

Generating Concept Map Exercises from Textbooks

Andrew M. Olney, Whitney L. Cade, and Claire Williams

Institute for Intelligent Systems

University of Memphis

365 Innovation Drive, Memphis, TN 38152

aolney@memphis.edu

Abstract

In this paper we present a methodology for creating concept map exercises for students. Concept mapping is a common pedagogical exercise in which students generate a graphical model of some domain. Our method automatically extracts knowledge representations from a textbook and uses them to generate concept maps. The purpose of the study is to generate and evaluate these concept maps according to their accuracy, completeness, and pedagogy.

1 Introduction

Concept mapping is an increasingly common educational activity, particularly in K-12 settings. Concept maps are graphical knowledge representations that represent a concept, question or process (Novak and Canas, 2006). A recent meta-analysis of 55 studies involving over five thousand participants found the students *creating* concept maps had increased learning gains ($d = .82$) and students *studying* concept maps had increased learning gains ($d = .37$) (Nesbit and Adesope, 2006). In comparison, novice tutoring across many studies have had more modest learning gains ($d = .40$) (Cohen et al., 1982) – comparable to studying concept maps but not to creating them.

For difficult topics, or for students new to concept mapping, some researchers propose so-called expert skeleton concept maps (Novak and Canas, 2006). These are partially specified concept maps that may have some existing structure and then a “word bank” of concepts, properties, and relations that can be used to fill in the rest of the map. This

approach is consistent with concept maps as instructional scaffolds for student learning (O’Donnell et al., 2002). As students increase in ability, they can move from expert skeleton concept maps to self-generated maps.

Because concept maps are essentially knowledge representations based in words, analysis and synthesis of concept maps are theoretically amenable to knowledge-rich computational linguistic techniques. This paper presents an approach to extracting concept maps from textbooks to create educational materials for students. The concept maps can be used as expert skeleton concept maps. The rest of the paper is organized as follows. Section 2 presents a brief overview of concept maps from the AI, psychological, and education literatures and motivates a particular representation used in later sections. Section 3 presents a general technique for extracting concept maps from textbooks and generating graphical depictions of these as student exercises. Section 4 describes a comparative evaluation of maps extracted by the model to gold-standard human generated concept maps. Section 5 discusses these results and their significance for generating concept map exercises for students.

2 Perspectives on Concept Maps

There are many different kinds of concept maps, and each variation imposes different computational demands. One prominent perspective comes from the AI literature in formal reasoning, as an extension of work done a century ago by Pierce on existential graphs (Sowa, 2007; Sowa, 2009). In this formulation (which is now an ISO standard), so-called *con-*

ceptual graphs are interchangeable with predicate calculus. Of particular importance to the current discussion is *grain size*, that is the level of granularity given to nodes and relationships. In these conceptual graphs, grain size is very small, such that each argument, e.g. John, is connected to other arguments, e.g. Mary, through an arbitrary predicate, e.g. John *loves* Mary. Aside from the tight correspondence to logic, grain size turns out to be a relevant differentiator amongst conceptualizations of conceptual graphs amongst different fields, and one that leads to important design decisions when extracting graphs from a text.

Another prominent perspective comes from the psychology literature (Graesser and Clark, 1985), with some emphasis on modeling question asking and answering (Graesser and Franklin, 1990; Gordon et al., 1993). In this formulation of conceptual graphs, nodes *themselves* can be propositions, e.g. “a girl wants to play with a doll,” and relations are (as much as possible) limited to a generic set of propositions for a given domain. For example, one such categorization consists of 21 relations including *is-a*, *has-property*, *has-consequence*, *reason*, *implies*, *outcome*, and *means* (Gordon et al., 1993). A particular advantage of limiting relations to these categories is that the categories can then be set into correspondence with certain question types, e.g. definitional, causal consequent, procedural, for both the purposes of answering questions (Graesser and Franklin, 1990) as well as generating them (Gordon et al., 1993).

Finally, concept maps are widely used in science education (Fisher et al., 2000; Mintzes et al., 2005) for both enhancing student learning and assessment. Even in this community, there are several formulations of concept maps. One such widely known map is a hierarchical map (Novak and Canas, 2006; Novak, 1990), in which a core concept/question at the root of the map drives the elaboration of the map to more and more specific details. In hierarchical maps, nodes are not propositions, and the edges linking nodes are not restricted (Novak and Canas, 2006). Alternative formulations to hierarchical maps include cluster maps, MindMaps, computer-generated associative networks, and concept-circle diagrams, amongst others (Fisher et al., 2000).

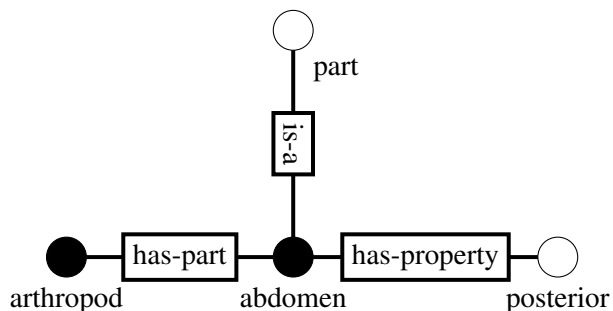


Figure 1: A concept map fragment. Key terms have black nodes.

Of particular interest is the SemNet formulation, which is characterized by a central concept (which has been determined as highly relevant in the domain) linked to other concepts using a relatively prescribed set of relations (Fisher, 2010). End nodes can be arbitrary, and cannot themselves be linked to unless they are another core concept in the domain. Interestingly, in the field of biology, 50% of all links are *is-a*, *part-of*, or *has-property* (Fisher et al., 2000), which suggests that generic relations may be able to account for a large percentage of links in any domain, with only some customization to be performed for specific domains. An example SemNet triple (start node/relation/end node) is “prophase *includes* process chromosomes become visible.” Several thousand of such triples are available online for biology, illustrating the viability of this representational scheme for biology (Fisher, 2010).

3 Computational Model

Our approach for extracting concept maps from a biology textbook follows the general SemNet formulation with some elements of the conceptual graphs of Graesser and Clark (1985). There are two primary reasons for adopting this formulation, rather than the others described in Section 2. By using a highly comparable formulation to the original SemNets, one can compare generated graphs with several thousand, expert-generated triples that are freely available. Second, by making just a few modifications to the SemNet formalism, we can create a formalism that is more closely aligned with question answering/question generation, which we believe is a fruitful avenue for future research.

Our concept map representation has two significant structural elements. The first is key terms, shown as black nodes in Figure 1. These are terms in our domain that are pedagogically significant. Only key terms can be the start of a triple, e.g. *abdomen is-a part*. End nodes can contain key terms, other words, or complete propositions. This structural element is aligned with SemNets. The second central aspect of our representation is labeled edges, shown as boxes in Figure 1. As noted by (Fisher et al., 2000), a small set of edges can account for a large percentage of relationships in a domain. Thus this second structural element aligns better with psychological conceptual graphs (Gordon et al., 1993; Graesser and Clark, 1985), but remains consistent with the spirit of the SemNet representation. The next sections outline the techniques and models used for defining key terms and edges, followed by our method of graph extraction.

3.1 Key Terms

General purpose key term extraction procedures are the subject of current research (Medelyan et al., 2009), but they are less relevant in a pedagogical context where key terms are often already provided in learning materials. For example, both glossaries (Navigli and Velardi, 2008), and textbook indices (Larrañaga et al., 2004) have previously been used as resources in constructing domain models and ontologies. To develop our key terms, we used the glossary and index from a textbook in the domain of biology (Miller and Levine, 2002) as well as the keywords given in a test-prep study guide (Cypress Curriculum Services, 2008). Thus we can skip the keyword extraction step of previous work on concept map extraction (Valerio and Leake, 2008; Zouaq and Nkambou, 2009) and the various errors associated with that process.

3.2 Edge Relations

Since edge relations used in conceptual graphs often depict abstract, domain-independent relationships (Graesser and Clark, 1985; Gordon et al., 1993), it might be inferred that these types of relationships, e.g. *is-a*, *has-part*, *has-property*, are exhaustive. While such abstract relationships may be able to cover a sizable percentage of all relationships previous work suggests new content can drive

new additions to that set (Fisher et al., 2000). In order to verify the completeness of our edge relations, we undertook an analysis of concept maps from biology.

Over a few hours, we manually clustered 4371 biology triples available on the Internet¹ that span the two topics of molecules & cells and population biology. Although these two topics represent a small subset of biology topics, we hypothesize that as the extremes of levels of description in biology, their relations will be representative of the levels between them.

Consistent with previous reported concept map research in biology (Fisher et al., 2000), our cluster analysis revealed that 50% of all relations were either *is-a*, *has-part*, or *has-property*. Overall, 252 relation types clustered into 20 relations shown in Table 1. The reduction from 252 relation types to 20 clusters generally lost little information because the original set of relations included many specific subclass relationships, e.g. *part-of* had the subclasses *composed of*, *has organelle*, *organelle of*, *component in*, *subcellular structure of*, *has subcellular structure*. In most cases subclassing of this kind is recoverable from information distributed across nodes. For example, if we know that *golgi body is-a organelle* and we know that *eukaryotic cell has-part golgi body*, then the original relation *golgi body organelle of eukaryotic cell* is implied.

Additional edge relations were added based on the psychology literature (Graesser and Clark, 1985; Gordon et al., 1993) as well as adjunct information gleaned from the parser described in the next section, raising the total number of edge relations to 30. As indicated by Table 1 a great deal of overlap exists between the clustered edge relations and those in the psychological literature. However, neither goal-oriented relationships nor logical relationships (and/or) were included as these did not seem appropriate for the domain (a cell divides because it must, not because it “wants to”). We also removed general relations that overlapped with more specific ones, e.g. *temporal* is replaced by *before*, *during*, *after*. We hypothesize that the edge relation scheme

¹<http://www.biologylessons.sdsu.edu>

Relation	Clustered	Gordon	Adjunct	Relation	Clustered	Gordon	Adjunct
after		*		has-consequence	*	*	*
before		*		has-part	*	*	
combine	*			has-property	*	*	
connect	*	*		implies		*	
contain	*	*		isa	*	*	
contrast	*			lack	*		
convert	*			location	*		*
definition	*			manner		*	*
direction			*	not			*
during	*	*		possibility			*
enable	*			produce	*		
example	*			purpose			*
extent			*	reciprocal			*
follow	*			require	*		
function	*			same-as	*	*	

Table 1: Edge relations from cluster analysis, Gordon et al. (1993), and parser adjunct labels

in Table 1 would be portable to other domains, but some additional tuning would be necessary to capture fine-grained, domain specific relationships.

3.3 Automatic Extraction

According to the representational scheme defined above, triples always begin with a key term that is connected by a relation to either another key term or a propositional phrase. In other words, each key term is the center of a radial graph. Triples beginning and ending with key terms bridge these radial graphs. The automatic extraction process follows this representational scheme. Additionally, the following process was developed using a biology glossary and biology study guide as a development data set, so training and testing data were kept separate in this study.

We processed a high school biology text (Miller and Levine, 2002), using its index and glossary as sources of key terms as described above, using the LTH SRL² parser. The LTH SRL parser is a semantic role labeling parser that outputs a dependency parse annotated with PropBank and NomBank predicate/argument structures (Johansson and Nugues, 2008; Meyers et al., 2004; Palmer et al., 2005). For each word token in a parse, the parser returns in-

²The Swedish “Lunds Tekniska Högskola” translates as “Faculty of Engineering”

formation about the word token’s part of speech, lemma, head, and relation to the head. Moreover, it uses PropBank and NomBank to identify predicates in the parse, either verbal predicates (PropBank) or nominal predicates (NomBank), and their associated arguments. A slightly abbreviated example parse corresponding to the concept map in Figure 1 is shown in Table 2.

In Table 2 the root of the sentence is “is,” whose head is token 0 (the implied root token) and whose dependents are “abdomen” and “part,” the subject and predicate, respectively. Predicate “part.01,” being a noun, refers to the Nombank predicate “part” roleset 1. This predicate has a single argument of type A1, i.e. *theme*, which is the phrase dominated by “of,” i.e. “of an arthropod’s body.” Predicate “body.03” refers to Nombank predicate “body” roleset 3 and also has a single argument of type A1, “arthropod,” dominating the phrase “an arthropod’s.” Potentially each of these semantic predicates represents a relation, e.g. *has-part*, and the syntactic information in the parse also suggests relations, e.g. ABDOMEN *is-a*.

The LTH parser also marks adjunct arguments. For example, consider the sentence “During electron transport, H⁺ ions build up in the intermembrane space, making it positively charged.” There are four adjuncts in this sentence: “During electron trans-

port” is a temporal adjunct, “in the intermembrane space” is a locative adjunct, “making it positively charged” is an adverbial adjunct, and “positively” is a manner adjunct. The abundance of these adjuncts led to the pragmatic decision to include them as edge relation indicators in Table 1.

After parsing, four triple extractor algorithms are applied to each sentence, targeting specific syntactic/semantic features of the parse, *is-a*, adjectives, prepositions, and predicates. Each extractor first attempts to identify a key term as a possible start node. The search for key terms is greedy, attempting to match an entire phrase if possible, e.g. “abiotic factor” rather than “factor,” by searching the dependents of an argument and applying morphological rules for pluralization. If no key term can be found, the prospective triple is discarded. Potentially, some unwanted loss can occur at this stage because of unresolved anaphora. However, it appears that the writing style of the particular textbook used, Miller and Levine (2002), generally minimizes anaphoric reference.

As exemplified by Figure 1 and Table 2, several edge relations are handled purely syntactically. The *is-a* extractor considers when the root verb of the sentence is “be,” but not a helping verb. *Is-a* relations can create a special context for processing additional relations. For example, in the sentence, “An abdomen is a posterior part of an arthropod’s body,” “posterior” modifies “part,” but the desired triple is *abdomen has-property posterior*. This is an example of the adjective extraction algorithm running in the context of an *is-a* relation: rather than always using the head of the adjective as the start of the triple, the adjective extractor considers whether the head is a predicate nominative. Prepositions can create a variety of edge relations. For example, if the preposition has part of speech IN and has a LOC dependency relation to its head (a locative relation), then the appropriate relation is *location*, e.g. “by migrating whales in the Pacific Ocean.” becomes *whales location in the Pacific Ocean*.

The predicates from PropBank and NomBank use specialized extractors that consider both their argument structure as well as the specific sense of the predicate used. As illustrated in some of the preceding examples, not all predicates have an A0. Likewise not all predicates have patient/instrument roles

like A1 and A2. Ideally, every predicate would start with A0 and end with A1, but the variability in predicate arguments makes simple mapping unrealistic. To assist the predicate extractors, we created a manual mapping between predicates, arguments, and edge relations, for every predicate that occurred more than 40 times in the textbook. Table 3 lists the four most common predicates and their mappings.

Predicate	Edge Relation	Start	End
have.03	HAS_PROPERTY	A0	Span
use.01	USE	A0	Span
produce.01	PRODUCE	A0	Span
call.01	HAS_DEFINITION	A1	A2

Table 3: Predicate map examples

The label “Span” in the last column indicates that the end node of the triple should be the text dominated by the predicate. Consider the example, “The menstrual cycle has four phases” has A0 *cycle* and A1 *phases*. Using just A0 and A1, the extracted triple would be *menstrual cycle has-property phases*. Using the span dominated by the predicate yields *menstrual cycle has-property four phases*, which is more correct in this situation. As can be seen in this example, end nodes based on predicate spans tend to contain more words and therefore have closer fidelity to the original sentence.

After triples are extracted from the parse, they are filtered to remove triples that are not particularly useful for generating concept map exercises. Filters are applied on the back end rather than during the extraction process because the triples discarded at this stage might be usefully used for other applications such as student modeling or question generation. The first three filters used are straightforward and require little explanation: the repetition filter, the adjective filter, and the nominal filter. The repetition filter considers the number of words in common between the start and end nodes. If the number of shared words is more than half the words in the end node, the triple is filtered. This helps alleviate redundant triples such as *cell has-property cell*. The adjective filter removes any triple whose key term is an adjective. These triples violate the assumption by the question generator that all key terms are nouns.

Id	Form	Lemma	POS	Head	Dependency Relation	Predicate	Arg 1	Arg 2
1	abdomen	abdomen	NN	2	SBJ	-	-	-
2	is	be	VBZ	0	ROOT	-	-	-
3	a	-	DT	5	NMOD	-	-	-
4	posterior	posterior	JJ	5	NMOD	-	-	-
5	part	part	NN	2	PRD	part.01	-	-
6	of	-	IN	5	NMOD	-	A1	-
7	an	-	DT	8	NMOD	-	-	-
8	arthropod	arthropod	NN	10	NMOD	-	-	A1
9	s	-	POS	8	SUFFIX	-	-	-
10	body	body	NN	6	PMOD	body.03	-	-
11	.	-	.	2	P	-	-	-

Table 2: A slightly simplified semantic parse

Has-property edge relations based on adjectives were also filtered because they tend to overgenerate. Finally the nominal filter removes all NomBank predicates except *has-part* predicates, since these often have Span end nodes and so contain themselves, e.g. *light has-property the energy of sunlight*.

The final filter uses likelihood ratios to establish whether the relation between start and end nodes is meaningful, i.e. something not likely to occur by chance. This filter measures the association between the start and end node using likelihood ratios (Dunning, 1993) and a χ^2 significance criterion to remove triples with insignificant association. As a first step in the filter, words from the end node that have low log entropy are removed prior to calculation. This penalizes non-distinctive words that occur in many contexts. Next, the remaining words from start and end nodes are pooled into bags of words, and the likelihood ratio calculated. By transforming the likelihood ratio to be χ^2 distributed (Manning and Schütze, 1999), and applying a statistical significance threshold of .0001, triples with a weak association between start and end nodes were filtered out. The likelihood ratio filter helps prevent sentences related to specific examples from being integrated into concept maps for a general concept. For example, the sentence “In most houses, heat is supplied by a furnace that burns oil or natural gas.” from the textbook is part of a larger discussion about homeostasis. An invalid triple implied by the sentence is *heat has-property supplied by a furnace*. Since *heat* and *furnace* do not have a strong association in

the textbook overall, the likelihood ratio filter would discard this triple.

After filtering, triples belonging to a graph are rendered to image files using the NodeXL³ graphing library. In each image file, a key term defines the center of a radial graph. To prevent visual clutter, triples that have the same edge type can be merged into a single node as is depicted in Figure 2.

4 Evaluation

A comparison study using gold-standard, human generated maps was performed to test the quality of the concept maps generated by the method described in Section 3. The gold-standard maps were taken from Fisher (2010). Since these maps cover only a small section of biology, only the corresponding chapters from Miller and Levine (2002), chapters two and seven, were used to generate concept maps. All possible concept maps were generated from these two chapters, and then 60 of these concept maps that had a corresponding map in the gold-standard set were selected for evaluation.

Two judges having background in biology and pedagogy were recruited to rate both the gold standard and generated maps. Each map was rated on the following three dimensions: the coverage/completeness of the map with respect to the key term (*Coverage*), the accuracy of the map (*Accuracy*), and the pedagogical value of the map (*Pedagogy*). A consistent four item scale was used for

³<http://nodexl.codeplex.com/>

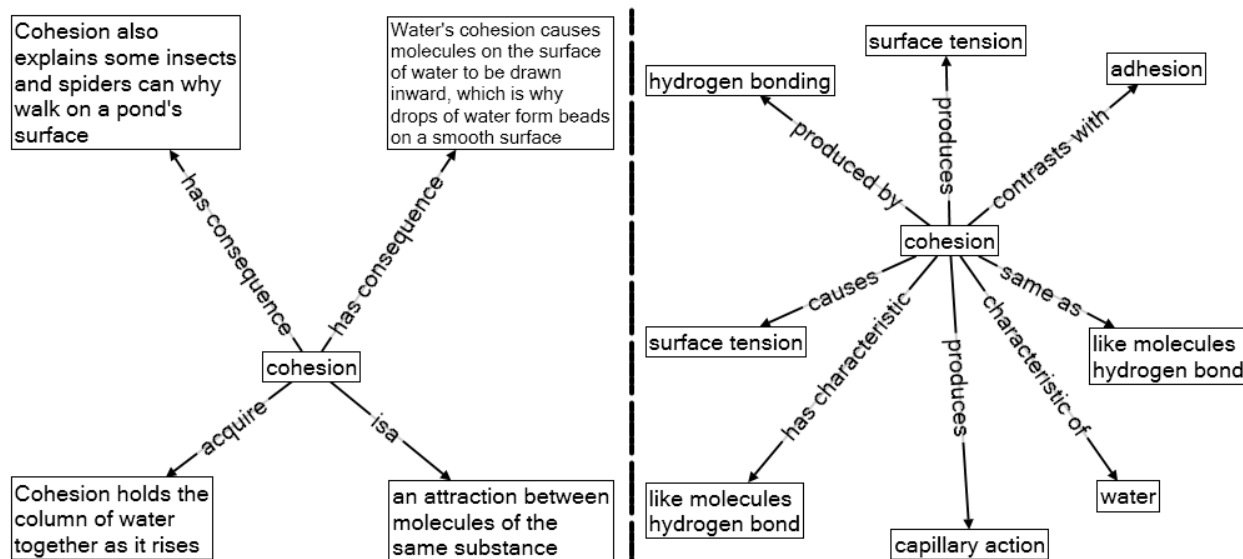


Figure 2: Comparison of computer and human generated concept maps for “cohesion.” The computer generated concept map is on the left, and the human generated map is on the right.

all ratings dimensions. An example of the four item scale is shown in Table 4.

Score	Criteria
1	The map covers the concept.
2	The map mostly covers the concept.
3	The map only slightly covers the concept.
4	The map is unrelated to the concept.

Table 4: Rating scale for coverage

Judges rated half the items, compared their scores, and then rated the second half of the items. Inter-rater reliability was calculated on each of the three measures using Cronbach’s α . Cronbach’s α is more appropriate than Cohen’s κ because the ratings are ordinal rather than categorical. A Cronbach’s α for each measure is presented in Table 5. Most of the reliability scores in Table 5 are close to .70, which is typically considered satisfactory reliability. However, reliability for accuracy was poor at $\alpha = .41$.

Scale	Cronbach’s α
Coverage	.75
Accuracy	.41
Pedagogy	.71

Table 5: Inter-rater reliability

Scale	Computer		Human	
	Mean	SD	Mean	SD
Coverage	2.47	.55	1.67	.82
Accuracy	1.87	.67	1.47	.55
Pedagogy	2.53	.74	1.83	.90

Table 6: Inter-rater reliability and mean ratings for computer and human generated maps

Means and standard deviations were computed for each measure per condition as shown in Table 6. In general, the means for the computer generated maps were in between 2 and 3 on the respective scales, while the human generated maps were between 1 and 2. The outlier is accuracy for the computer generated maps, which was significantly higher than for the other scales. However, since the inter-rater reliability for this scale was relatively low, the mean for accuracy requires closer analysis. Inspection of the individual means for each judge revealed that judge A had the same mean accuracy for both human and computer generated maps, ($M = 1.73$), while judge B rated the human maps higher ($M = 1.2$) and the computer generated maps lower ($M = 2$). Thus it is reasonable to use this more conservative lower mean, ($M = 2$), as the estimate of accuracy for the computer-generated concept maps.

Wilcoxon signed ranks tests pairing computer and human generated maps based on their key terms were computed for each of the three scales. There was a significant effect for coverage, $Z = 2.95$, $p < .003$, a significant effect for accuracy, $Z = 2.13$, $p < .03$, and a significant effect for pedagogy $Z = 2.46$, $p < .01$.

Since the purpose of the computer generated maps is to help students learn, pedagogy is clearly the most important of the three scales. In order to assess how the other scales were related to pedagogy, correlations were calculated. Accuracy and pedagogy were strongly correlated, $r(28) = .57$, $p < .001$. Coverage and pedagogy were even more strongly correlated, $r(28) = .86$, $p < .001$.

The strong relationship between coverage and pedagogy suggests that the number of the triples in the map might be strongly contributing to the judges ratings. An inspection of the number of triples in the human maps compared to the computer generated maps reveals that there are approximately 3.5 times as many triples in the human maps as the computer generated maps. To further explore this relationship, a linear regression was conducted using the log of number of triples in each graph to predict the mean pedagogy score for that graph. The log number of triples in a graph significantly predicted pedagogy ratings, $b = -.96$, $t(28) = -3.47$, $p < .002$. The log number of triples in the graph explained a significant proportion of variance in pedagogy ratings, $r^2 = .30$, $F(1, 28) = 12.02$, $p < .002$.

These results are encouraging on two fronts. First, the computer generated maps are on average “mostly accurate.” Secondly, the computer generated maps fare less well for coverage and pedagogy, but these two scale are highly correlated, suggesting that judges are using a criterion largely based on completeness when scoring maps. The strength of the log number of triples in a graph as a predictor of pedagogy likewise indicates that increasing the number of triples in each graph, which would require access to a larger sample of texts on these topics, would increase the pedagogical ratings for the computer generated maps. However, while gaps in the maps would be problematic if the students were using the maps as an authoritative source for study, gaps are perfectly acceptable for expert skeleton concept maps.

5 Conclusion

In this paper we have presented a methodology for creating expert skeleton concept maps from textbooks. Our comparative analysis using human generated concept maps as a gold standard suggests that our maps are mostly accurate and are appropriate for use as expert skeleton concept maps.

Ideally student concept maps that extend these skeleton maps would be automatically scored and feedback given as is already done in intelligent tutoring systems like Betty’s Brain and CIRCSIM Tutor (Biswas et al., 2005; Evens et al., 2001). Both of these systems use expert-generated maps as gold standards by which to evaluate student maps. Therefore automatic scoring of our expert skeleton concept maps would require a more complete map in the background.

In future work we will examine increasing the number of knowledge sources to see if this will increase the pedagogical value of the concept maps and allow for automatic scoring. However, increasing the knowledge sources will also likely lead to an increase not only in total information but also in redundant information. Thus extending this work to include more knowledge sources will likely require incorporating techniques from the summarization and entailment literatures to remove redundant information.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080594 and by the National Science Foundation, through Grant BCS-0826825, to the University of Memphis. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education or the National Science Foundation.

References

- Gautam Biswas, Daniel Schwartz, Krittaya Leelawong, and Nancy Vye. 2005. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19:363–392, March.
- Peter A. Cohen, James A. Kulik, and Chen-Lin C. Kulik. 1982. Educational outcomes of tutoring: a meta

- analysis of findings. *American Educational Research Journal*, 19:237–248.
- LLC Cypress Curriculum Services. 2008. *Tennessee Gateway Coach, Biology*. Triumph Learning, New York, NY.
- Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, March.
- Martha W. Evens, Stefan Brandle, Ru-Charn Chang, Reva Freedman, Micheal Glass, Yoon Hee Lee, Leem Seop Shim, Chong Woo Woo, Yuemei Zhang, Yujian Zhou, Joel A. Michael, and Allen A. Rovick. 2001. CIRCSIM-Tutor: An intelligent tutoring system using natural language dialogue. In *Proceedings of the 12th Midwest AI and Cognitive Science Conference (MAICS 2001)*, pages 16–23, Oxford, OH.
- Kathleen M. Fisher, James H. Wandersee, and David E. Moody. 2000. *Mapping biology knowledge*. Kluwer Academic Pub.
- Kathleen Fisher. 2010. Biology Lessons at SDSU. <http://www.biologylessons.sdsu.edu>, January.
- Sallie E. Gordon, Kimberly A. Schmierer, and Richard T. Gill. 1993. Conceptual graph analysis: Knowledge acquisition for instructional system design. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 35(3):459–481.
- Arthur C. Graesser and Leslie C. Clark. 1985. *Structures and procedures of implicit knowledge*. Ablex, Norwood, NJ.
- Arthur C. Graesser and Stanley P. Franklin. 1990. Quest: A cognitive model of question answering. *Discourse Processes*, 13:279–303.
- Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic-semantic analysis with PropBank and NomBank. In *CoNLL '08: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 183–187, Morristown, NJ, USA. Association for Computational Linguistics.
- Mikel Larrañaga, Urko Rueda, Jon A. Elorriaga, and Ana Arruarte Lasa. 2004. Acquisition of the domain structure from document indexes using heuristic reasoning. In *Intelligent Tutoring Systems*, pages 175–186.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327, Singapore, August. Association for Computational Linguistics.
- Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.
- Kenneth R. Miller and Joseph S. Levine. 2002. *Prentice Hall Biology*. Pearson Education, New Jersey.
- Joel J. Mintzes, James H. Wandersee, and Joseph D. Novak. 2005. *Assessing science understanding: A human constructivist view*. Academic Press.
- Roberto Navigli and Paola Velardi. 2008. From glossaries to ontologies: Extracting semantic structure from textual definitions. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 71–87, Amsterdam, The Netherlands, The Netherlands. IOS Press.
- John C. Nesbit and Olusola O. Adesope. 2006. Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3):413–448.
- Joeseeph D. Novak and Alberto J. Canas. 2006. The theory underlying concept maps and how to construct them. Technical report, Institute for Human and Machine Cognition, January.
- Joeseeph D. Novak. 1990. Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, 27(10):937–49.
- Angela O'Donnell, Donald Dansereau, and Richard Hall. 2002. Knowledge maps as scaffolds for cognitive processing. *Educational Psychology Review*, 14:71–86. 10.1023/A:1013132527007.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.
- John F. Sowa. 2007. Conceptual graphs. In F. Van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of knowledge representation*, pages 213–237. Elsevier Science, San Diego, USA.
- John F. Sowa. 2009. Conceptual graphs for representing conceptual structures. In P. Hitzler and H. Scharfe, editors, *Conceptual Structures in Practice*, pages 101–136. Chapman & Hall/CRC.
- Alejandro Valerio and David B. Leake. 2008. Associating documents to concept maps in context. In A. J. Canas, P. Reiska, M. Ahlberg, and J. D. Novak, editors, *Proceedings of the Third International Conference on Concept Mapping*.
- Amal Zouaq and Roger Nkambou. 2009. Evaluating the generation of domain ontologies in the knowledge puzzle project. *IEEE Trans. on Knowl. and Data Eng.*, 21(11):1559–1572.