Extraction of Concept Maps from Textbooks for Domain Modeling

Andrew M. Olney**

University of Memphis, Memphis TN 38152, USA,
aolney@memphis.edu,
WWW home page: http://iis.memphis.edu

Abstract. Previous research using concept maps as domain models in intelligent tutoring systems has demonstrated their power and flexibility. However, these concept maps must still be authored by a domain expert, creating a development bottleneck. We present a new, streamlined methodology for automatically extracting concept maps from textbooks using term extraction, semantic parsing, and relation classification.

Key words: concept map, domain model, semantic parsing

1 Introduction

In an intelligent tutoring system (ITS), the representation of subject matter knowledge is often referred to as a domain model. The domain model is an integral part of an ITS and is typically strongly connected both the model of the student's knowledge (student model) and the model of how to teach the subject matter (pedagogical/expert model). As exemplified by CIRCSIM-Tutor and Betty's Brain, concept maps can be used as both domain models and overlay student models as well as to interpret student utterances, generate explanations, and perform qualitative reasoning [1, 2]. However, in both CIRCSIM-Tutor and Betty's Brain, expert concept maps need to be authored. In this paper we outline a new approach to concept map extraction from textbooks. Our approach is significantly streamlined relative to previous approaches [3, 4]. In what follows we define our concept map representation and then describe our approach to extracting concept maps from a textbook.

2 Definitions

Our concept map definition is a blend of previous work in the psychology and education literatures [5–7]. We adopt a formulation of concept maps largely

** The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080594 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.

consistent with the SemNet formulation [7]. This leads to maps with one layer of links radiating out of a core concept. In our representation, only key terms can be the start of a triple (equivalently the center of a map). End nodes can contain key terms, other words, or complete propositions. In addition, our representation uses a restricted set of labeled edges. As noted by Fisher [7], a small set of edges account for a large percentage of relationships. Moreover, having a prescribed set of edges facilitates linkages between the map representation and question asking/answering [5, 6].

While generalized term extraction procedures exist [8], such methods tend to be less relevant in a pedagogical context where key terms are often already provided, whether in glossaries [9], textbook indices [10], study guides, etc. To develop our key terms, we used the glossary and index from a textbook as well as the keywords in a test-prep study guide, yielding approximately 3,000 terms. Thus we can skip the keyword extraction step of previous work on concept map extraction [3, 4].

We obtained general relations from previous work [5–7, 11] but augmented these with relations specific to our biology domain through manual analysis. We manually analyzed and clustered 4371 biology triples available on the Internet[1] to a set of 30 relations, including *after, contrast, enable, has-consequence, lack, produce, before, convert, example, has-part, location, purpose, combine, definition, extent, has-property, manner, reciprocal, connect, direction, follow, implies, not, require, contain, during, function, isa, possibility,* and *same-as*. Previous approaches to concept map extraction do not appear to use a fixed set of relations [3, 4], making them less suited to question asking/answering techniques [5, 6].

3 Extraction

Given a textbook as input, the LTH SRL Parser [11] outputs a dependency parse annotated with semantic roles derived from Propbank and Nombank, i.e. part of speech, lemma, head, and relation to the head, verbal predicates, nominal predicates, and associated arguments. Use of a semantic parser significantly streamlines the extraction of concept maps, which previous approaches have addressed using syntactic parsing with regular expressions, or word/relation extraction through phrase identification [3, 4].

For each syntactic or semantic relation found by the parser, we require that the start node be a key term. Several relations are handled purely syntactically, including *is-a* via "be," *has-property* via adjectives, and *location* via prepositions. Relations from Propbank and Nombank require examination of several features in order to determine the relationship between the arguments, including the lexical form, the gloss for the roleset of the predicate, the label given to the argument, and the gloss given to the argument. These features are input to a manually designed decision tree, which inspects the features by priority and assigns a relation.

[1] http://www.biologylessons.sdsu.edu

Using this approach, we extracted 28,994 relations from a thousand page textbook. These relations were distributed around 1,886 key terms out of approximately 3,000. The mean number of relations per term is 15.4, but the variation is quite high (min 1, max 552, sd 31.7). The five most connected key terms are *animals, cell, species, genes,* and *blood*. We extracted 27 relations of 30, excluding *lack*, *requires*, and *same-as*. The top five relations extracted were *has-property has-consequence, isa, manner, and location*, making up roughly 80% of the total relations. The relations *has-property*, *is-a*, and *has-part* make up 52% of the total, which is consistent with reported human concept maps for biology domains [7]. Our future work will expand our evaluation of the raw maps as well as their application to question asking/answering and domain models.

References

1. Evens, M., Brandle, S., Chang, R., Freedman, R., Glass, M., Lee, Y., Shim, L., Woo, C., Zhang, Y., Zhou, Y., Michael, J., Rovick, A.: CIRCSIM-Tutor: An intelligent tutoring system using natural language dialogue. In: Proceedings of the 12th Midwest AI and Cognitive Science Conference (MAICS 2001), Oxford, OH (2001) 16–23
2. Leelawong, K., Biswas, G.: Designing learning by teaching agents: The Betty's Brain system. Int. J. Artif. Intell. Ed. **18**(3) (2008) 181–208
3. Zouaq, A., Nkambou, R.: Evaluating the generation of domain ontologies in the knowledge puzzle project. IEEE Trans. on Knowl. and Data Eng. **21**(11) (2009) 1559–1572
4. Valerio, A., Leake, D.B.: Associating documents to concept maps in context. In Canas, A.J., Reiska, P., Ahlberg, M., Novak, J.D., eds.: Proceedings of the Third International Conference on Concept Mapping. (2008)
5. Graesser, A.C., Franklin, S.P.: Quest: A cognitive model of question answering. Discourse Processes **13** (1990) 279–303
6. Gordon, S., Schmierer, K., Gill, R.: Conceptual graph analysis: Knowledge acquisition for instructional system design. Human Factors: The Journal of the Human Factors and Ergonomics Society **35**(3) (1993) 459–481
7. Fisher, K., Wandersee, J., Moody, D.: Mapping biology knowledge. Kluwer Academic Pub (2000)
8. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, Association for Computational Linguistics (August 2009) 1318–1327
9. Navigli, R., Velardi, P.: 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, Amsterdam, The Netherlands, The Netherlands, IOS Press (2008) 71–87
10. Larrañaga, M., Rueda, U., Elorriaga, J.A., Lasa, A.A.: Acquisition of the domain structure from document indexes using heuristic reasoning. In: Intelligent Tutoring Systems. (2004) 175–186
11. Johansson, R., Nugues, P.: Dependency-based syntactic-semantic analysis with PropBank and NomBank. In: CoNLL '08: Proceedings of the Twelfth Conference on Computational Natural Language Learning, Morristown, NJ, USA, Association for Computational Linguistics (2008) 183–187