Identifying the structure of students' explanatory essays

Simon Hughes[1 [0000−0002−7923−3506]⋆], Peter Hastings[1 [0000−0002−0183−001X]], and M. Anne Britt[2 [0000−0003−2328−4096]]

[1] DePaul University School of Computing
[2] Northern Illinois University Psychology Department

Abstract. Recent educational standards stress that students should learn how to read and understand scientific explanations and create explanations of their own. But these skills are difficult for teachers to evaluate, so they often assess them at a shallow level or avoid giving such assignments. Previous approaches for automatically evaluating explanatory and other types of structured essays have relied on the use of shallow features or bag-of-words methods. These methods might allow for a reasonable holistic assessment of an essay, but they fail to identify which concepts students included and which causal connections they made. In this paper, we investigate which natural language processing methods are most successful at locating conceptual information in student explanations and the causal connections between them. We found that a combination of a recurrent neural network for identifying concepts along with a novel causal relation parser produced very good accuracy in two different scientific domains, significantly improving on the prior state-of-the-art.

1 Introduction

The US Common Core standards and the Next-Generation Science Standards reflect an increasing emphasis in education on how important it is for students to learn how to read and comprehend science theories, models, and explanations, integrate information from multiple sources, and to create their own explanations [6, 1]. Teachers often find it challenging to evaluate such texts in more than a cursory manner [13, 22]. Automated Essay Scoring mechanisms could be used to reduce the load on teachers, but they tend to rely on surface-level features of text aggregated across the essay [14] or bag-of-words approaches like LSA [8], correlated with expert scores or pre-scored essays. These approaches are not sophisticated enough to identify the structure of the students' explanations. In other words, they cannot determine which components of an ideal explanation

⋆ The assessment project described in this article was 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.

the students have included, and how they have connected them together. In this paper, we attempt to determine the optimal natural language processing (NLP) techniques for identifying conceptual information and causal relations in explanations, including a novel relation-parsing method.

2 Materials

As part of a larger project focused on understanding students' reading processes, approximately 1,300 14–15 year old students in a large U.S. city were asked to read a small set of documents on a particular scientific phenomenon and create their own explanation of that phenomenon. Two different topics were used: skin cancer and coral bleaching. Each student worked with both topics. There were 5 documents of less than 1 page for each topic. One gave a general overview and the others gave related information including images, maps, and charts. With input from topic experts, a causal model was created for each topic, indicating the important concepts described in the documents and the causal connections made between them. The coral bleaching causal model included 13 concepts, and the skin cancer model had 9.

Over 1100 student essays were collected for each topic. The `brat` tool [23, 24] was used to annotate word spans as concepts and explicit connections between them as causal relations. Inter-rater reliability was high, with κ values of 93%. In the next sections, we present the evaluation of several successful NLP techniques for identifying the concepts and causal relations in the essays.

3 Concept Identification

The five techniques we compared have each been previously shown to produce state-of-the-art results on various NLP tasks. Each had different representational approaches to handling the challenges of ambiguity in text, interrelationships between words, and relative probabilities of classification. We compared the approaches using micro-averaged F_1 scores, because they capture performance with the relative frequencies of the codes in actual texts. All were tested with 5-fold cross validation

- **Window-based taggers** [21, for example] classify an item using that item and features about its neighboring items as inputs. Previously, we evaluated a window-based method with an SVM classifier, yielding an F_1 score of 0.73 [10]. Here, we extended that approach, finding the best performance by using logistic regression on positional stemmed unigrams, non-positional unigrams, Brown cluster labels [3], and dependency parser relations.
- A **Conditional Random Field** (CRF) [12] learns a graphical model which constitutes a linear chain of probabilities, expressing relationships between random variables [12]. We used the CRFSuite [16] implementation and trained the model with the L-BFGS gradient descent method.

– A **Hidden Markov Model** (HMM) is also a linear chain probabilistic model [18, 25], but it is a *generative* model. It learns to predict the probability of observing a particular *word* based on the label from the training set and the label of the previous word.
– A **Structured Perceptron** was used to perform multi-class classification [5, for example]. Being an online model allows this approach to more easily incorporate its own previous predictions as features to predict the next label in the sentence.
– A **Recurrent Neural Network** learns to build its own representation as it iterates through the words in a sentence [7]. We used the bi-directional Gated Recurrent Unit (GRU) variant of RNN, with 100-dimensional GloVe embeddings [17] as inputs. The best-performing network followed the inputs with two bi-directional GRU layers of 256 units, then a softmax output layer, and it was trained with the Adam optimizer [11].

The performance metrics for the five different concept identification methods on the testset are shown in the top of Table 1. Averaging across topics, the RNN performed best. In comparison with previous results, the average F_1 of 0.84 found here was significantly higher than the 0.73 previously reported.

Table 1. Testset accuracy for Concept and Causal Relation Identification

	Coral Bleaching			Skin Cancer		
	Recall	Precision	F_1	Recall	Precision	F_1
Window-based Tagger	0.802	**0.885**	**0.842**	0.779	0.853	0.814
CRF	0.797	0.787	0.835	0.759	0.855	0.804
HMM	0.799	0.702	0.747	0.731	0.628	0.675
Structured Perceptron	0.794	0.884	0.837	0.773	0.860	0.814
Bi-directional RNN	**0.830**	0.855	**0.842**	**0.807**	**0.869**	**0.837**
RNN Word Tagger	0.656	0.698	0.676	**0.798**	0.786	**0.792**
Stacked Model	0.674	**0.736**	0.704	0.719	0.816	0.765
Dependency Parser	**0.766**	0.693	**0.728**	0.760	**0.823**	0.790

4 Causal Relation Identification

Causal relation identification is a much more challenging task than concept identification because a concept tends to be described by a relatively small set of contiguous words, whereas causal relations are inherently spread across a wider range of words and variety of patterns. Previous work on detecting causal relations in text reflects the difficulty of the problem, either restricting the forms of relations that were considered [2, 9] or achieving rather low performance (e.g., $F_1 = 0.41$ [19], $F_1 = 0.39$ [20]). Our previous work with an SVM classifier

achieved $F_1 = 0.63$ for the two topics [10], but it was limited to detecting only the presence or absence of *any* causal relation within a sentence. Here, we evaluated three techniques:

- **RNN Word Tagger**: We trained a bi-directional RNN to predict, for each word, the label of the the causal connection that it was involved in (if any). The same RNN architecture described above performed best.
- As a **Stacked Model** [15, for example], we used predictions for all codes in a sentence, and their combinations from the best concept identifier, the RNN, as inputs to a logistic regression classifier, because it is robust to overfitting and can learn from arbitrary input features.
- **Transition-based Dependency Parser**: We developed a novel parsing mechanism which learn to detect causal relations between concept codes predicted by the RNN model. The parsing mechanism was adapted from dependency parsers, such as [4].

The performance of the different causal relation identification techniques on the test sets for both topics is shown in the bottom part of Table 1. The dependency parser produced the top combined performance with an average F_1 score of 0.759, compared to 0.734 for the RNN Word Tagger and 0.735 for the stacked model. The parser's advantages are reflected in the pattern of results. In the coral bleaching topic, students mentioned 85 different relations, compared to 49 relations between the smaller set of concepts in the skin cancer topic. Accordingly, the average number of examples of each causal relation was much higher in the skin cancer topic (20.3 compared to 7.0). The parser learns when it can combine two concept codes into a causal relation instead of treating each relation as a separate label. This allows it to generalize better over all of the relations, as reflected in the higher recall scores for the parser over the other models on the coral bleaching topic. The higher precision for the parser on the skin cancer topic than on coral bleaching can be attributed to the higher number of training examples.

5 Conclusions

In this paper, we compared the performance of several highly competitive techniques for identifying explanation structure, including a novel adaptation of a parsing mechanism to the task of causal relation identification. The bi-directional RNN showed the best performance on the concept identification task, achieving an average F_1 score of 0.84, significantly higher than that found in previous research. Although the Word-Tagging RNN achieved slightly higher performance than the Dependency Parser for causal relation identification on the skin cancer topic, overall the parser provided better performance, with an average F_1 of 0.76. Here too, we have achieved a significant increase in accuracy over previous research. This level of performance indicates that these techniques can be confidently used by an intelligent system to give feedback on the concepts and causal structure in students' scientific explanations.

References

1. Achieve, Inc: Next Generation Science Standards (2013)
2. Blanco, E., Castell, N., Moldovan, D.: Causal relation extraction. In: LREC (2008)
3. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational linguistics **18**(4), 467–479 (1992)
4. Collins, M.: Head-driven statistical methods for natural language parsing. Unpublished PhD thesis, University of Pennsylvannia (1999)
5. Collins, M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 1–8. Association for Computational Linguistics (2002)
6. Council of Chief State School Officers (CCSSO): The Common Core Standards for English Language Arts and Literacy in History/Social Studies and Science and Technical Subjects (2010), downloaded from http://www.corestandards.org
7. Dietterich, T.G.: Machine learning for sequential data: A review. In: Structural, syntactic, and statistical pattern recognition, pp. 15–30. Springer (2002)
8. Foltz, P.: Latent Semantic Analysis for text-based research. Behavior Research Methods, Instruments, and Computers **28**, 197–202 (1996)
9. 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. 13–18 (2007), `http://acl.ldc.upenn.edu/W/W07/W07-2003.pdf`
10. Hughes, S., Hastings, P., Britt, M.A., Wallace, P., Blaum, D.: Machine learning for holistic evaluation of scientific essays. In: Proceedings of Artificial Intelligence in Education 2015. Springer, Berlin (2015)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). pp. 282–289 (2001)
13. Magliano, J.P., Graesser, A.C.: Computer-based assessment of student-constructed responses. Behavior Research Methods **44**(3), 608–621 (2012)
14. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press (2014)
15. Menahem, E., Rokach, L., Elovici, Y.: Troika–an improved stacking schema for classification tasks. Information Sciences **179**(24), 4097–4122 (2009)
16. Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs) (2007), `http://www.chokkan.org/software/crfsuite/`
17. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
18. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
19. Riaz, M., Girju, R.: Recognizing causality in verb-noun pairs via noun and verb semantics. EACL 2014 p. 48 (2014)
20. 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)

21. Sánchez-Villamil, E., Forcada, M.L., Carrasco, R.C.: Unsupervised training of a finite-state sliding-window part-of-speech tagger. In: Advances in Natural Language Processing, pp. 454–463. Springer (2004)
22. Shermis, M.D., Burstein, J.: Handbook of automated essay evaluation: Current applications and new directions. Routledge (2013)
23. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations Session at EACL 2012. Association for Computational Linguistics, Avignon, France (April 2012), http://brat.nlplab.org
24. Stenetorp, P., Topić, G., Pyysalo, S., Ohta, T., Kim, J.D., Tsujii, J.: Bionlp shared task 2011: Supporting resources. In: Proceedings of BioNLP Shared Task 2011 Workshop. pp. 112–120. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/W11-1816
25. Welch, L.R.: Hidden markov models and the Baum-Welch algorithm. IEEE Information Theory Society Newsletter **53**(4), 10–13 (2003)