Machine Learning for Holistic Evaluation of Scientific Essays

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

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

Abstract. In the US in particular, there is an increasing emphasis on the importance of science in education. To better understand a scientific topic, students need to compile information from multiple sources and determine the principal causal factors involved. We describe an approach for automatically inferring the quality and completeness of causal reasoning in essays on two separate scientific topics using a novel, two-phase machine learning approach for detecting causal relations. For each core essay concept, we initially trained a window-based tagging model to predict which individual words belonged to that concept. Using the predictions from this first set of models, we then trained a second stacked model on all the predicted word tags present in a sentence to predict inferences between essay concepts. The results indicate we could use such a system to provide explicit feedback to students to improve reasoning and essay writing skills.

Keywords: Reading, Argumentation, Causal Relation, Natural Language Inference, Machine Learning, Natural Language Processing, NLP

1 Introduction

Educational standards in the US have increased considerably in accordance with the Common Core standards and Next-Generation science standards [15, 19]. These standards call for a focus on comprehending and evaluating science models, theories, explanations, and evidence, and learning from multiple documents and representation formats. However, middle and high school students and even many undergraduates have difficulty learning from multiple documents in science or history [4, 25]. One explanation is students fail to develop an adequate schema for this genre to guide their reasoning. According to Kintsch and Van Dijk [16], effective integration and summarization requires learning genre-specific macro-rules that differentiate the more salient information from the less.

* The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100007 to University of Illinois at Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education

To explain scientific phenomena, one must explain how a sequence of causal factors leads to an outcome via a sequence of intervening concepts — a causal chain. For example, in Figure 1 we can see how the 2 separate initiating factors of DECREASE IN TRADE WINDS and STORMS/RAINFALL lead via two different causal chains to the outcome of CORAL BLEACHING. A schema for such a process would thus involve slots for the initiating factors, concepts, the final outcome and for the causal links by which the factors and concepts drive the outcome.

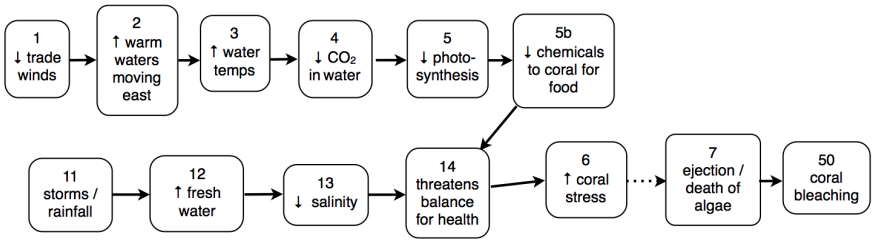

Fig. 1. Causal model for coral bleaching

Essay writing is an important learning skill, and has been shown to promote deeper understanding and integration of information from different sources [3, 26]. Thus, as part of a much larger project[3][7], we are examining student comprehension of scientific explanations (e.g., explain how and why coral bleaching rates vary at different times) from multiple documents of a variety of types (e.g., descriptive texts, images, graphs and maps) as measured primarily from an essay written with the available documents. A complete and coherent explanation requires integrating information from the entire document set to form a causal model of the phenomena. To improve reading and writing skills in this domain, students need practice developing schemas for understanding this sort of material. This study investigates the use of machine learning to map written essays to a complete causal model of the essay topic. This technique can then lead to the creation of tools that can help guide students in the development of such schemas to aid comprehension and develop better writing and reasoning skills.

2 Related Work

Explaining causality has long been a focus in science education [24, 5, for example]. However, little research has been carried out to automatically detect causal inference in essays. In 1987, Cohen [6] described a theoretical framework detailing

[3] The assessment was created by the READI science design team which includes D. Blaum, M. A. Britt, W. Brown, C. Burkett, M. George, S. R. Goldman, C. Greenleaf, K. James, M. Ko, K. Lawless, S. Marple, U. Sexton, P. Wallace, & M. Yukhymenko.

the different problems that need to be addressed to understand argumentative discourse.

Most work in this area has focused on the use of surface lexical features and syntactic patterns to extract causal relations from open domain text. Previous work was mostly restricted to certain sub-types of causal relations involving noun phrases [2], such as between nominals, and between nouns and verbs only. In 2002, Girju and Moldovan [8] used lexico-syntactic patterns to detect a particular form of causal relation between two noun phrases of the form <NP1 Verb NP2>, where the verb was a simple causative. They achieved an accuracy of 65.6% compared to the average of two human annotators. In Semeval 2007, Task 4 focused on "the classification of semantic relations between nominals" [9], which included the detection of causal relations. The highest F_1 score on this category was 0.82 with an accuracy of 77.5% on a system that used features based on WordNet, VerbNet, lexico-syntactic features and parse features. To move beyond surface lexical features, in 2014 Riaz and Girju [21], used Integer Linear Programming to combine verb and noun semantics with the predictions of a supervised machine-learning model trained only on linguistic features. They focused on detecting only causal relations between noun and verbs, achieving an F_1 score of 0.41 and an accuracy of 80.73%, a 15% improvement over using linguistic features alone. In a very different approach, Rink et al [22] built a system to detect the presence or absence of causality only in sentences containing verbal events joined with a conjunction. They created graphical models of causal sentences encoding syntactic and hypernymy information, and dependences from a dependency parser. They then extracted sub-graph patterns that occured in causal sentences and used a constraint satisfaction solver to detect these patterns from new sentences, attaining an F_1 score of 0.39.

In contrast, our work does not restrict the type of causal relation based on the syntactic categories involved. However, we only concern ourselves with the causal relations defined in the causal model, which can take the form of any combination of syntactic categories observed in the essays, but are restricted to relations relevant to the essay topics. We also investigate automatic inference of full causal chains, a topic that has received little attention in the literature.

3 The Essay Annotation Procedure

Two document sets assess students abilities to integrate information and develop an understanding about two scientific phenomena: coral bleaching and skin cancer. The documents were prepared from reputable sources (e.g. the NASA earth observatory, the US Geological Survey, and online science textbooks), and each started with some short background material to provide framing, necessary vocabulary, and relevant background knowledge. The sources were compiled such that the students needed to combine information from multiple sources to fully answer the question.

In conjunction with the development of the document sets, a causal model of each scientific phenomenon was created (see Figure 1) that represents the rel-

evant scientific phenomena in the source documents and the causal connections between them, from initiating factors (e.g., decreased trade winds, storms and rainfall, decreased salinity) via various intervening concepts to the final outcome (coral bleaching, or increased skin cancer rates). This can be thought of as a representation of the causal structure of the ideal essay according to the viewpoint of the researchers. There are 2 possible full causal chains in each model, each starting with a different initiating factor but resulting in the same final outcome.

Each student was provided with the essay prompt (e.g. explain the causes of coral bleaching), and asked to answer using the source material. 105 middle and high-school students were assigned 2 essay questions. The essays were then annotated by two different annotators according to how well they aligned with the corresponding causal model. Inter-rater reliability was high ($\kappa = 0.85$). Words or phrases indicating concepts from the model were tagged, and causal links were made between them where it was explicitly stated in the essay. For example "How coral are bleaching because the water temperature is increasing, the solubility of carbon dioxide (CO_2) in water decreases" would be coded as concept 3 (INCREASING WATER TEMPS) causing concept 4 (DECREASING CO_2 IN WATER), which causes concept 50 (CORAL BLEACHING). Here the student missed the intervening concepts in the chain, going straight from concept 4 to 50.

4 The Tagging Problem: Identifying Concept Codes

4.1 Previous Work

In previous work, we experimented with a number of different machine learning techniques to detect the core concepts and claims in student essays, [10, 12, 14], including a support vector machine (SVM), a regular expression learning algorithm [10, 14], and the k-nearest neighbor algorithm run on Latent Semantic Analysis (LSA) projections of the sentences. Under this approach, we found that training a separate binary classifier per code was more effective than a multi-class classifier trained on all codes due to the degree of semantic overlap between concept codes in the domain.

In some of these previous studies, we were only able to utilize annotations on individual sentences, although many of the concept codes only covered a few words or short phrases within each sentence. For instance, in the coral bleaching domain, the most common concept code is the concept CORAL BLEACHING, which is most often referred to using exactly those terms, or "coral whitening", or similar. When such a phrase occurs in a sentence with 10 or 20 other words, it is difficult for the algorithm to learn which words denote the particular phrase. Another limitation to our previous work was the use of a bag of word representation for the sentences fed to the SVM and k-nearest neighbor classifiers. This ignores word order, and means the classifier cannot distinguish between individual words and phrases. The regular expression learner was able to learn multi-word phrases over the sentences, but that had lower accuracy on most tasks. In this work, we use a tagging model which overcomes these 2 limitations.

Curriculum learning is the technique of training machine learning models on gradually harder training examples, and lowers the generalization error in some problems [1]. We take a similar approach for detecting causality. We first train a set of tagging models, one per concept, to learn tags for individual words. We then feed the predicted tags into a second set of classification algorithms to solve a much harder problem, detecting inferences within a sentence. We use the same approach to detect concept codes at the sentence level, as this allows the classifier to utilize information about the presence of other tags in the sentence.

4.2 A Window-Based Tagger

Tagging models are commonly used in natural language processing to tag words with one or more labels, such as in identifying parts of speech tags such as nouns or verbs, or in named entity recognition. To effectively determine the concept code of a word in this domain, it's important to include the surrounding context of the word as that denotes the word's meaning and can disambiguate its word sense. For this reason, we feed a window of words around the target word to an SVM to predict the concept code for the individual word. Each word and position forms a separate feature, so if the word 'melanin' occurs in window position 1 and in position 5, those are treated as separate features. We experimented with a range of window sizes, 1, 3, 5, 7 and 9 and found 7 to be optimal. In addition, we used all bi-grams (consecutive word pairs) appearing in the window as additional features, again encoded with their relative window position. We experimented with trigrams, but that lowered tagging accuracy. Table 1 shows the recall, precision, F_1 score and classification accuracy for the word-level tagging task on the validation data using 5 fold cross-validation. The mean values were computed over all the concept codes, the weighted mean was weighted by the number of occurrences of each code.

Table 1. Word tagging results

	Coral Bleaching			Skin Cancer		
	Recall	Precision	F_1	Recall	Precision	F_1
Mean	0.57	0.67	0.58	0.62	0.72	0.66
Weighted Mean	0.70	0.80	0.73	0.70	0.75	0.72

In preparing the data, we replaced words occurring only once in each document set with a special token, padded the start and ends of the sentence with special start and stop tokens, and used a spelling corrector based on [18] to correct typos. We used stemming only on the coral bleaching data as it reduced accuracy on the skin cancer data set. We did not remove stop words as they can aid classification accuracy when tagging words based on their context. In previous work [11] we compared the performance of Linear Discriminant Analysis, an

SVM, Random Forest and a Decision Tree on this task and found the SVM to be superior in terms of F_1 score, and so that technique was used in this paper.

In addition to training the tagging model to detect concept codes, we also trained it on 3 additional tags, "Causal Tags", associated with the causal chains - CAUSER, RESULT and EXPLICIT. CAUSER was an additional tag applied to concepts denoted as causal. Similarly, RESULT was applied to concepts denoted as the result of a causal link. EXPLICIT describes words or phrases that link CAUSE concepts with RESULT concepts and often includes terms denoting causality, such as "can cause", "because of" and "affects". Predictions for these 3 tags, along with the other concept codes from the tagger were then used to build a predictive model to predict which sentences contained causal links.

5 Sentence-Level Concepts and Causality

Causal links within the essays occur within a sentence, and as such cannot be detected using a tagging model alone, as they occur between concepts. To detect causal links, we used a form of stacking, or "stacked generalization" [27], to train a binary classifier on the output of the tagging models. Having trained a different tagging model for each of the codes, including the Causal Tags, we first aggregated the word-level predictions at the sentence level in 3 different ways to produce 3 sets of features. (1) For each code, we took the maximum distance either side of the margin (positive and negative) learned by the SVM classifier over all words in the sentence. This measures how strongly parts of each sentence are positive and negative exemplars of each code. (2) We created a binary feature for each code that was true if any of the words were tagged with that code. (3) We took all the unique codes predicted over the sentence and created a feature for each unique pairwise combination of codes. We experimented with several other sets of features, including computing the mean distance from the margin per word, computing 3-way combinations of predicted codes, and features representing the order in which pairs of codes appeared in the sentence. However, these three feature sets proved optimal in terms of classification accuracy. A separate binary SVM classifier was then trained using these three feature sets as input to predict three types of causal relation for each sentence: Cause-Explicit-Result, Cause-Explicit and Explicit-Result, as well as the individual concept codes at the sentence level. The last two causal relationships represent incomplete cause-effect relations. Table 2 shows the classification accuracy at the sentence level for the three relations and the average classification accuracy over all the non-causal concept codes on the validation data using 5-fold cross-validation.

6 Evaluating Writing Quality

In prior work, we have used tools to automatically identify core concepts [10, 12, 14], and, starting with human scoring of core concepts, automatically identify causal chains [11]. As the logical next step, we identified four levels of explanation

Table 2. Sentence level classification accuracy

	Coral Bleaching			Skin Cancer		
	Recall	Precision	F_1	Recall	Precision	F_1
Cause-Explicit	0.64	0.66	0.65	0.66	0.61	0.61
Explicit-Result	0.61	0.65	0.63	0.68	0.62	0.63
Cause-Explicit-Res.	0.63	0.68	0.65	0.67	0.62	0.62
Concept Mean	0.74	0.72	0.71	0.81	0.77	0.78
Concept Weighted	0.88	0.84	0.85	0.88	0.84	0.85

quality that capture general goals for an explanation (e.g., accuracy, completeness, coherence). When reading, especially multiple documents, readers' goals determine what is relevant which, in turn, influences how information is processed [17, 23]. Therefore, we selected categories that could point to feedback that could be used to help students refine the goals for the task.

The four quality levels were (1) No core content, (2) No causal chains, (3) Causal chain with no intervening factors, (4) Chain with intervening. The "No core content" essays did not have any core concepts other than the final outcome of the causal chain that was given. Students who received no credit for core concepts generally focused on statements that were supporting but not part of the explanation, or were too vague. Feedback for these students could encourage them to begin to identify elements of the explanation and to make their statement of the concept more explicit and complete. For example, merely saying that "wind affects the water" does not help the reader distinguish wind conditions that would lead to coral bleaching from those that would lead to healthy coral. Students may not understand the importance of attending to directional modifiers. In the "No causal chains" essays, students focused on at least one important element of the causal model but did not explicitly connect this information to the final outcome. These students are reproducing some correct concepts but could be instructed how to connect these concepts via intervening concepts to the final outcome. The final two types of essays actually have some degree of structure that is required in the essay question. The difference is whether there is some success in connecting initiating factors and intervening concepts. Students' writing that included a causal chain but with no intervening concepts could be encouraged to examine whether concepts across documents could be connected as intervening causes.

To assess entire essays and for the pedagogically appropriate categories, we had to convert binary causal predictions into predictions for specific concepts, and aggregate the predictions over all the sentences in each essay. For this, we employed simple heuristics. For each sentence in which a causal connection was predicted, we assumed that it was between the first two codes (numerically) identified in the sentence. When there are exactly two concepts in the sentence, this heuristic works well. If there are 1 or 0 codes identified, it results in a partial chain (which also happens in student essays). If there are more than two concepts, this results in additional, unconnected concept codes. For example,

if the system predicted a causal connection and codes 1, 2 and 50, this was identified by the heuristics as a causal chain between 1 and 2 with an extra 50.

Aggregating the sentence codes for an essay required resolving vague concepts and linking chains together. For each of the concept codes, the human coders could mark it as "vague" if it only partially matched the concept. But we only want to give an essay credit for claims which are fully specified. Because many writers start out general and get more specific, when aggregating, we converted vague concepts to non-vague if they were specified fully elsewhere in the essay.

Multiple metrics were defined to evaluate the relative completeness of each student's essay, and to compute the quality level defined above. The evaluation of the inferred causal chains with respect to these pedagogically useful attributes is shown in Table 3. The first four numerical columns show the correspondence between four measures computed from the human coders and those from the machine learning approach: the number of unique codes in the essay, the number of unique causal chains, the maximum chain length, and the number of distinct paths leading to the final outcome. The Quality column shows the correspondence between the explanation quality level assigned by the coders and by the system. Because these attributes are interval and ordinal (Quality) we measured correspondence with accuracy, adjacent accuracy (which includes misses by 1), and Krippendorff's alpha (agreement), instead of Recall, Precision, and F_1.

Table 3. Essay-level accuracy and correlations

	Codes	Chains	Length	Paths	Quality
	Coral Bleaching				
Accuracy	0.57	0.62	0.63	0.63	0.51
Adjacent	0.91	0.90	0.85	0.91	0.85
Agreement	0.89	0.37	0.20	0.36	0.56
	Skin Cancer				
Accuracy	0.38	0.41	0.48	0.45	0.43
Adjacent	0.87	0.90	0.81	0.93	0.88
Agreement	0.82	0.34	0.27	0.33	0.47

It is important to note that each of these measures has a relatively wide range of values (0–9, 0–4, 0–5, 0–3, and 1–4, respectively), so the expected accuracy of random guessing for any of them would be from 0.1 to 0.25. As these results show, the aspects of the classification that rely only on inference of codes are fairly reliable. Even if the machine learning identification of codes doesn't exactly match the human calculation, the agreement is high. The rest of the attributes and the quality level rely on identifying the causal connections, which is notoriously hard, and that is reflected in our results. Nevertheless, the high adjacent accuracies make us confident that feedback generated on the basis of these evaluations would be helpful for the students.

7 Conclusions and Further Work

Prior work in this area [2, 8, 9, 21, 20, 22] has demonstrated the difficulty of automatically inferring causality from text, and has primarily focused on detecting sub-types of causal relations. Instead we focused on a particular domain, and our approach was able to detect partial and full causal chains, a topic without much precedence in the literature. Furthermore, we presented a novel scoring rubric for assessing the accuracy, completeness and coherence of the explanations present in the essays. We then demonstrated reasonable accuracy in automating this assessment task. This approach has the potential to be used to develop intelligent tutoring systems to assist students with the development of mental schema to help them refine their goals for this task, and thus aid comprehension, reasoning skills and writing ability. For example, if a student skips a section of the causal chain, such as going from concept 4 to 50 as described in section 3, a tutoring system could show the graph of the causal model, highlighting the missing inferences to guide the student to produce a more complete essay.

There are a few limitations to our approach that could be addressed in future work. A limitation of the window-based tagging model is that it can only consider relationships between words that occur together within a window, and cannot consider longer-term dependencies. Long-Short-Term-Memory Recurrent Neural Networks can learn long-term dependencies between sequential items [13], and could be a promising future direction for this work. We also plan to explore the use of coreference resolution to help identify causal connections that are linked anaphorically across sentences. Finally, it would be interesting to investigate whether the causal patterns learned by our system could predict causal relations in open domain text.

References

1. Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.
2. E. Blanco, N. Castell, and D. Moldovan. Causal relation extraction. In *LREC*, 2008.
3. M. A. Britt, P. Wiemer-Hastings, A. Larson, and C. Perfetti. Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education*, 14:359–374, 2004.
4. M.A. Britt and C. Aglinskas. Improving students' ability to identify and use source information. *Cognition and Instruction*, 20(4):485–522, 2002.
5. M. Chi, R. Roscoe, J. Slotta, M. Roy, and C. Chase. Misconceived causal explanations for emergent processes. *Cognitive Science*, 36:1–61, 2012.
6. R. Cohen. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13(1-2):11–24, 1987.
7. Institute for Education Sciences. Reading for understanding across grades 6 through 12: Evidence-based argumentation for disciplinary learning. washington, d.c.: National center for education research, 2010. retrieved from `http://www.ies.ed.gov/ncer/projects/results.asp?ProgID=62&NameID-351` last accessed 2015-01-20.

8. R. Girju and D. Moldovan. Mining answers for causation questions. In *Proc. The AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.
9. R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, page 1318, 2007.
10. P. Hastings, S. Hughes, J. Magliano, S. Goldman, and K. Lawless. Text categorization for assessing multiple documents integration, or John Henry visits a data mine. In *Proceedings of the 15th AIED Conference*, 2011.
11. Peter Hastings, Simon Hughes, Anne Britt, Dylan Blaum, and Patty Wallace. Toward automatic inference of causal structure in student essays. In *Intelligent Tutoring Systems*, pages 266–271. Springer, 2014.
12. Peter Hastings, Simon Hughes, Joseph Magliano, Susan Goldman, and Kimberly Lawless. Assessing the use of multiple sources in student essays. *Behavior Research Methods*, 44(3):622–633, 2012.
13. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
14. Simon Hughes, Peter Hastings, Joseph Magliano, Susan Goldman, and Kimberly Lawless. Automated approaches for detecting integration in student essays. In S. Cerri, W. Clancey, G. Papadourakis, and K. Panourgia, editors, *Proceedings of Intelligent Tutoring Systems 2012*, 2012.
15. Achieve Inc. *Next Generation Science Standards*. Achieve Inc., 2013.
16. W. Kintsch and T. A. Van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394, 1978.
17. S. Lehman M. McCrudden, G. Schraw and A. Poliquin. The effect of causal diagrams on text learning. *Contemporary Educational Psychology*, 33:367–388, 2007.
18. How to write a spelling corrector. `http://norvig.com/spell-correct.html`.
19. The Council of Chief State School Officers. The common core standards for english language arts and literacy in history/social studies and science and technical subjects. Washington, DC: National Governors Association for Best Practices, 2010. `http://www.corestandards.org`.
20. Mehwish Riaz and Roxana Girju. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 161, 2014.
21. Mehwish Riaz and Roxana Girju. Recognizing causality in verb-noun pairs via noun and verb semantics. *EACL 2014*, page 48, 2014.
22. Bryan Rink, Cosmin Adrian Bejan, and Sanda M. Harabagiu. Learning textual graph patterns to detect causal event relations. In Hans W. Guesgen and R. Charles Murray, editors, *FLAIRS Conference*. AAAI Press, 2010.
23. J. F. Rouet and M. A. Britt. Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, and G. Schraw, editors, *Text Relevance and Learning from Text*. Information Age Publishing, Greenwich, CT, in press.
24. B. White and J. Frederiksen. Causal model progressions as a foundation for intelligent learning environments. *Artificial Intelligence*, 42:99–157, 1990.
25. J. Wiley, S.R. Goldman, A. Graesser, C. Sanchez, I. Ash, and J. Hemmerich. Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal*, 46(4):1060–1106, 2009.
26. J. Wiley and J. F. Voss. Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, 91:301–311, 1999.
27. David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.