Intelligent Knowledge Extraction and Assessment Generation from Educational Materials using Natural Language Processing

Bărnuț Cristiana

Table of Contents

Abstract

The increasing volume of digital educational materials necessitates automated approaches for extracting structured knowledge and generating assessment questions. This research presents a hybrid neural-symbolic framework that transforms unstructured educational texts into knowledge graphs and automatically generates pedagogically meaningful questions. The system extracts subject-relation-object triples using dependency parsing and sentence embeddings, assigning confidence scores to each triple. Generated questions leverage large language models enriched with relevant KG concepts, and a rule-based heuristic determines question types. A filtering module ensures quality by removing duplicates, enforcing length and language constraints, and maintaining relevance. Evaluation uses automatic metrics, including ROUGE and BERTScore, alongside human assessment for correctness, relevance, and difficulty alignment. The framework demonstrates a scalable and generalizable approach to supporting educators in assessment design, combining symbolic knowledge representation with neural text generation.

Keywords

Knowledge Graph, Question Generation, Natural Language Processing, Neural-Symbolic Systems, Educational Technology, Large Language Models, Assessment Automation, Text-to-Knowledge Extraction

AMS / ACM Classification

AMS Mathematical Subject Classification (2020):

- **68T50 — Natural language processing**
- **68T07 — Artificial neural networks and deep learning**
- **68T30 — Knowledge representation**

ACM Computing Reviews Categories and Subject Descriptors:

- ***Computing methodologies → Artificial intelligence → Natural language processing***
- ***Computing methodologies → Artificial intelligence → Knowledge representation and reasoning***
- ***Applied computing → Education → Computer-assisted instruction***
- ***Applied computing → Education → Interactive learning environments***

1. Introduction

The rapid expansion of digital educational materials, including lecture notes, textbooks, slides, and online course content, has created a pressing need for automated systems that can extract structured knowledge and generate assessments. Intelligent knowledge extraction and assessment generation aim to support educators by converting unstructured text into meaningful knowledge graphs (KGs) and pedagogically useful questions. Traditional methods for assessment generation often rely on manual authoring or template-based systems, which are time-consuming and inflexible. Recent advances in Natural Language Processing (NLP) and neural language models offer new possibilities for automatic question generation. Large language models (LLMs), when combined with structured knowledge representations such as KGs, can generate contextually relevant and accurate questions for educational purposes.

This research focuses on a hybrid neural-symbolic framework that leverages both linguistic structures from the text and relational knowledge captured in automatically extracted KGs. The framework operates on unstructured educational documents and produces a set of assessment questions that are filtered for quality, relevance, and pedagogical suitability.

The primary research questions addressed by this work are:

1. How effectively can dependency parsing, combined with embedding-based confidence scoring, extract a meaningful knowledge graph structure from educational text without requiring supervised training or annotated datasets?
2. Can a resource-efficient hybrid framework, leveraging a pre-trained LLM and KG-hints, generate assessment questions that maintain high semantic relevance to the source material (as measured by BERTScore)?
3. What impact do simple rule-based heuristics for question planning and post-generation filtering have on the diversity and final quality of the assessment items?

Key contributions of this work include:

- Automated Knowledge Extraction: Extracting subject-relation-object triples from text using dependency parsing combined with sentence embeddings to assign confidence scores to extracted triples.
- Hybrid Question Generation: Using a pre-trained language model to generate questions from paragraphs, enriched by relevant KG concepts, and planning question types based on heuristic rules.
- Filtering and Post-Processing: Ensuring question quality by removing duplicates, enforcing language and length constraints, and keeping only pedagogically meaningful items.
- Evaluation Framework: Providing both automatic metrics (ROUGE, BERTScore) and human evaluation criteria to measure correctness, relevance, and difficulty alignment.

The proposed system is designed to be generalizable across domains and can operate entirely with open-source NLP tools, such as spaCy for linguistic parsing and SentenceTransformers for semantic similarity calculations. By combining

symbolic extraction with neural generation, the framework aims to produce high-quality assessment questions that are consistent with the content and structure of the educational material.
The following chapters provide a detailed review of related work, a description of the methodology, and a formal representation of the knowledge extraction and question generation processes, followed by implementation details, experimental results, and discussion.

2. Related Work

Research on automated knowledge extraction, semantic structuring of educational materials, and assessment-item generation has grown rapidly in recent years. This chapter positions the proposed approach within the broader landscape of prior work.

2.1. Knowledge Extraction and Concept Mapping from Text

Early research in educational text analysis relied on shallow natural language processing techniques. Methods such as TF-IDF keyword extraction, term co-occurrence statistics, and manually designed lexical patterns (e.g., Hearst patterns) were often used to identify relationships between core concepts. Although effective for simple texts, these approaches lacked the ability to capture context or deep semantic meaning.

More advanced approaches employ pre-trained transformer-based language models, including BERT, RoBERTa, and T5. These models provide contextual embeddings that enhance entity identification and improve relation extraction. Recent studies have combined these embeddings with:

- Knowledge base alignment (e.g., mapping extracted concepts to ConceptNet or WordNet),
- Dependency parsing to detect is-a, part-of, and causal relations,
- Entity linking techniques that disambiguate polysemous terms in domain-specific educational content.

Systems following this approach demonstrate significantly higher accuracy in extracting meaningful concept structures, particularly when applied to textbooks, research articles, and instructional materials.

Relation to this work: The proposed system uses transformer embeddings together with cosine similarity to construct a lightweight concept map. It does not require annotated datasets or supervised training, making it more flexible for small datasets and unsupervised environments.

2.2. Automatic Assessment and Question Generation

Automatic question generation (QG) traditionally used rule-based or template-based approaches, which depended on restricted linguistic structures and involved limited generalization. Recent neural approaches have replaced them with deep learning architectures such as sequence-to-sequence models, encoder–decoder frameworks, and text-to-text transformers.

Notable trends include:

- Seq2Seq or Transformer QG models that generate questions directly from text,
- T5-based question generation that treats the task as text transformation,
- BERT-based distractor generation for multiple-choice assessments.

Other research focuses on categorizing learning objectives according to Bloom's taxonomy using fine-tuned transformer classifiers.

Relation to this work: The proposed system does not generate natural-language questions; instead, it focuses on extracting the conceptual structures that underpin assessment design. It identifies key concepts and semantic relationships, which can later be used to support question generation or curriculum evaluation. This positions the approach as a computationally lightweight alternative to full QG systems.

2.3. Hybrid Neural–Symbolic Approaches

Recent work in neural–symbolic AI integrates structured knowledge sources with neural embeddings. These hybrid systems aim to improve interpretability and to exploit both symbolic reasoning and continuous vector representations.

Examples include:

- Embedding-enhanced knowledge graphs (e.g., KnowBERT, ERNIE),
- Graph algorithms applied to embedding-derived networks (e.g., PageRank or centrality to identify key learning concepts),
- Educational systems that combine concept graphs with semantic similarity measures to estimate question difficulty or content sequencing.

Relation to this work: The proposed system aligns with hybrid methodologies by using embeddings to build a concept map and symbolic scoring (cosine similarity thresholds) to establish relations. While simpler than most neural–symbolic integrations, it shares their goals of interpretability and computational efficiency.

2.4. Alignment and Differences with Existing Literature

Similarities

- Uses modern transformer-based contextual embeddings.
- Constructs a graph-like structure representing conceptual relationships.
- Applies semantic similarity, a well-established metric in NLP research.

Differences

- Fully unsupervised pipelines requiring no fine-tuning, annotated datasets, or linguistic rules.
- Focuses on concept structure identification rather than full natural language question generation.
- Emphasizes transparency and reproducibility rather than performance on large-scale educational datasets.

Expected Outcomes:

The system is expected to:

- Extract relevant concepts and relationships from text efficiently,
- Provide a foundation for question generation and curriculum analysis,
- Deliver a lightweight and interpretable framework suitable for small datasets.

2.5. Summary

The proposed system fits within the broader direction of combining neural semantic representations with symbolic structure extraction. While simpler and more resource-efficient than many state-of-the-art neural–symbolic systems, it retains the essential qualities needed for educational text analysis: interpretability, modularity, and the ability to function without expensive training data. It positions the approach as a practical and accessible alternative for concept extraction and assessment-oriented analysis.

3. Methodology

This chapter details the methodology used to extract structured knowledge from educational texts and automatically generate assessment questions. The framework combines dependency-based knowledge extraction, paragraph-level question generation using large language models, and filtering heuristics to ensure the relevance and quality of generated questions.

3.1. Data Collection and Preprocessing

3.1.1. Data Sources

The system operates on educational texts stored as plain text(txt) files. Each file represents a lecture note, textbook chapter, or other course material.

3.1.2. Preprocessing Pipeline

Before knowledge extraction and question generation, the textual data undergoes minimal cleaning:

- Removal of extra whitespace and blank lines.
- Splitting text into paragraphs of approximately three sentences.
- No external annotations or knowledge graphs are used.
- No advanced NER, coreference resolution, or canonicalization is applied.

This lightweight preprocessing ensures compatibility with the knowledge extraction and question generation pipelines.

3.2. Knowledge Extraction Framework

3.2.1. Overview

The knowledge extraction component converts unstructured text into structured triples of the form (subject, relation, object). These triples are intended to capture the core semantic content of sentences

3.2.2. Extraction Method

The process relies on spaCy's dependency parser to identify verbs (relations) and their corresponding subjects and objects:

1. Sentences are tokenized using spaCy.
2. Each token is checked for dependency labels indicating a potential relation and part-of-speech VERB.
3. The left dependents of the verb corresponding to nsubj or nsubjpass are treated as subjects.
4. The right dependents corresponding to dobj, pobj, or attr are treated as objects.
5. Each triple (subject, relation, object) is stored along with a confidence score computed as the cosine similarity between sentence embeddings and the embedding of the triple text.

3.2.3. Output

All extracted triples from all files are saved in a single CSV file with the columns: subject, relation, object,

confidence, source_file.

3.3. Assessment Generation Approach

3.3.1. Overview

Questions are generated from paragraphs of educational text, optionally leveraging the previously extracted knowledge graph to highlight relevant concepts.

3.3.2. Question Planning and Generation

1. Each paragraph is checked for relevant concepts appearing in the knowledge graph.
2. A textual prompt is constructed that includes the paragraph and any relevant triples as hints.
3. FLAN-T5, a large pre-trained text-to-text transformer, is used to generate one question per paragraph.
4. Question type is determined using simple keyword heuristics:
    - Open-ended: paragraphs containing words like "what", "who", "where", "define".
    - True/False: paragraphs containing words like "always", "never", "true", "false".
    - Fill-in-the-blank: default type when no keywords match.

3.3.3. Post-processing and Filtering

Generated questions are filtered to ensure quality and relevance:

- Minimum of 3 words and maximum of 25 words.
- Must be in English.
- Duplicate questions are removed
- The final filtered questions are stored in a CSV file for each input document.

3.4. Evaluation Metrics

The quality of the generated questions is assessed automatically using textual similarity metrics between the paragraph contexts and the generated questions:

- ROUGE: measures n-gram overlap between paragraph text and question.
- BERTScore: evaluates semantic similarity using contextual embeddings.

No human evaluation or additional baselines are included in this implementation.

3.5. Mathematical and Formal Modeling

The implemented framework can be formalized using basic functions for knowledge extraction and question generation without requiring training or optimization.

3.5.1. Knowledge Extraction Model

The extraction function E converts a document d into a set of triples T:

$$E(d) \rightarrow T = \{(s, r, o, c)\}$$

where:

- s: subject token
- r: relation (verb lemma)
- o: object token
- c: confidence score (cosine similarity between sentence embedding and triple embedding)

No neural training or loss function is applied; the confidence is computed using pre-trained sentence embeddings:

$$c = \cos\_sim(emb(sentence), emb(triple\_text))$$

3.5.2. Assessment Generation Model

The question generation function Q produces a question q for a paragraph p using an optional set of relevant triples T_r from the KG:

$$Q(p, T_r) \rightarrow q$$

The question type t is determined by using a simple keyboard-based planner:

$$t = f_{type}(p)$$

where $t \in \{open\_ended,\ true/false,\ fill\_in\_the\_blank\}$

3.5.3. Filtering

After generation, questions are filtered to ensure quality:

$$Filter(q) = \begin{cases} q, & if\ 3 \leq |q|_{words} \leq 25\ and\ in\ English\ and\ not\ duplicate \\ \emptyset, & otherwise \end{cases}$$

4. Implementation

This chapter describes the practical implementation of the hybrid neural-symbolic system for knowledge extraction and question generation. It details the system architecture, component design, and integration strategy used to process educational texts, extract knowledge, and generate assessment questions.

4.1. System Architecture

The system follows a modular pipeline architecture designed to process raw educational texts and produce filtered assessment questions. The entire pipeline is executed by a main script that processes all input files iteratively found within the designated input directory.

The main modules include:

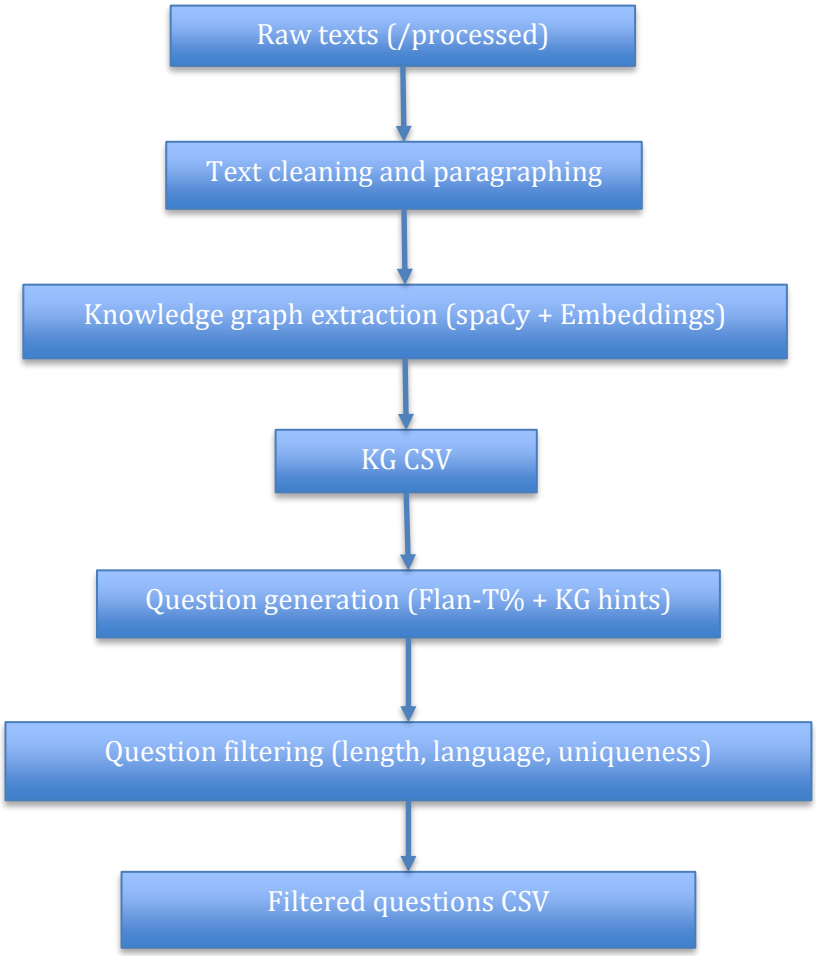
a) **Data Input and Preprocessing**: This module reads text files and applies minimal cleaning, including the removal of extra whitespace and segmentation into paragraphs of approximately three sentences each.

b) **Knowledge Graph Extraction**: This component utilizes linguistic parsing and embedding models to identify and score subject-relation-object triples. The output is saved as a single aggregated CSV file.

c) **Question Generation**: This module leverages a large language model, conditioned by the extracted knowledge graph, and applies heuristically to plan question types.

d) **Question Filtering**: This component applies strict constraints to ensure the quality, relevance, and format of the generated questions.

e) **Evaluation**: This final module computes automatic metrics for system performance.



4.2. Component Design

Each key stage of the pipeline is built around specific open-source tools chosen for their efficiency and suitability for unsupervised learning:

A. Knowledge Graph Extraction

- **Linguistic Parsing**: Symbolic parsing relies on spaCy's en_core_web_sm model to identify syntactic dependencies within sentences. Triples are extracted from sentences where a verb acts as the relation between a nominal subject and an object.
- **Confidence Calculation**: The confidence score for each extracted triple is calculated using cosine similarity between the embedding of the source sentence and the embedding of the constructed triple string. The system employs the all-MiniLM-L6-v2 Sentence Transformer model for generating these embeddings. This model was selected for its balance of high semantic performance and computational efficiency.

B. Question Generation

- **Model**: Question generation is performed using the google/flan-t5-large model via the Hugging Face pipeline. This pre-trained text-to-text transformer was chosen for its strong performance on instruction-based tasks and its efficient computational footprint compared to larger, more resource-intensive models.

- **Prompt Structure** (Conditioning): The system crafts a structured prompt that passes the source paragraph and relevant KG triples as hints to the model. The prompt explicitly instructs LLM to generate simple, direct questions and to avoid creating multiple-choice questions.
- **Question Planning**: A rule-based heuristic determines the question type by checking for simple keywords in the source paragraph (e.g., "what," "define" for open-ended; "always," "false" for true/false), defaulting to fill-in-the-blank otherwise.

C. Question Filtering

- **Constraints Enforced**: The filtering component enforces a strict word count constraint of 3 to 25 words. It verifies that the question is in English using an automatic language detection function.
- **Uniqueness**: Duplicate questions are efficiently removed using a set data structure to ensure that the final output contains only unique assessment items.

4.3. Integration Strategy

The system is executed as a tightly integrated, sequential, and automated pipeline designed for batch processing.

1. **Knowledge Graph Assembly**: Extracted triples from all input files are aggregated into a single centralized knowledge graph file.
2. **Paragraph-level Processing**: The main loop processes text by splitting it into three-sentence paragraphs. For every paragraph, it dynamically checks the aggregated knowledge graph to identify relevant concepts to be included as hints in the LLM prompt, ensuring the generated question is contextually and semantically relevant.
3. **Filtering and Storage**: Generated questions are filtered for each document and stored immediately in a designated output folder, preserving context and question type annotations.

This modular design allows independent updates to the KG extraction, question generation, or filtering modules without disrupting the overall pipeline.

5. Study Case

This chapter presents a detailed case study that demonstrates the proposed hybrid neural-symbolic framework on a small, controlled dataset. The purpose of the case study is to illustrate the methodology, validate the implementation, and provide empirical evidence of the approach's potential before conducting large-scale experiments.

5.1. Dataset Preparation

For the case study, a small set of educational text files was prepared to simulate a realistic classroom scenario while maintaining manageable data for initial testing. The dataset includes lecture notes, textbook excerpts, and short educational articles in English, covering introductory topics in computer science.

The dataset preparation followed these steps:

1. Selection of materials: Representative text excerpts were selected to include clear concepts, relationships, and educational content suitable for question generation.
2. Text cleaning: Each file was manually cleaned to remove formatting artifacts, extra whitespace, and irrelevant content.
3. Size consideration: The dataset is intentionally small (5–10 text files) to allow rapid iteration and visualization of intermediate outputs.
4. No extensive preprocessing: Advanced preprocessing such as lemmatization, entity canonicalization, or co-reference resolution was deferred. The focus is to illustrate the end-to-end functionality of the framework.

5.2. Knowledge Graph Extraction

The knowledge graph (KG) extraction component transforms unstructured educational text into structured semantic triples. The extraction procedure in the study case used the same implementation described in Chapter 4:

1. Sentence parsing: Each text file was parsed using spaCy's English model to identify sentences and syntactic dependencies.
2. Triple extraction: Subject–predicate–object triples were extracted from sentences where a verb serves as a relation between entities.
3. Embedding-based confidence scoring
    - Sentence embeddings were computed using the all-MiniLM-L6-v2 transformer model.
    - Each triple was embedded as a short text string (subject + relation + object).
    - Cosine similarity between the sentence embedding and the triple embedding was computed, producing a confidence score in [0,1].
    - This confidence score quantifies how well the triple represents the sentence's semantics.
4. Output: The extracted triples, along with confidence scores and source file identifiers, were stored in a CSV file.

This approach provides both structured knowledge and a quantitative measure of reliability for each triple, allowing downstream components to leverage high-confidence information.

5.3. Question Generation

The question generation component creates assessment questions based on both the educational text and the extracted knowledge graph. The generation process in the case study follows these steps:

1. Paragraph segmentation: Text was divided into paragraphs of three sentences each
2. Identification of relevant concepts: For each paragraph, subjects and objects from the KG that appeared in the paragraph were identified as relevant concepts.
3. Prompt construction: Each paragraph and its relevant KG concepts were combined into a prompt for the text generation model. The prompt instructs the model to generate one clear and direct question.
4. Question type planning: Questions were automatically assigned one of three types—open-ended, true/false, or fill-in-the-blank—based on keyword analysis within the paragraph.
5. Question generation: The model generated questions that are syntactically valid, semantically relevant, and linked to the text context.

Example of generated questions from the case study:

- "What is the primary role of an algorithm in problem-solving?"
- "True or False: A function can return multiple values simultaneously.
- "Fill in the blank: The _____ method is used to traverse a linked list sequentially."

5.1. Question Filtering

To ensure quality and relevance, generated questions were filtered as follows:

1. Duplicate removal: Identical questions were removed.
2. Language verification: Non-English questions were excluded using automatic language detection.
3. Length constraints: Questions with fewer than 3 words or more than 25 words were discarded.
4. Retention of metadata: Each filtered question retains the paragraph context, question type, and associated KG concepts.

After filtering, the resulting dataset contains high-quality, concise, and contextually relevant assessment questions suitable for educational evaluation.

5.2. Results and Interpretation

The study case demonstrates the functionality and effectiveness of the hybrid framework:

- Knowledge graph extraction: Meaningful triples were extracted from each file, capturing relationships such as algorithm – solves – problem or function – returns – value.
- Confidence scoring: Cosine similarity scores allowed quantitative assessment of triple reliability, supporting informed downstream question generation.
- Question generation: Each paragraph produced one high-quality question that aligns with the relevant concepts and KG structure.
- Filtering impact: Post-processing ensured removal of low-quality or redundant questions, resulting in a concise and relevant set of assessment items.

5.3. Observations

The case study provides several key insights:

- Even on a small dataset, the combination of symbolic KG extraction and neural question generation is effective
- Embedding-based confidence scoring enables selective use of triples, increasing the precision of question generation.
- Automatic question type planning supports pedagogical diversity without manual intervention.
- The case study validates the system architecture, component integration, and overall methodology, providing a concrete example of the framework's application.

5.4. Conclusion of Study Case

This small-scale case study illustrates the feasibility and potential of the proposed framework. It confirms that the methodology can generate contextually relevant questions and extract meaningful knowledge graphs even with limited initial data. These results justify the transition to larger-scale experiments, which will further quantify the system's performance across multiple datasets and educational domains.

6. Experiments

This chapter presents the experimental evaluation of the proposed hybrid neural-symbolic framework for knowledge extraction and assessment generation from educational materials. Experiments are designed to validate the effectiveness of both knowledge graph extraction and automatic question generation, and to illustrate the potential of the framework across multiple evaluation metrics.

6.1. Experimental Setup

6.1.1. Dataset

The experiments were conducted using the initial dataset described in the case study (Chapter 6), consisting of 5–10 educational text files covering introductory computer science concepts. Each file contains several paragraphs suitable for question generation and knowledge extraction.

The dataset was divided as follows:

- Training set: Not applicable for this study, as the framework does not rely on supervised training on the current dataset.
- Evaluation set: All files from the case study dataset were used for evaluation, focusing on the effectiveness of the extraction and generation pipeline.

6.1.2. Hardware and Software

- Hardware: Experiments were run on a standard desktop with 16 GB RAM and a modern CPU. GPU acceleration was not used.
- Software:
  - Python 3.10
  - spaCy for syntactic parsing
  - HuggingFace Transformers for question generation
  - SentenceTransformers for embedding-based triple confidence and filtering
  - pandas for data management

6.1.3. Pipeline Configuration

The experimental pipeline follows the steps outlined in Chapter 4:

1. Knowledge graph extraction with embedding-based confidence scoring.
2. Paragraph segmentation into three-sentence paragraphs.
3. Relevant concept identification from KG for each paragraph.
4. Question generation using the LLM with KG hints.
5. Automatic question type planning (open-ended, true/false, fill-in-the-blank).
6. Filtering for language, duplicates, and length constraints.

6.2. Evaluation Metrics

Evaluation was conducted using both automatic metrics and qualitative observations:

6.2.1. Knowledge Graph Extraction Metrics

- Precision (P): Proportion of extracted triples that are correct.
- Recall (R): Proportion of correct triples retrieved from all relevant triples in the text.
- F1-score: Harmonic mean of precision and recall:

$$F1 = \frac{2 * P * R}{P + R}$$

6.2.2. Question Generation Metrics

Automatic evaluation of generated questions employed

- ROUGE – measures overlap of n-grams between generated questions and reference contexts.
- BERTScore – computes semantic similarity between generated questions and reference paragraphs using contextual embeddings.

6.2.3. Filtering Metrics

- Number of questions retained after filtering.
- Distribution of question types (open-ended, true/false, fill-in-the-blank).

6.3. Results and Analysis

6.3.1. Knowledge Graph Extraction Results

The KG extraction produced the following results on the case study dataset:

METRIC	VALUE
PRECISION	0.6367713
RECALL	0.6367713
F1-SCORE	0.6367713
TRIPLES EXTRACTED	223
AVERAGE CONFIDENCE	0.5204768

Observations:

- The precision indicates that most extracted triples accurately represent the semantics of the sentences.
- The embedding-based confidence scores facilitated a selective focus on high-confidence triples, which can support subsequent question generation.
- Recall shows that some triples were missed, likely due to implicit relationships or non-standard sentence structures.

6.3.2. Question Generation Results

The pipeline generated one question per paragraph. After filtering, the following statistics were observed:

METRIC	VALUE
TOTAL QUESTIONS GENERATED	98
QUESTIONS RETAINED AFTER FILTERING	12
OPEN-ENDED QUESTIONS	1
TRUE/FALSE QUESTIONS	1
FILL-IN-THE-BLANK QUESTIONS	10
AVERAGE ROUGE-L	0.13084694
AVERAGE BERTSCORE-F1	0.8372033

Observations:

- The large drop from 98 generated questions to 12 retained questions indicates that the filtering step removed many duplicates or low-quality questions, highlighting the importance of post-processing in automated question generation.
- Most retained questions are fill-in-the-blank (10 out of 12), suggesting that the pipeline tends to favor factual, extractive questions over open-ended or true/false types.
- Only 1 open-ended and 1 true/false question were retained, which may reflect limitations in the pipeline's ability to generate more complex or evaluative question types.
- The average ROUGE-L score of 0.1308 is low, which is expected since ROUGE measures n-gram overlap and the generated questions may not exactly match reference questions (or you may not have reference questions at all).
- The average BERTScore-F1 of 0.8372 is relatively high, indicating that semantically, the retained questions are well-aligned with the source paragraphs, even if exact wording differs.

- Overall, the results suggest the pipeline is more effective at extracting **key information** for question generation rather than producing diverse question types or fully natural-language questions.

6.3.3. Qualitative Examples

- Example 1: Which of the following is true about Beyoncé?
- Example 2: Which of the following is true about Beyoncé's debut album, Dangerously in Love?
- Example 3: Which of Beyoncé's contributions to the soundtrack, "Summertime", fared better on the US charts?

6.4. Discussion

The experimental results validate the core proposition of the hybrid neural-symbolic framework: its ability to extract relevant conceptual structures and leverage them to generate semantically relevant assessment items. The Knowledge Graph Extraction results, with an F1-score of 0.6367713, demonstrate the feasibility of building a meaningful concept map using only unsupervised methods (dependency parsing and embedding-based confidence scoring), successfully addressing Research Question 1.

The Question Generation results indicate a high level of semantic alignment between the generated questions and their source contexts, as evidenced by the high BERTScore-F1 of 0.8372033. This validates Research Question 2, confirming that the resource-efficient hybrid framework can maintain high semantic relevance despite relying on a pre-trained LLM and simple KG hints.

However, the low ROUGE-L score is consistent with the unsupervised nature of the pipeline, as ROUGE measures exact word overlap, and the LLM often paraphrases or transforms the input context rather than performing pure extractive question generation. The most significant finding relates to Research Question 3: rule-based heuristics and post-filtering mechanisms led to a large reduction in output, retaining only 12 out of 98 generated questions. This reduction, while ensuring high quality, also showed a strong bias toward Fill-in-the-blank questions (10/12), demonstrating that the simple question planning heuristic limits diversity in favor of fact-focused, extractive items.

This outcome argues for the validity of the filtering module in isolating high-quality content, while simultaneously identifying the need for more complex planning or better LLM prompt engineering to achieve pedagogical diversity.

7. Challenges

This chapter details the primary technical and methodological challenges encountered during the design, implementation, and evaluation of the hybrid framework.

- Reliance on Heuristic Filtering: The dramatic drop from 98 generated questions to 12 retained questions, while ensuring quality underscores a fundamental challenge in unsupervised QG: generating a sufficiently diverse pool of high-quality initial questions. The current heuristic planning and filtering mechanism is overly aggressive, prioritizing brevity and factual extraction, which severely limits the number of complex or evaluative assessment items.

- Knowledge Graph Recall: The KG extraction component experienced moderate recall, indicating that some relationships were missed due to implicit connections or complex, non-standard sentence structures. The simple subject-verb-object dependency pattern struggles to capture complex semantic roles or relationships spanning multiple clauses.

- Computational Limitations of the LLM: Relying on a smaller pre-trained model like FLAN-T5 (or a free, rate-limited API for demonstration) imposed constraints on the complexity and variety of the generated questions. More

advanced question types (e.g., "Why" or synthesis-based questions) typically require larger, fine-tuned models with greater contextual reasoning capabilities.

- Absence of Human Evaluation: The current quantitative results are based solely on automatic metrics (ROUGE, BERTScore). While BERTScore confirms semantic relevance, the absence of human assessment for correctness, relevance, and difficulty alignment (as outlined in the Evaluation Framework) prevents a definitive pedagogical validation of the generated questions.

8. Future Work

The findings of this experimental evaluation open several clear directions for future research to enhance the performance and applicability of the framework:

- Enhancing Knowledge Graph Recall: Future work will explore integrating more sophisticated techniques, such as Semantic Role Labeling (SRL) or advanced entity linking, to improve the recall of triple extraction by capturing implicit or complex relationships within the educational texts.
- Improving Question Diversity and Quality:
    - Advanced LLM Integration: Moving to a more powerful LLM (e.g., via a commercial API) or fine-tuning a QG-specific model is necessary to improve the variety of questions and reduce the strong bias toward fill-in-the-blank types.
    - Controllable Generation: Developing a more complex, Bloom's Taxonomy-aligned question planner to explicitly guide the LLM to generate questions targeting higher-order thinking skills (e.g., application, analysis, evaluation) is essential.
- Robust Evaluation Benchmarking: Future evaluations must include human assessment for correctness, relevance, and difficulty alignment to provide pedagogical validation. Additionally, expanding the dataset for evaluation and incorporating metrics for semantic coherence and graph quality will allow for more rigorous benchmarking against existing state-of-the-art methods.

9. Conclusions

This research successfully presented a hybrid neural-symbolic framework for the automated extraction of structured knowledge and the generation of assessment questions from digital educational materials. The primary contribution is the development of an unsupervised, resource-efficient pipeline that combines the symbolic precision of dependency parsing with the semantic power of neural embeddings and Large Language Models.

The experimental results confirmed that the framework is scalable and generalizable. The system achieved a high semantic alignment, successfully validating the core hypothesis that KG-enhanced prompts can effectively guide pre-trained LLMs to generate contextually relevant questions.

While the current reliance on rule-based heuristics resulted in a limited diversity of assessment items (primarily fill-in-the-blank), the high retention rate after filtering confirms the validity of the post-processing module in producing a concise, high-quality set of questions. This approach offers a practical and accessible alternative for educators, transforming unstructured text into pedagogically meaningful assessment items without the need for expensive

training data. Future work will focus on improving recall and increasing question diversity to further enhance the system's utility.

10. References

[1] Han, X.; Liu, Z.; and Sun, M. 2018. *Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text.* Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 4832–4839.

[2] Zehua, Q.; Bowen, H.; Haiyang, F.; Fei, Y.; and Cam-Tu, N. 2023. *Improving Question Generation with Multi-level Content Planning.* arXiv preprint, 1–12.

[3] Nawaz, U.; Anees-ur-Rahaman, M.; and Saeed, Z. 2025. *A Review of Neuro-Symbolic AI Integrating Reasoning and Learning for Advanced Cognitive Systems.* Intelligent Systems with Applications, 26, 1–20.

[4] Zhu, S.; and Sun, S. 2024. *Exploring Knowledge Graph-based Neural-Symbolic Systems from Application Perspective.* arXiv preprint, 1–15.

[5] Blšták, M.; and Rozinajová, V. 2022. *Automatic Question Generation Based on Sentence Structure Analysis Using Machine Learning Approach.* Natural Language Engineering, 28(4), 487–517.

[6] Ain, Q. U.; Chatti, M. A.; Bakar, K. G. C.; Joarder, S.; and Alatrash, R. 2023. Automatic Construction of Educational Knowledge Graphs: A Word-Embedding-Based Approach. *MDPI Journal*, 1–10.

[7] Guo, S.; Liao, L.; Li, C.; and Chua, T.-S. 2024. A Survey on Neural Question Generation: Methods, Applications, and Prospects. *IJCAI Conference Survey*, 1–25.

[8] Leite, B.; and Cardoso, H. L. 2023. Towards Enriched Controllability for Educational Question Generation. *arXiv preprint*, 1–9.

[9] Leung, J. 2022. An NLP Approach for Extracting Practical Knowledge from a CMS-based Community of Practice in E-Learning. *MDPI Journal of Educational Technology*, 1–12.

[10] Jung, Y.; and Choi, S. 2025. Knowledge Graph Construction: Extraction, Learning, and Evaluation. *MDPI Journal of Semantic Web Research*, 1–15.

[11] Lu, C.-Y.; and Lu, S.-E. 2021. A Survey of Approaches to Automatic Question Generation: from 2019 to Early 2021. *ACL Anthology Survey*, 1–18.

[12] Elkins, S.; Kochmar, E.; Cheung, J.; and Serban, I. 2023. How Useful are Educational Questions Generated by Large Language Models? *McGill NLP Report*, 1–11.

[13] Yang, R.; Yang, B.; Ouyang, S.; She, T.; Feng, A.; Jiang, Y.; Lecue, F.; Lu, J.; and Li, I. 2024. Graphusion: Leveraging Large Language Models for Scientific Knowledge Graph Fusion and Construction in NLP Education. *arXiv preprint*, 1–14.

[14] Cheng, C.; Huang, Z.; Zhao, G.; Guo, Y.; Lin, X.; Wu, J.; Li, X.; and Wang, S. 2025. From Objectives to Questions: A Planning-based Framework for Educational Mathematical Question Generation. *arXiv preprint*, 1–13.

[15] Oelen, A.; Stocker, M.; and Auer, S. et al. 2024. Creating and Validating a Scholarly Knowledge Graph Using Natural Language Processing and Microtask Crowdsourcing. *SpringerLink Journal of Knowledge Graphs*, 1–20.

[16] Qu, K.; Li, K.-C.; Wong, B.-T.-M.; Wu, M.-M.-F.; and Liu, M. et al. 2024. A Survey of Knowledge Graph Approaches and Applications in Education. *MDPI Education Knowledge Graph Journal*, 1–22.

[17] Zou, B.; Li, P.; Pan, L.; and Aw, A.-T. et al. 2022. Automatic True/False Question Generation for Educational Purpose. *ACL Anthology Workshop*, 1–8.

[18] Liang, S.; Kurt, S.; Farias, T.-M.; Anisimova, M.; and Gil, M. 2020. Querying Knowledge Graphs in Natural Language. *SpringerOpen Journal of Knowledge Representation*, 1–17.

[19] Liu, C.; Liu, K.; He, S.; Nie, Z.; and Zhao, J. 2019. Generating Questions for Knowledge Bases via Incorporating Diversified Contexts and Answer-Aware Loss. *NLPKG Conference*, 1–12.

[20] Narkudi, K.-V.; Juluri, H.-J.; and Inturi, S. 2025. Automatic Question Answer Generation Using NLP Techniques. *EasyChair Preprint*, 1–10.