Welcome to

instats

The Session Will Begin Shortly

START

Using and Evaluating LLMs in Academic Work

Session 3: From Text to Structure: Knowledge Graphs for LLM Evaluation

Moses Boudourides

Faculty, Graduate Program on Data Science
Northwestern University

Moses.Boudourides@northwestern.edu

Moses.Boudourides@gmail.com

instats Seminar

Thursday, February 26, 2026
4:00 PM – 5:30 PM UTC

Session 3: From Text to Structure: Knowledge Graphs for LLM Evaluation

What are Knowledge Graphs?

- To address the challenge of structural hallucination, since
  - structural hallucination cannot be detected reliably at the sentence level,
- We propose a methodological framework that translates unstructured text into a structured, analyzable format.
- Our framework provides such a methodology by translating text—whether produced by an LLM or written by a human with LLM assistance—into **knowledge graphs**.

Why Networks?

- **Text is linear; knowledge is relational**.
- In particular, knowledge is:
    - Hierarchical
    - Interdependent
    - Clustered
    - Historically layered
- Linear reading makes it difficult to spot structural gaps.
- A network model restores this relational architecture allowing us to represent, visualize and quantify the "skeleton" of an argument.

What is a Knowledge Graph/Network?

- A **knowledge graph** is a network representation of knowledge, where concepts, entities, or documents are represented as **nodes**, and the relationships between them are represented as **edges**.
- Thus, a knowledge graph transforms a document from a linear sequence of sentences into a pattern–network of concepts and relationships.
- Such a structured representation provides a basis for qualitative, quantitative analysis and mixed–analysis that is not possible with the text alone.

Making Structural Hallucination Empirically Detectable

- To render structural hallucination empirically detectable, we propose a methodological framework that translates unstructured text into explicit network representations.
- Doing so, such a translation of scholarly texts enables the inspection of conceptual and bibliographic structure independently of sentence-level fluency.

The Evaluation Strategy

- Objective: Compare the LLM's "mental map" against an established "reference map."
- Step 1: Extract concepts (Nodes).
- Step 2: Extract relationships (Edges).
- Step 3: Map the topology.
- Step 4: Evaluate and validate the LLM's map on the basis of an expert-curated map.

Step 1: Named Entity Recognition (NER)

- Identifying the "Who" and "What" in the text.
- Identifying authors, locations, events, theories, dates, and technical terms.
- Filtering noise: Removing generic words to focus on domain-specific nodes.

Defining the Node

- A Node represents a discrete unit of knowledge.
- Example: "Large Language Models," "Social Network Analysis," "Boudourides (2024)."
- Nodes must be "normalized" (e.g., "LLM" and "Large Language Models" are the same node).
- Nodes must be hierarchically categorized (e.g., "Large Language Models" belong to the broader category of "AI," "Social Network Analysis" to "Network Statistics").

Step 2: Relation Extraction

- Determining how nodes interact.
- Types of relations:
  - **Causal**: "A implies or leads to B."
  - **Associative**: "A is related or affiliated to B."
  - **Bibliographic**: "A is cited by B."
  - **Co-Occurrence**: "A and B co-occur or co-participate into something."

Syntactic Dependency Parsing

- **Syntactic Dependency Parsing** analyzes the grammatical structure of sentences.
- It can reveal more specific relationships, such as subject-verb-object **Triples**, which form the fundamental unit of a Knowledge Graph.
- Triples can be translated into either directed edges in the graph (e.g., “Researcher A Proposes Theory X”) or undirected edges (“Researcher A” collaborates with “Researcher B” in “Project C”).
- Structure of (directed) triples: (Subject) → [Predicate] → (Object).
- Example of a (directed) triple: (LLM) → [generates] → (Hallucination).

Visualizing Triples

- Triples form the building blocks of the entire network.
- Millions of triples can be visualized to show a discipline's complexity.

Semantic Similarity and Edges

- What is Semantic Similarity?
  - Sometimes relations are not explicit and they need to unravel.
  - **Semantic Similarity**, using word embeddings, can be used to draw edges between concepts that are semantically close, even if they do not co-occur directly in the text.
  - Nodes close in semantic similarity are connected by weighted edges.
- Example of Semantic Similarity:
  - Word embeddings for *"king"* and *"queen"* would be located close to each other in the vector space, indicating a strong semantic relationship.
  - In a knowledge graph, this could be represented as an edge between the two nodes, even if they never appear in the same sentence.

Mathematical Formulation for Similarity Edge Weights

- The similarity weight of an edge between two concepts can be calculated using various methods, such as:
  - **Pointwise Mutual Information (PMI):** Measures the association between two concepts.
  - **Cosine Similarity:** Measures the similarity between the vector representations of two concepts.
- These similarity weights can be used to represent the strength of the relationship between concepts.

From Text to Incidence and Adjacency Matrices

- To analyze mathematically, we convert the graph to a matrix.
  - In a unipartite graph, the **adjacency matrix** A is a square matrix such that, if Node i connects to Node j, the matrix cell $A_{ij} = 1$ (or possibly equal to the weight of the connection) and $A_{ij} = 0$ otherwise.
  - In a bpartite graph (or a hypergraph), the **incidence matrix** B is a generally non-square matrix such that, if Node i is affiliated to Node j, the matrix cell $B_{ij} = 1$ (or possibly equal to the weight of the affiliation) and $B_{ij} = 0$ otherwise.
- This allows us to use linear (or tensor) algebra to find "hidden" structures (blocks).

Building the Reference Graph

- How do we know if the LLM is right?
- We need a "Ground Truth."
- Why network comparison?
  - The power of the network analysis comes from the comparison of the knowledge network extracted from an LLM-assisted (artificial) corpus and reference networks derived from a trusted corpus.
  - To evaluate an LLM-generated corpus, we can compare the knowledge graph to one or more “reference graphs” constructed from sources that are considered authoritative in the field.

Sources for Reference Graphs

- These reference graphs can be constructed from various sources, including:
  - Canonical textbooks
  - Curated bibliographies, overview articles, course syllabi, etc.
  - Encyclopedias and lexicons
  - Bibliographic databases
    - Standard databeses: Web of Science, Scopus, Dimensions
    - Other databases: OpenAlex, PubMed, Google Scholar etc.
  - Diaries and archives
  - Expert-curated documents
  - Course syllabi

Extracting the LLM-Generated Graph

- We prompt the LLM to generate a response document (e.g., summary, highlights, focal points etc.) through an appropriately phrased contextual query.
- We run the same extraction protocol on the LLM generated respond (to derive, for instance, a triple).
- Result: A network (knowledge graph) representation of the LLM's output.

Assessing the Alignment of the Two Graphs

- We overlay the LLM graph on the Reference graph.
- Comparing nodes:
  - **True Positive Nodes**: Node exists in both
  - **False Positive Nodes**: Node exists only in LLM (**Conceptual hallucination!**)
  - **False Negative Nodes**: Node missing in LLM (Conceptual omission)
- Comparing edges:
  - **True Positive Relation**: Relation exists in both
  - **False Positive Relation**: Relation exists only in LLM (**Structural hallucination!**)
  - **False Negative Relation**: Relation missing in LLM (Structural omission)

Detecting Structural Divergence

- By comparing the LLM-generated graph to these reference graphs, we can identify areas of structural divergence.
- Structural divergence between these networks can be treated as an indicator of conceptual distortion or bibliographic misalignment, thereby operationalizing the assessment of structural integrity.

Evaluating the Alignment of the Two Graphs

- Visualizing Graph Overlap
  - Discrepancies immediately highlight where the LLM's logic deviates from the field.
- Syntactic vs. Semantic Extraction
  - **Syntactic**: Based on sentence grammar (Subject-Verb-Object).
  - **Semantic**: Based on the meaning and context of terms.
  - A robust evaluation uses both to capture nuances.
- Challenges: Entity Resolution
  - Problem: "AI," "Artificial Intelligence," and "Machine Learning" might refer to the same concept.
  - Solution: Entity Linking to a standard ontology (like DBpedia or Wikidata).
- Challenges: Predicate Normalization
  - Problem: "leads to," "causes," and "results in" mean the same thing.
  - Solution: Mapping predicates to a simplified set of relational labels.

Practical Implementation Details

- The process of creating and comparing knowledge graphs can be automated using a combination of scripting and existing software libraries.
- The results of the comparison can be visualized to provide a clear and intuitive representation of the structural differences between the LLM-generated text and the reference corpus.
- Scalability of the Method:
    - This can be applied to a single abstract or a corpus of thousands or millions of units
    - Consistent methodology across different scales

Measuring the Alignment of the Two Graphs

- Quantifying Graph Density
  - Density = actual connections / potential connections.
  - High density in LLM graphs may indicate "reliable representation."
  - Low density may indicate "fragmented knowledge."
- Identifying Structural Connectivity
  - Articulation points or bridges are nodes or edges (resp.) that connect two separate clusters in a graph.
  - If an LLM misses an articulation point or a bridge, it fails to connect two sub-fields.
- Intellectual Lineages or Trajectories as Paths
  - A lineage or a trajectory is a directed path through a graph: $A \to B \to C$.
  - Evaluating if the LLM preserves the "chronological" and "logical" flow of ideas.

Bibliometric Graphs

- Networks = Co-authorship/Citation/Shared keywords or thematic proximity.
- Nodes = Authors/Papers or authors/Concepts or keywords or research areas.
- Edges = Co-authorships/Citations/Semantic similarities.
- These bibliometric graphs:
  - show the "social" structure of science
  - reveal the flow of ideas and the influence of different works in a field
  - identify clusters of research on similar topics

Session 3 Summary

- Knowledge graphs provide a powerful tool for evaluating the structural integrity of LLM-generated text.
- They allow us to move beyond sentence-level analysis to a more holistic assessment of conceptual and bibliographic structure.
- By comparing LLM-generated graphs to reference graphs, we can empirically detect and measure structural hallucination.

Looking Ahead to Session 4: Educational and Academic Benefits

- From Challenges in Teaching:
  - Teaching students to build these graphs encourages "Structural Literacy."
  - Students learn how ideas connect, not just what the ideas are.
- To Diagnosis and Rectification:
  - Now that we have the graph, how do we "diagnose" it?
  - Next Session: Network Diagnostics and Centrality.
  - Applying the "Hallucination Stress Test."

Questions and Discussion

Thank you!

Questions?

Moses.Boudourides@northwestern.edu
Moses.Boudourides@gmail.com

STOP