A Relationship Selection Task*

Andrew E. Waters[1][0000-0002-9831-5435], Vinay K. Chaudhri[2][0000-0002-1363-645X], Debshila Basu Mallick[1][0000-0002-0597-3528], and Richard G. Baraniuk[1][0000-0002-0721-8999]

[1] OpenStax, Rice University, Houston TX 77005, USA
[2] Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

Abstract. An intelligent textbook is a traditional textbook enhanced with a knowledge graph (KG) making it a source of enhanced learning and instruction. The nodes of a KG are key terms in a textbook and the edges are the relationships between the terms. *Relationship selection* is the process of selecting the most appropriate relationship type between two different terms to be incorporated into a KG. We demonstrate a tool that allows learners to select relationships between terms embedded in a textbook sentence. We created this tool as a key component of a scalable infrastructure for KG construction through crowdsourcing of relationships between automatically extracted terms from a textbook. This task has the potential to be flexibly adapted to different textbooks and content domains. It is also suitable for encouraging relational processing and, we believe that it has instructional value. Therefore, our future work is focused on the pedagogical evaluation of the relationship selection task with students reading from a textbook.

Keywords: Knowledge Graph · Intelligent Textbooks · Relationship Selection · Concept Mapping

1 Introduction

Intelligent Textbooks (ITBs) using Artificial Intelligence (AI) and knowledge graphs (KG) allow students to dynamically interact with the textbook content, increasing their ability to understand concepts, raising engagement, and thereby, improving academic performance. Initial trials of ITBs that utilize KGs have been found to improve student grade outcomes by a full letter grade over the control group that was using a conventional textbook [3].

However, the process of constructing KGs that power such ITBs are time consuming and resource intensive. For example, the Inquire Biology ITB that was created by author Chaudhri for the popular introductory textbook, Campbell's Biology [10] required 5 person years from biology subject matter experts for knowledge engineering a KG for the first 10 chapters. Scaling this effort to all

* Supported by the National Science Foundation (NSF).
We thank Abhay Agarwal for deploying the task on AWS and making significant additions to the code documentation.

56 chapters of the textbook would have cost over $1.5M.

To create a scalable AI-crowdsourcing hybrid infrastructure for KG construction, we focused on three elements of the overall process. First, to capture the critical terms from textbook content, we used an adapted version of the BERT language deep learning model [5]. Second, to identify the relationships between term-pairs, we created a novel relationship selection task (RST) for crowdsourcing the relationships. Third, we developed de-noising methods to effectively fuse crowdsourced responses while accounting for task difficulty and participant competency. For the purposes of this demo, we highlight the development of the RST.

Relationship selection refers to the process of selecting the relationships between key terms in a textbook. To illustrate a relationship, consider the sentence *eukaryotic cells contain a nucleus.* In this example, the key terms are "eukaryotic cells" and "nucleus", and the relationship linking these two terms is "contains". Failure to identify important relationships affects the quality of a KG and ultimately limits the performance of the ITB.

The example relationships presented in Fig. 1 can be encoded computationally with tuples of the form (entity, relationship, entity), which is the standard structure used in knowledge graphs. The relationships needed for ITBs include taxonomy-based relationships (e.g., "prokaryotic cells" and "eukaryotic cells" are both *subset of* the class of "cells"), meronymic relationships (e.g., "nucleus" *is inside* "eukaryotic cell"), event structure relationships (e.g., "Anaphase" is a *sub step* of "mitosis") and causal relationships (e.g., "diffusion" *enables* "respiratory gas exchange") [2, 3].

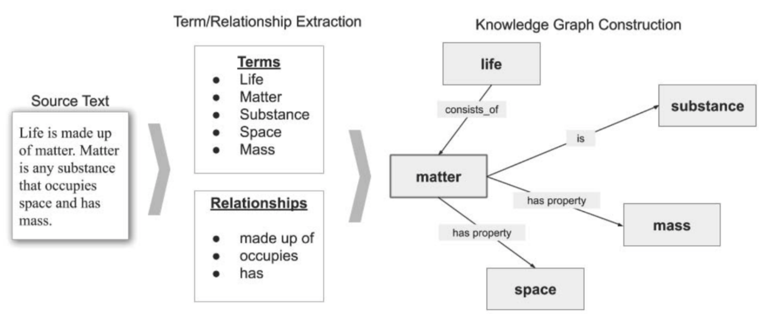

Fig. 1. Illustration of steps in Knowledge Graph (KG) construction. The process begins with extracting terms and then identifying relationships. Next, we connect the terms via relationships to build the KG.

There has been significant interest in the ML community in automatically identifying such relationships from natural language text. However, automatic extraction of relationships from text requires massive amounts of training data and rarely yields the high accuracy needed for an ITB. Therefore, our strategy was to develop a crowdsourcing task that could not only serve the purpose of creating the necessary relationship data, but also provide pedagogical benefits to student participants who are actively learning the material. To this end, we leveraged the popular educational task of concept mapping [9]). Concept mapping is an educational activity wherein a student takes individual concepts, represented as nodes, and defines the labeled edges between them. The task is ideal for present purposes, as the end product aligns with the KGs we ultimately hope to develop. Moreover, the process of creating the concept maps is believed to be beneficial for student learning [8]. Specifically, concept maps have been shown to be effective at promoting relational processing [7] during learning. Hence, we created a modified concept mapping activity, specifically designed to foster relational processing.

2 The development of the Relationship Selection Task

We describe the key steps in the development of RST: identifying relationship vocabulary, identifying terms, and developing a crowdsourcing tool.

2.1 Identifying an appropriate set of relationships

Our first step in this process was to choose an appropriate set of relationships also known as relationship vocabulary. Our relationship vocabulary is based on an upper ontology called Component Library or CLIB [1]. As CLIB was used extensively for constructing KB Bio 101 [4], we analyzed the most frequently used relationships. Such relationships included relations for describing the structure and function of entities, structure of processes, and causal relationships between processes. We were also informed by the empirical experience of the effectiveness of these relationships in practice as well as more recent work on linguistic analysis of relations [6]. For example, linguistic analysis suggests that some relationships from the CLIB are confusing, such as *agent, object*, and *base*. We replaced these confusing relationships with a general, but clearly understood relationship *participant*. The relationships we currently support include taxonomic relationships for classes and instances, structural relationships such as *has part* and *material*, spatial relationships such as *is inside* and *is above*, functional relationships such as *has function* and *facilitates*, event structure relationships, such as *subevent* and *next event*, and causal relationships such as *enables* and *prevents*. We also allow the possibility that no direct relationship may exist, as well as opportunities for crowd workers to define new relationships.

2.2 Automatically extracting terms and creating tasks

We used automated term extraction to identify all the terms in the textbook. As the precision and recall of the automated method is not perfect, the biologists on the team validated the terms. We then parsed the textbook section into individual sentences and automatically identified all term pairs that existed in each sentence. A sentence that contained N terms would have $\binom{n}{2}$ possible pairings, with each pairing considered to be a single task. After generating all possible tasks, we presented them to crowd workers using the tool that we describe next.

2.3 Developing the Crowdsourcing Tool

We developed a tool to guide the user in choosing the correct relationship between a pair of terms in the context of a sentence. The intended user of this tool is a student who does not have any formal training in knowledge engineering. During the development of the tool, we iteratively validated our designs through rapid prototyping with such users. The user is first asked to read a section from the textbook and then to undergo a short training on the relationships. We designed the training using simple common-sense examples that new users would find easy to understand. For example, we explain the *is inside* relationship using a visual in which a cat is shown hiding inside a box (Fig. 2). We developed similar illustrations for all the different relationships supported by the tool. After the training, the user completes a series of tasks through an interactive dialog to identify relationships between various textbook terms. All possible relationships between a pair of concepts can be extremely large. Some simple insights make our task tractable, namely: most term pairs are not related (i.e., the final graph is sparse), the terms that are connected are also likely to co-occur closely in the text, and that we can group the relationships into families so the user first chooses a relationship family before choosing the actual relationship (Fig. 3).

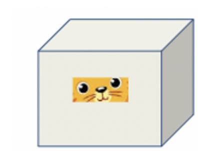

Fig. 2. Illustration of the *inside* relationship.

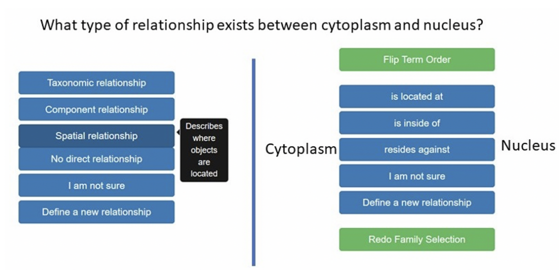

Fig. 3. Respondents on the Relationship Selection Task first select the correct family of relationship, followed by the actual relationship.

As a concrete example, consider this two-step selection in the dialog shown in Figure 3 where the user is asked to relate the terms "cytoplasm" and "nucleus". The user first chooses the appropriate relationship family for the terms, including taxonomic, spatial, and component-based relationships. The user further can select that the terms have no relationship between them, that they are unsure of the relationship, or that they would like to define a new relationship to relate the terms. In this example, the correct relationship family is a spatial relationship and clicking on this option takes them to a second set of options to specify which spatial family relationship is correct. In this dialog, they have an option to flip the order of the terms to ensure that the chosen relationship applies in the correct direction. Once they flip the order of terms, they can correctly indicate that the nucleus *is inside* cytoplasm.

3 Future directions

Using the RST, we successfully crowdsourced relationships between terms in sections of college level biology and psychology textbooks respectively, and are currently investigating its pedagogical efficacy.

References

1. Barker, K., Porter, B., Clark, P.: A library of generic concepts for composing knowledge bases. In: Proceedings of the 1st international conference on Knowledge capture. pp. 14–21 (2001)
2. Chaudhri, V.K., Cheng, B., Overtholtzer, A., Roschelle, J., Spaulding, A., Clark, P., Greaves, M., Gunning, D.: Inquire biology: A textbook that answers questions. AI Magazine **34**(3), 55–72 (2013)
3. Chaudhri, V.K., Dinesh, N., Inclezan, D., EDU, M.: Three lessons for creating a knowledge base to enable explanation, reasoning and dialog. In: Proceedings of the Second Annual Conference on Advances in Cognitive Systems ACS. vol. 187, p. 203. Citeseer (2013)
4. Chaudhri, V.K., Wessel, M.A., Heymans, S.: Kb bio 101: A challenge for tptp first-order reasoners. In: CADE-24 Workshop on Knowledge Intensive Automated Reasoning. Citeseer (2013)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Gisborne, N., Donaldson, J.: Thematic roles and events. In: The Oxford Handbook of Event Structure (2019)
7. Grimaldi, P.J., Poston, L., Karpicke, J.D.: How does creating a concept map affect item-specific encoding? Journal of Experimental Psychology: Learning, Memory, and Cognition **41**(4), 1049 (2015)
8. Nesbit, J.C., Adesope, O.O.: Learning with concept and knowledge maps: A meta-analysis. Review of educational research **76**(3), 413–448 (2006)
9. Novak, J.D., Cañas, A.J.: The origins of the concept mapping tool and the continuing evolution of the tool. Information visualization **5**(3), 175–184 (2006)
10. Reece, J.B., Meyers, N., Urry, L.A., Cain, M.L., Wasserman, S.A., Minorsky, P.V.: Campbell Biology Australian and New Zealand Edition, vol. 10. Pearson Higher Education AU (2015)