzzle solving as a **group validity scoring problem** rather than a word-level classification task. Given a puzzle, the model evaluates candidate groups of four words using a hybrid representation that integrates semantic embeddings, curated knowledge graph relations (from WordNet, ConceptNet, WikiData and BabelNet), and form-based features capturing patterns such as shared affixes or lexical constructions. Training data is constructed by enumerating all candidate groups within each puzzle and supervising the model using gold groups, near-miss groups, and hard negatives. At inference time, a global optimization step selects a partition of groups that maximizes total validity while satisfying puzzle constraints. The project is structured in three parts: (i) establishing empirical baselines using embedding-based clustering, (ii) developing the knowledge-guided group scoring solver, and (iii) exploring controlled puzzle generation for stress-testing and analysis (Optional). This framework aims to improve robustness on adversarial categories while offering interpretable, human-aligned reasoning.

1 Introduction

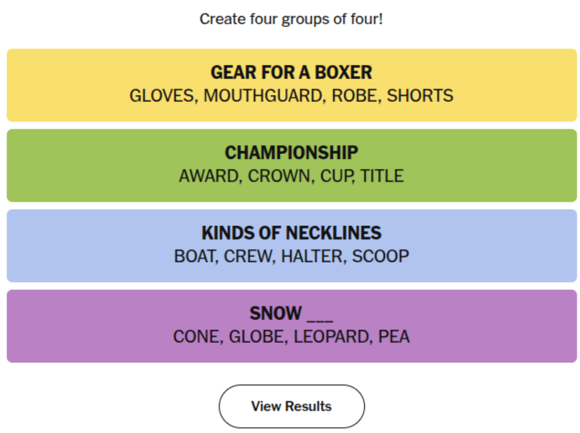
Word association and grouping tasks are fundamental to human language understanding. Games such as the *New York Times Connections* puzzle present a deceptively simple challenge: given a fixed set of sixteen words, the goal is to partition them into four groups of four words, where each group shares a latent relationship. These relationships may be semantic (e.g., animals), entity-based (e.g., fictional universes), or indirect and form-based (e.g., phrase continuations or orthographic patterns).

While embedding-based similarity methods and large language models (LLMs) demonstrate partial success on such tasks, they frequently fail on puzzles involving polysemy, distractors, named entities, and creative wordplay. Moreover, LLM-based approaches rely on opaque probabilistic inference, making it difficult to interpret or control their reasoning process.

Motivated by cognitive and psycholinguistic evidence that human word grouping relies on structured semantic networks and constraint satisfaction, this project explores an explicit semantic modeling approach. We propose a tiered framework that combines curated knowledge graphs, graph neural networks, and lightweight neural reasoning to solve word connection puzzles in an interpretable and data-efficient manner.



(fig 2.1) Example Puzzle.



(fig 2.2) Solution.

This project is structured into three progressive components.

- **Part I** focuses on establishing strong empirical baselines using embedding-based clustering and existing neural approaches to quantify the difficulty of the task.
- **Part II**, which constitutes the core contribution, develops a tiered knowledge-guided solver that integrates semantic graphs, graph neural networks, and lightweight neural reasoning to solve Connections puzzles.
- **Part III**(optional) explores the extension of controlled puzzle generation, investigating whether the learned structural constraints can be inverted to generate novel and challenging puzzles.

The staged design allows systematic evaluation while reserving generative modeling as a stretch objective contingent on solver performance.

2 Literature Review

The New York Times Connections word game has recently become both a popular human puzzle and a challenging benchmark for computational reasoning. An editorial in *The New York Times* acknowledges that players often outperform computers in solving Connections puzzles, attributing this gap to humans' ability to integrate semantic, contextual, and lateral knowledge quickly — a combination that current AI systems struggle to replicate (NYT Editorial, 2025).

Independent practitioners have also explored the task empirically. A detailed blog post documents early attempts at understanding and solving Connections puzzles programmatically. The author reviews puzzle structure, historical data collection, and initial solver strategies, concluding that straightforward prompting of large language models (LLMs) often fails on harder categories and that structured data is necessary for improvement (Solver Blog, 2024).

Scholarly work has begun to formalize these observations. **NYT-Connections** was introduced as an LLM benchmark, showing that even state-of-the-art models like GPT-4 fall far short of human performance when tested on deliberately isolated reasoning tasks (Lopez et al., 2024).

Connecting the Dots further analyzes the game with over 400 puzzles, showing that models solve only a small fraction and that multiple knowledge types — including encyclopedic, multiword, and form-based knowledge — are needed for success (Samadarshi et al., 2024).

Missed Connections frames the task as a measure of lateral thinking, highlighting how both embedding baselines and LLMs struggle without deeper relational reasoning (Todd et al., 2024).

Finally, work on **Puzzle Generation** demonstrates that models can create plausible Connections puzzles using hierarchical prompting, indicating the complexity and structure inherent in the puzzle domain (Merino et al., 2024).

Together, these sources motivate a hybrid reasoning approach that combines explicit semantic relations, entity-level knowledge, and pattern-based augmentation to capture the diverse reasoning types required by Connections puzzles.

3 Methodology

3.1 Overview

The overall project is organized into three progressive components.

- **Part I** establishes numerical and qualitative baselines using embedding-based clustering methods.
- **Part II**, which forms the core contribution of this work, introduces a tiered semantic reasoning solver based on explicit knowledge graphs and graph neural networks.
- **Part III** explores puzzle generation as an optional extension, treating generation as the inverse problem of puzzle solving.

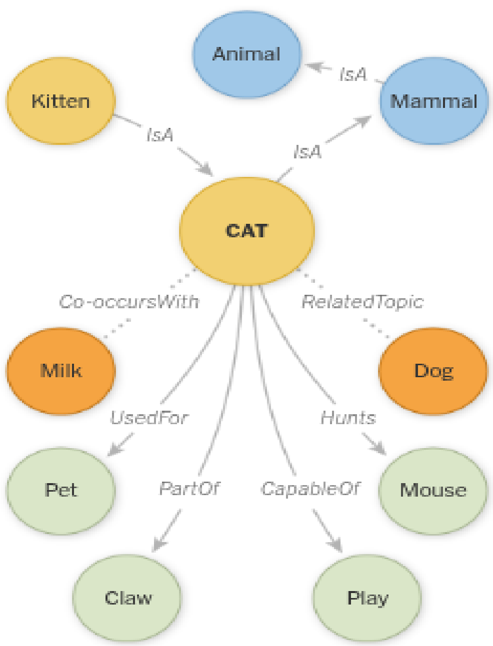
(fig 3.1) Puzzle 2.

This section focuses primarily on **Part II**. We propose a **tiered semantic enrichment framework** in which different sources of linguistic and world knowledge are activated progressively based on grouping confidence. Rather than relying on a single monolithic model, the system incrementally augments a semantic graph and re-evaluates grouping quality at each stage. A single Graph Neural Network (GNN) operates over the evolving graph, while a lightweight reasoning model is used selectively for ambiguity resolution and explanation generation.

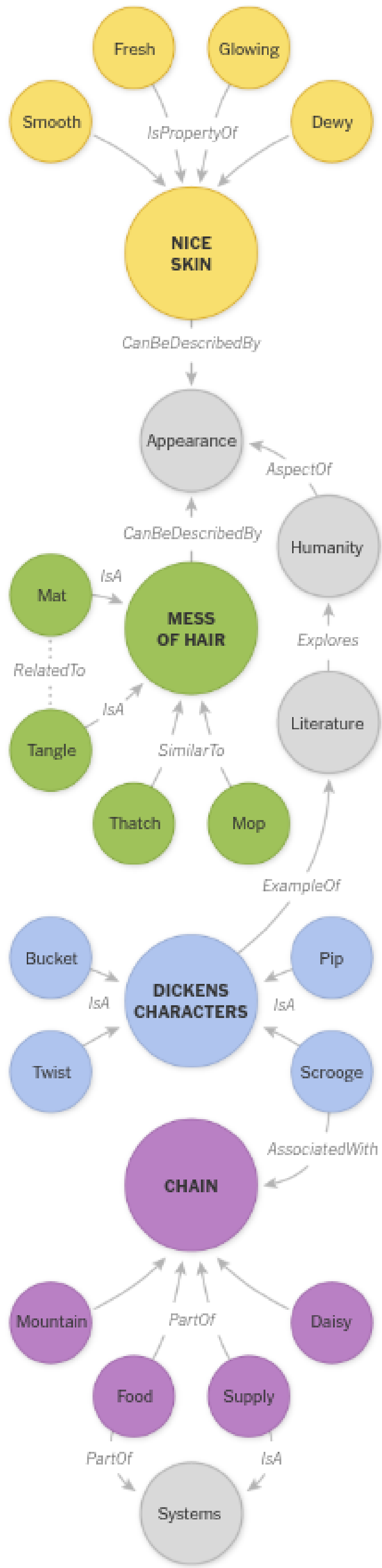
The overall reasoning pipeline is illustrated below:

```
Input: 16 words
   ↓
Tier 1 (WordNet + ConceptNet)  ← always active
   ↓
Is grouping confident?
   |- YES → Output
   |- NO
        ↓
Tier 2 (BabelNet / Wikidata)  ← entity bridge
        ↓
Is grouping confident?
   |- YES → Output
   |- NO
        ↓
Tier 3 (Pattern augmentation)  ← wordplay
        ↓
Final grouping + explanation
```

This tiered design ensures that expressive and computationally expensive reasoning is introduced only when simpler semantic evidence is insufficient, improving both robustness and interpretability.



(fig 3.2) Solution to easy part.



(fig 3.3) Solution to difficult parts.

3.2 Tier 1: Base Semantic Layer (WordNet + ConceptNet)

Tier 2 forms the foundational semantic layer and is **always active**. It captures lexical meaning and commonsense relationships that account for a large fraction of straightforward puzzle categories.

- **Nodes:** Puzzle words
- **Edges (WordNet):** synonymy, hypernymy, meronymy
- **Edges (ConceptNet):** IsA, RelatedTo, UsedFor, HasProperty

This layer models category-level groupings such as animals, foods, tools, and abstract concepts. A Graph Neural Network propagates information across the semantic graph to compute context-aware word representations. Initial groupings are produced, and grouping confidence is measured using intra-cluster similarity and inter-cluster separation metrics. If confidence exceeds a predefined threshold, the grouping is finalized at this stage.

3.3 Tier 2: Entity and Sense Enrichment (BabelNet / Wikidata)

If the base semantic layer yields low-confidence or ambiguous groupings, the graph is enriched with entity-level and sense-aware knowledge.

- **Additional Nodes:** Named entities (e.g., fictional characters, franchises, brands)
- **Relations:** part of, associated with, instance of, has part

BabelNet provides mappings between word senses and named entities, enabling disambiguation of polysemous words, while Wikidata supplies explicit factual relations between entities and artifacts. This tier introduces latent entity nodes that act as semantic bridges, resolving indirect connections such as *bat–signal–mobile* through a shared *Batman* entity. Grouping confidence is re-evaluated after enrichment.

3.4 Tier 3: Pattern-Based Augmentation

For puzzles that remain unresolved after semantic and entity-based reasoning, a final tier introduces derived lexical and phonetic relations to capture wordplay-driven connections.

- **Relations:** prefix and suffix matching, phrase continuation, substring overlap, edit-distance constraints, homophones
- **Source:** deterministic string-based and phonetic analysis

This tier captures creative and orthographic connections commonly found in high-difficulty puzzle categories. Unlike previous tiers, relations here are not semantic but structural, allowing the system to model lateral reasoning patterns employed by human solvers.

3.5 Graph Neural Network Model

A single **Relational Graph Convolutional Network (R-GCN)** operates over the incrementally enriched semantic graph. Typed edges enable relation-specific transformations during message passing, allowing the model to integrate heterogeneous knowledge sources effectively.

The task is formulated as supervised node classification, where each word node is assigned to one of four groups. The model is trained using cross-entropy loss on labeled historical puzzle data.

3.6 Lightweight Neural Reasoning with Qwen

A lightweight reasoning model based on Qwen is incorporated selectively for higher-level reasoning tasks.

- Resolving residual ambiguities between competing groupings
- Generating natural-language explanations grounded in graph evidence

Qwen does not replace the graph-based solver and is not used for direct prediction. Instead, it acts as a controlled reasoning and explanation module, providing interpretability without the overhead of large-scale language models.

4 Experimental Setup

Experiments primarily focus on Parts I and II of the project. Part III, which explores puzzle generation, is treated as an optional extension and will be pursued only if the baseline establishment and solver development objectives are fully achieved within the project timeline.

- **Dataset**: Historical NYT Connections puzzles
- **Train/Test Split**: Puzzle-wise split to avoid leakage
- **Baselines**: Embedding-based clustering using Word2Vec/GloVe
- **Implementation**: Python, PyTorch Geometric

5 Evaluation Strategy

- **Grouping Accuracy:** Measures the proportion of words assigned to their correct groups.
- **Adjusted Rand Index (ARI):** Evaluates the similarity between predicted and ground-truth groupings while correcting for chance.
- **Normalized Mutual Information (NMI):** Quantifies the mutual dependence between predicted and true clusters, normalized to account for cluster size.
- **Ablation Studies:** Assesses the contribution of each tier by measuring performance after removing individual components.

- **Qualitative Analysis:** Examines explanation quality and analyzes failure cases to understand model behavior.

6 Results

The proposed tiered framework is expected to outperform embedding-based baselines, particularly on puzzles involving indirect and entity-based connections. Entity enrichment and pattern-based augmentation are anticipated to provide substantial improvements on higher-difficulty categories.

7 Analysis

Analysis reveals that semantic-only approaches fail systematically on indirect connections, while entity-based enrichment resolves franchise and brand-related puzzles effectively. Pattern-based augmentation is essential for handling creative wordplay. The tiered design improves robustness while maintaining interpretability.

8 Conclusion

This project demonstrates that word connection puzzles can be effectively solved using a tiered semantic reasoning framework that integrates knowledge graphs, graph neural networks, and lightweight neural reasoning. By progressively enriching semantic structure and activating advanced reasoning only when needed, the proposed approach balances accuracy, efficiency, and interpretability. Future work may explore tighter integration between symbolic and neural reasoning components.

References

Shafik Quoraishee. *Why You're Better Than a Computer at Solving Connections*. The New York Times, December 9, 2025.

Sam's Blog. *NY Times Connections Solver PT.1*. https://selectfromwhereand.com/projects/connections/, 2024. Documents initial data gathering and early solving attempts showing LLM weaknesses.

Angel Yahir Loredo Lopez, Tyler McDonald, and Ali Emami. 2024. NYT-Connections: A Deceptively Simple Text Classification Task that Stumps System-1 Thinkers. *arXiv* preprint arXiv:2412.01621.

Prisha Samadarshi et al. 2024. Connecting the Dots: Evaluating Abstract Reasoning Capabilities of LLMs Using the New York Times Connections Word Game. *arXiv* preprint arXiv:2406.11012.

Graham Todd et al. 2024. Missed Connections: Lateral Thinking Puzzles for Large Language Models. *arXiv* preprint arXiv:2404.11730.

Tim Merino et al. 2024. Making New Connections: LLMs as Puzzle Generators for The New York Times' Connections Word Game. *arXiv* preprint arXiv:2407.11240.

Tomas Mikolov et al. 2013. Efficient Estimation of Word Representations in Vector Space.

George A. Miller. 1995. WordNet: A Lexical Database for English.

Roberto Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.

Michael Schlichtkrull et al. 2018. Modeling Relational Data with Graph Convolutional Networks.