Using Digital Textbooks as Knowledge Base to Detect Student Idea Development in Collaborative Discourse

Jiuning Zhong[1], Guangji Yuan[1], Jiangwei Zhang[1,] and Mei-Hwa Chen[1]

[1] University at Albany, State University of New York, Albany, 12203, NY, USA
{jzhong,gyuan,jzhang1,mchen}@albany.edu

Abstract. This paper presents a novel framework to monitor idea progression and novelty from learners' discourse by using digital textbooks as the knowledge base by means of knowledge graphs. A knowledge graph depicts an idea in a tripartite consisting of two concepts and their relations extracted from students' discourse and digital textbooks. A progressive mapping between the knowledge graphs extracted from the digital textbooks and the ones continuously collected from the students' discourse can be used to monitor the idea progression and detect idea novelty, which can greatly improve the teachers' and the students' awareness of learning outcome.

1. Introduction

Contemporary pedagogies encourage students to build deep knowledge in core curriculum areas through collaborative discourse and inquiry, making productive use of textbooks and other sources to scaffold, not to limit, their thinking. This study aims to design AI-empowered techniques to assess student idea development in collaborative online discourse in relation to core disciplinary concepts presented in textbooks, focusing on retrieving core concepts from textbooks and student discourse entries and analyzing the novelty of student ideas and questions.

Our approach draws upon a knowledge graph method to conduct automated textual analysis of learner discourse contributions. This paper presents an automated framework for extracting key concepts and ideas from textbooks and student online discourse, tracking student idea development using knowledge graph, and gauging idea novelty. The key elements of our framework include: (i) constructing and leveraging a knowledge base in the format of knowledge graphs using the triple units extracted from digital textbooks, (ii) constructing both personal (individual-centric) and collective (group-centric) temporal knowledge graphs from learner discourse for further idea analysis, and (iii) analyzing the progression of ideas and capturing novel ideas using multi-dimensional measures.

2. Knowledge Graph

We define a knowledge graph as $G = \{S, R, T\}$, where S, R, and T are sets of source entities, relations, and target entities, respectively. An idea from either online discourse or digital textbooks is a proposition with a triple semantic unit of two entities and the relation between them, where a proposition is denoted as a triple $(s, r, t) \in P$, which is presented by a union of two entities (source entity and target entity) and a relation, which forms a meaningful statement.

Our definition of an idea focuses not only on specific concepts but also on relations, which can also be extended to idea contexts in the form of a subgraph. Because, surrounding entities of an entity can represent the local information. Furthermore, our approach focuses on the relation between entities and also the context of ideas.

2.1 Idea Novelty

In collaborative discourse, students do not merely share and rephrase what they have learned from textbooks but make connected and non-redundant contributions to advance the group's knowledge and go beyond what they already know. They continually generate deeper questions and build on one another's ideas to develop higher levels of understandings. Thus, our analytics detect the novelty of student ideas posted in online discourse over time.

We define idea novelty as the extent to which an online post presents unique and relevant information that goes beyond what has already been posted in a temporal thread of conversation. The new contribution may be in the form of a new idea (thought) or question, and uniqueness may be gauged at a personal or group level. Novel ideas hidden in the discourse text have high dimensional textual features that can be evaluated using multidimensional measures. In addition, novelty can be out of knowledge base information or novel alias for existing information in the knowledge base. During online discussions, we monitor student idea development and identifying their novel ideas based on the mapping of constructed knowledge graphs. Our novelty rubric includes new concept, new question, new relation, and new context.

3. The Design of Novel Automated Framework

Our novel automated framework (shown in Figure 1) consists of two systems, online system and offline system, both sharing the Natural Language Understanding Component that processes raw unstructured textual data into triple units for knowledge graph construction. The offline system focuses on knowledge acquisition and knowledge base graph construction from digital textbooks and external sources on the web, which are highlighted in dark blue. The online system (highlighted in yellow) mainly supports many learners concurrently and connects with the novelty analysis component that uses incoming information from three knowledge graphs to conduct deeper analysis and sends the idea analysis feedback back to learners.

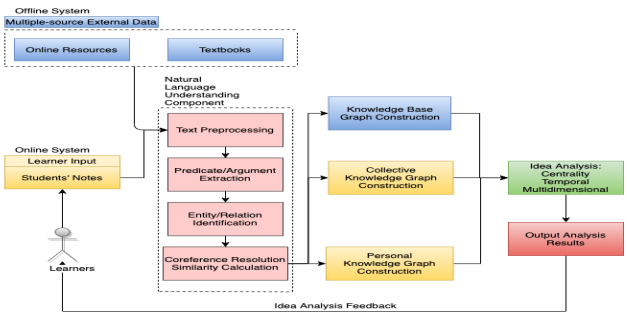

Fig. 1. A Novel Automated Framework of Idea Novelty Detection.

The significance of our novel automated framework is: 1) the framework connects each learner's input into collective and personal knowledge graphs simultaneously and joins with knowledge base graph for idea analysis; 2) to meet our novelty rubric, the idea analysis component relies on analyzing multidimensional features of the learner input, which include knowledge graph context feature, entity semantic related features, entity centrality feature, and temporal feature; 3) the output of the idea analysis has two parts, personal level and collective (group) level, which enables evaluation for individuals as well as for the whole class. The multidimensional features are described in the following:

Knowledge graph context feature: Given the constructed knowledge graph and new learner's input, we adopt the two-stage embedding scheme [2] that takes into account both contextual connectivity patterns and local connectivity patterns.

Entity semantic related features: Given the content of digital textbooks and learner input, our system constructs appropriate representations for entities and relations. To get a low-dimensional, continuous, and dense semantic representation, we apply entity word embeddings [1] from textbook content and learner notes. Given an entity pair, the semantic relevance of the two entities can be represented by their cosine similarity and their Euclidean distance in the vector space.

Entity Centrality feature: Centrality is shown to be intimately connected with the cohesive subgroup structure of a graph [4], which has been suggested as a good indicator of the importance of novelty.

Temporal feature: Decay operates on the assumption that the learner note in a discussion has a certain level of coherence, and therefore, show some cognitive continuity [3], and longer exposure of entities or relations would have diminishing influence on learners [5]. Each entity and relation contain the timestamp *t* of their creation, which would be the input of our novelty decay function.

We have been testing and refining our framework based on an extensive dataset of online discussions collected from a set of Grade 5 science classrooms. Figures 2 and 3 show the mapping between the knowledge base graph and the temporal knowledge graph generated based on student online discussion. In this example the knowledge base graph (shown in Figure 2) was constructed from textual data in a digital textbook: "Pass the Energy Please"; and the corresponding online discourse graph (shown in Figure 3) was generated based on a fifth-grade classroom discussion.

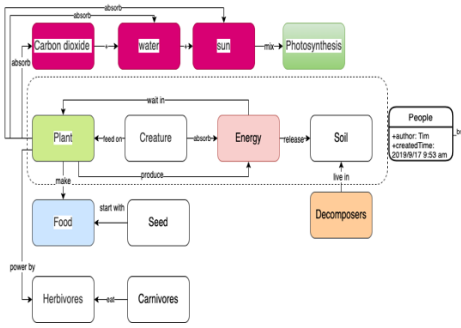

Fig. 2. Knowledge Base Graph.

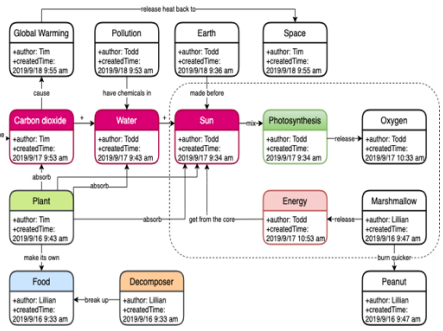

Fig. 3. Collective Temporal Graph.

When a triple unit is added to the graph, the framework checks if the entities and their relation exist on the knowledge base graph by applying distance-based and similarity matching based scoring functions to check entities. If there is a match, it selects and maps entities as highlighted in color on the two graphs, such as “Food” and “Plant” on the base graph. Otherwise, entities are not colored in the graph, like “Space” and “Global Warming” on the collective graph. Also, the new entities and relation are considered new concepts and relation. Each entity and relation on the temporal graph contain additional temporal information such as “createdTime” that can be used by the novelty decay function to calculate an importance score and author’s name for tracking individual learner’s idea progression. Meanwhile, a novel context detection function compares the subgraph of each matched entity against the neighbors of the same entity in the base graph in the format of embedding scheme [2] that takes into account both contextual connectivity patterns and local connectivity patterns to see if a context novelty exists, for instance, “energy” in the dashed line rectangle on the collective graph (Figure 3) has a novel context as the same entity has a distinguishing subgraph on the base graph. On the base graph, “energy” is around “soil”, “creature”, and “plant”, which indicates a context of food chain concept. While on the collective graph, “energy” is surrounded by entities as “Marshmallow”, “Sun”, “Photosynthesis”, “Oxygen”, which indicates a new and novel context about energy generation by photosynthesis and energy release.

4. Conclusion

We have proposed a novel automated framework consisting of a knowledge graph construction task from learner discussions and digital textbooks, an idea assessment task including tracking idea development and identifying novel ideas based on the mapping of knowledge graphs. Each important idea from digital textbooks or learner discourse is extracted as a semantic triple unit of a proposition with two entities and a relation between them and further used to construct the corresponding knowledge graphs. The idea reasoning task analyzes the progression of ideas and captures idea novelty from multidimensional measures.

During the construction of knowledge graphs, we faced many challenges, especially in subtasks like entity recognition and alignment, relation extraction due to colloquial language in learner discourse. We will apply additional training on the data and more advanced machine learning techniques to improve the precision and recall of the outcome. Upon all the work, we are creating visual abstractions of knowledge and ideas as knowledge graphs and multidimensional analytics of knowledge-building discourse that include idea progression, novelty, and digital textbook relevancy, which provide very valuable feedback to students and teachers.

References

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
2. Luo, Y., Wang, Q., Wang, B., & Guo, L. Context-dependent knowledge graph embedding. in EMNLP, pp. 1656–1661 (2015).
3. Liu, H., Lieberman, H., & Selker, T. A model of textual affect sensing using real-world knowledge. In Proceedings of the Seventh International Conference on Intelligent User Interfaces, pages 125–132 (2003).
4. Borgatti, S.P., Everett, M.G. A graph-theoretic perspective on centrality. Social Networks 28 (4), 466–484 (2006).
5. Feng, S., Chen, X., Cong, G., Zeng, Y., Chee, Y. M., & Xiang, Y. Influence maximization with novelty decay in social networks. In AAAI, pages 37–43 (2014).