Toward automatic inference of causal structure in student essays

Peter Hastings[1]*, Simon Hughes[1], Anne Britt[2], Dylan Blaum[2], and Patty Wallace[2]

[1] DePaul University, Chicago, Illinois
[2] Northern Illinois University, DeKalb, Illinois

Abstract. With an increasing focus on science and technology in education comes an awareness that students must be able to understand and integrate scientific explanations from multiple sources. As part of a larger project aimed at deepening our understanding of student processes for integrating multiple sources of information, we are developing machine learning and natural language processing techniques for evaluating students' argumentative essays. In previous work, we have focused on identifying conceptual elements of the essays. In this paper, we present a method for inferring the causal structure of student essays. We used a standard parser to derive grammatical dependencies of the essay and converted them to logic statements. Then a simple inference mechanism was used to identify concepts linked to syntactic connectors by these dependencies. The results suggest that we will soon be able to provide explicit feedback that enables teachers and students to improve comprehension.

Keywords: Reading, Argumentation, Natural language processing, Machine learning

1 Introduction

Recent science and literacy standards are increasing the demand for students to use multiple sources of information to understand explanations for phenomena and to use data to support these explanations. Thus, there is critical need for methods of evaluating students' explanations and argumentative support based on scientifically important criteria (e.g., coherence, completeness, and accuracy).

A scientific explanation, also called a causal chain, is a statement that makes clear how one or more factors lead to an outcome. For example, in Figure 1 below, the to-be-explained outcome is an "increase in recent average global temperatures", and there are two separate initiating factors (fossil fuel consumption and deforestation). It is expected that students need practice to become more

* The assessment project described in this article is funded, in part, by the Institute for Education Sciences, U.S. Department of Education (Grant R305G050091 and Grant R305F100007). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

facile using an explanation schema to guide both writing and reading. It would be very helpful for teachers to have a tool that supports student practice with feedback to help them develop this explanation schema. As a first step, we are examining whether we can automatically identify the causal structure of student essays in two different scientific domains.

This paper describes previous research done on this task, and then presents more fully the educational context of the current work. Then we describe our ongoing research in using machine learning to identify the conceptual elements of essays, and our initial efforts toward inferring causal structure.

2 Previous research

Although causal explanations have long been a focus for science education [15,3, for example], very little research has been done to automatically identify causal connections in student essays, but there has been some research with other types of texts. In 1987, Cohen [4] laid out a theoretical framework encompassing the many different challenges that need to be solved to fully understand argumentative discourse. Thirty years later, a SemEval-2007 workshop focused on sentences known to have one of seven different types of relations, including causation [7]. Accuracy in distinguishing betweeen the seven types ranged from 50 to 76%.

More recently, Rink et al. developed a system focused on identifying the presence or absence of a causal relationship within a sentence [12]. They used a graph representation of the sentence and trained a machine learning technique on 700 sentences (30% with a causal relation) to distinguish graphs with and without causal connections. Their best F_1 score was 0.39. This was on news texts rather than student essays but clearly demonstrates the difficulty of the task.

3 Educational context

To deepen our understanding of students' comprehension processes, we created two document sets describing the causes of two scientific phenomena: global warming and coral bleaching. Each document set was based on a causal model of the scientific phenomenon and used information from reputable websites (e.g. the United States Geological Survey). Each document contributed only a partial causal chain. Students were given a document set and asked to write an essay explaining the phenomenon using specific information from the documents to support their conclusions and ideas. A total of 183 middle (84%) and high school (16%) students wrote essays on the global warming document set, and 105 middle (73%) and high school (27%) students wrote essays on the coral bleaching document set.

As mentioned above, Figure 1 gives a graphical representation of the space of causal connections that students might make from the documents to the conclusion in the global warming domain. Thin black arrows indicate explicit connections made in the documents. Dotted lines indicate implicit connections — inferences that students might make between concepts. Red lines represent

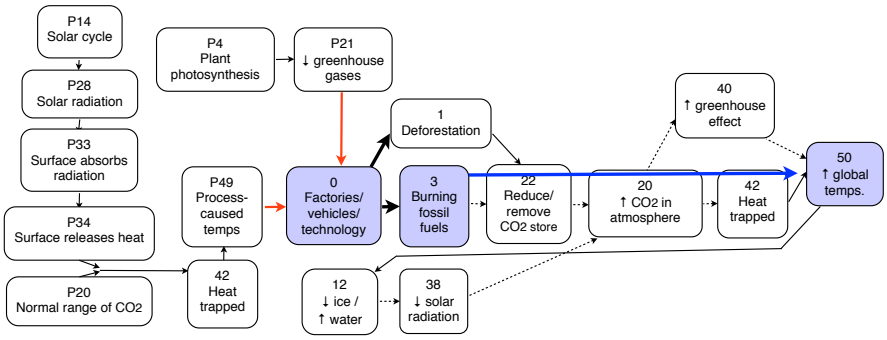

Fig. 1. Causal model with feedback for Global Warming

"counters", for example, "Normal temperature shifts happen *but* our use of cars and factories changes things." This graph also provides an example of how automatic assessment of the essays might be used to provide feedback to students. The thick black arrows mark explicit connections that were identified in the student's essay. The thick blue arrow shows where the student made a causal connection to the conclusion, but skipped some intermediate causal links. For explicit feedback, the student could be shown the graph to provide an indication of what was found and what was missing from her essay. For less guiding feedback, the student could be told that she has identified some links in the causal chain, but has omitted others.

Humans evalated the essays to identify which causes (nodes in Figure 1) were explicitly linked to the target effect (here, increase in global temperatures). Inter-rater reliability was high ($\kappa = 0.85$), and the method was useful in discriminating essays that provided coherent and complete answers. There was a difference in annotation for the two sets of essays. The global warming essays were annotated at the sentence level. Each sentence was associated with a set of codes indicating the concepts and causal connections found in that sentence. The coral bleaching essays were annotated later with a more sophisticated tool (brat.nlplab.org), identifying which specific words in the essay were associated with each concept and connection.

4 Concept identification

In previous work, we evaluated several different techniques for identifying conceptual material (i.e., the nodes in the graphs) in student essays, including simple pattern matching, latent semantic analysis (LSA), and support vector machines (SVMs) [8,9,10]. In general, we have found that the machine learning approaches do best at identifying high-level claims and specific details about the claims. Student sentences associated with these items tend to bear a striking similarity to

the original texts that they came from.[3] The machine learning techniques have had a much more difficult time identifying conceptual material related to inferences between documents. Examples of these items are rather infrequent in the students' essays (explaining why we need a system like this). They also combine information (and, therefore, words) from different documents and are thus less similar to the original sources [9]. We have recently begun evaluating a new machine learning approach, Deep Learning [13,1,5], which uses multilayer neural networks, but details of this approach are omitted due to space limitations.

5 Inferring causal connections

Once the conceptual content of an essay has been identified, the next step required for automatic structure evaluation is to find where the essay makes explicit connections between the concepts. For this step, it is clear that a "bag-of-words" approach would be severely handicapped because it would not be able to take advantage of the critical information provided by the linguistic structure of the text. To capture this structure, we applied the Stanford Compositional Vector Grammar parser [14] from Stanford CoreNLP (v.3.3.1) to tokenize and parse the essays and identify coreference relations [11].

We were particularly interested in taking advantage of the dependencies that the parser identifies in the text [6]. Dependencies are textual relations that are extracted from the parse tree, connecting different components. For example, the sentence, "The fat dog was chased by a cat," produces (among others) dependencies indicating that "fat" is an adjectival modifier for "dog", "dog" is the passive nominal subject of "chased", "cat" is the agent of "chased", and "chased" is the root of the sentence.

To enable inference of causal connections, we transformed the dependencies into clauses in Prolog, because Prolog seems especially well suited for specifying complex constraints. To evaluate the identification of connections independently from the identification of concepts, we started from the human annotations of concepts and connectors,[4] which were also converted into Prolog clauses.

A total of five Prolog rules were used to do the inference. Three of them handle different forms of representation. One of the two main inference rules searches for dependencies between connectors and *causes*, looking at three dependency types. The other rule looks for dependencies between connectors and *results*, looking at 7 types of dependencies.

[3] In fact, 25 – 30% of the student essay sentences had an LSA cosine greater that 0.75 with some sentence from the relevant document set. Ironically, this facilitates our job of classifying the student sentences. The effect on student learning, however, is subtle (analysis forthcoming).

[4] We do not include connectors as concept codes, but they are a critical part of identifying causal relations. Fortunately, students use fairly standard connectors. In coral bleaching, for example, of 134 coded connectors, 32 were "because (of)" and 15 had some variation of "cause". The rest, though less frequent, followed standard conventions.

Using this minimal inference mechanism, we calculated Recall, Precision, and F_1 scores, based on the whether the *inferred* causal connections matched the *annotated* ones. On the coral bleaching essays, the scores were: $Recall = 0.26$, $Precision = 0.59$, and $F_1 = 0.36$. On the 30 (out of 183) global warming essays we have fully annotated, this method achieved $Recall = 0.37$, $Precision = 0.53$, and $F_1 = 0.44$. At this early stage, the results are very encouraging. This technique outperformed the most similar previous research on inferring causal connections (although we did have the advantage of pre-identified concepts and connectors). Also, given that the Precision scores are high relative to the Recall scores, more sophisticated inference rules should be able to find items that our simple rules missed without producing too many false alarms.

6 Conclusions and future work

Clearly the work presented here is in its early stages, but the results so far have been extremely encouraging. Even though we have artificially boosted our results by starting with human-annotated concept codings, our very simple mechanism for identifying causal relations has already outperformed previous approaches. We are pursuing several different directions that should bring us closer to our ultimate goal of fully automatic causal relation identification so that we can provide reliable feedback to teachers and students.

With respect to concept classification, greedy sequence classification [2] could be used where a sequential classifier is trained to incorporate the tag it predicted for the previous word when tagging the next word. Neural Network Language Models (NNLMs) have recently become very popular due to their ability to learn a distributed representation for words at the same time as creating a language model to predict the likelihood of a sequence of words [1]. However little work has been done to investigate their use in creating sequential classifiers. An NNLM could be used to create a sequential classifier that predicts the concept tag for the central word in a word window instead of predicting the likelihood of the central word.

Another critical component for identifying causal relations is anaphora resolution. Students often use pronouns to refer to previously mentioned concepts. In the coral bleaching domain, 10% of the identified causal relations involved a pronominal reference. Because we included the human annotations for references in our evaluation, we were able to correctly identify a comparable percentage of causal relations with and without anaphora. As mentioned above, the Stanford CoreNLP parser returns coreference relations in addition to the dependencies. If these are reliable for student essays, they should allow us to successfully automate identification of relations across sentences.

The current inference rules for identifying causal relations are quite simple. It is quite likely that the hit-rate of these rules can be significantly improved by adding more dependencies, although it may well be that additional constraints are necessary to avoid over-generalization. We will also explore the use of machine learning techniques like Rink et al. used to automatically derive new inference

rules [12]. Finally, we are collecting additional student essay data in these and in new scientific explanation domains. This will support cross-domain validation of our techniques, to ensure that they can produce robust results.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. Journal of Machine Learning Research 3, 1137–1155 (2003)
2. Bird, S., Klein, E., Loper, E.: Natural Language processing with Python Analyzing Text with the Natural Language Toolkit. O'Reilly (2009)
3. Chi, M., Roscoe, R., Slotta, J., Roy, M., Chase, C.: Misconceived causal explanations for emergent processes. Cognitive Science 36, 1–61 (2012)
4. Cohen, R.: Analyzing the structure of argumentative discourse. Computational Linguistics 13(1-2), 11–24 (1987)
5. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Cohen, W., McCallum, A., Roweis, S. (eds.) ICML. vol. 307, pp. 160–167. ACM (2008)
6. de Marneffe, M., Manning, C.: The Stanford typed dependencies representation. In: COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation (2008), http://nlp.stanford.edu/pubs/dependencies-coling08.pdf
7. Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., Yuret, D.: Semeval-2007 task 04: Classification of semantic relations between nominals. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). p. 1318 (2007), http://acl.ldc.upenn.edu/W/W07/W07-2003.pdf
8. Hastings, P., Hughes, S., Magliano, J., Goldman, S., Lawless, K.: Text categorization for assessing multiple documents integration, or John Henry visits a data mine. In: Biswas, G., Bull, S. (eds.) Proceedings of the 15th International Conference on Artificial Intelligence in Education (2011)
9. Hastings, P., Hughes, S., Magliano, J., Goldman, S., Lawless, K.: Assessing the use of multiple sources in student essays. Behavior Research Methods 44(3), 622–633 (2012)
10. Hughes, S., Hastings, P., Magliano, J., Goldman, S., Lawless, K.: Automated approaches for detecting integration in student essays. In: Cerri, S., Clancey, W., Papadourakis, G., Panourgia, K. (eds.) Proceedings of Intelligent Tutoring Systems 2012 (2012)
11. Recasens, M., de Marneffe, M.C., Potts, C.: The life and death of discourse entities: Identifying singleton mentions. In: HLT-NAACL. pp. 627–633. The Association for Computational Linguistics (2013)
12. Rink, B., Bejan, C.A., Harabagiu, S.M.: Learning textual graph patterns to detect causal event relations. In: Guesgen, H.W., Murray, R.C. (eds.) FLAIRS Conference. AAAI Press (2010)
13. Socher, R., Pennington, J., Huang, E., Ng, A., Manning, C.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: EMNLP. pp. 151–161. ACL (2011)
14. Socher, R., Bauer, J., Manning, C.D., Ng, A.Y.: Parsing with compositional vector grammars. In: ACL (1). pp. 455–465. Association for Computer Linguistics (2013)
15. White, B., Frederiksen, J.: Causal model progressions as a foundation for intelligent learning environments. Artificial Intelligence 42, 99–157 (1990)