Matching Hypothesis Text in Diagrams and Essays

Collin F. Lynch
Center for Educational Informatics
North Carolina State University
Raleigh, North Carolina, U.S.A.
collinl@cs.pitt.edu

Mohammad Falakmasir
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, Pennsylvania, U.S.A.
falakmasir@cs.pitt.edu

Kevin D. Ashley
Learning Research & Development Center
University of Pittsburgh
Pittsburgh, Pennsylvania, U.S.A.
ashley@pitt.edu

Keywords
Argument Diagramming, Ill-Defined Domains, Intelligent Tutoring Systems, Text-Mining, Multiple-representations

1. INTRODUCTION
We have previously shown that argument diagrams can help students both to read existing arguments [8] and to plan new ones [4]. We have also shown that student-produced diagrams can be graded reliably and evaluated, both by human graders and automatic analysis, to predict subsequent essay grades [4, 6, 5]. Argument diagrams are advantageous for tutoring as they focus students' attention on key structural features of otherwise implicit or opaque arguments as well as supporting empirically-valid automatic assessment and feedback [4]. It has not yet been shown that the content of the argument diagrams closely matches the essay text or that the two can be automatically aligned. Here we show that automatic alignment of hypothesis statements and hypothesis nodes is possible.

A sample hypothesis node is shown in Fig. 1. The ontology used here included nodes representing hypotheses, citations, claims, and the current study. Students added these nodes to a flexible workspace and connected them using supporting, opposing, undefined, and comparison arcs. Hypothesis nodes frame the discussion via a simple *if-then* format. The hypothesis statement tagged in the associated essay is:

> When presented with text-based signs versus signs with text and images or symbols, individuals will be more likely to respond to those signs with both images and text.

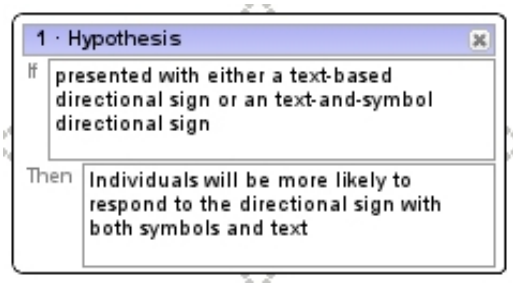

Figure 1: A sample hypothesis node drawn from a student-produced LASAD diagram.

2. ANALYSIS
Data for this study was drawn from work on argument planning for writing described in [4]. In that study we collected a set of 105 paired diagrams and essays collected in a course on Research Methods at the University of Pittsburgh. Students in this course conducted a group research project. As part of the assignment students produced an argument diagram when planning their essay. The diagrams and essays were graded independently by an expert grader who also annotated the hypothesis statements within the text. The reliability of the grading and annotation was evaluated via a separate inter-grader reliability study where we found 70% agreement on hypothesis tags. 85 of the pairs contained one or more hypothesis nodes in the diagram and one or more tagged hypothesis statements in the essay.

We assessed our primary hypothesis via two types of analyses. In the first analysis we focused on sentence *classification* with the goal of determining whether the textual information from the hypothesis nodes can be used to train efficient classifiers or to improve upon existing techniques. We split the papers into individual sentences using the grader tagging to annotate hypothesis statements. We then extracted two feature vectors for each sentence. We extracted a *static* feature vector for each sentence that reflects the 6 most frequent keyword stems: 'would,' 'likely,' 'hypothe,' 'study,' 'expect,' and 'predict.' Ironically the most predictive single feature was 'predict.'

We then matched each of the candidate sentences with the text drawn from the hypothesis node in the associated diagram and calculated a *similarity* vector. This vector con-

tained five features each of which reflected the output of an existing sentence similarity metric. The features were: Levenshtein distance [3, 11], Jaro-Winkler distance [12, 10], Ratcliff & Obershelp score [9], and two semantic metrics based upon WordNet [1]: Path [2], and Wu & Palmer [13]. For these latter measures, the similarity scores were calculated using only the first sense of the words in each sentence. For diagrams with more than one hypothesis node we computed one similarity vector per node and chose the best result based upon the Ratcliff & Obershelp score.

We trained two sets of classifiers via 10-fold cross-validation using these features. One set of classifiers was trained solely on the *static* vectors and reflected the predictiveness of the individual cue terms while the second *combined* both the static and similarity vectors for each sentence. We chose five standard classification algorithms for this purpose: Naïve Bayes, Nearest Neighbor, Maximum Entropy, Support Vector Machines, and Linear Regression. All of the classifiers were trained and evaluated using the RapidMiner toolkit [7]. For the Linear Regression model we tagged the sentences with a binary output variable and made predictions based upon a fixed cutoff of $\frac{1}{2}$. RapidMiner performs some mechanical filtering of collinear terms.

The most precise classifier was a maximum-entropy model based upon the static features which had a precision of 0.7, a recall of 0.48, and an F1 score of 0.56. The best overall classifier was an SVN model based upon the combined features which also achieved the highest recall and F1 Scores. The precision, recall, and F1 scores for this model were 0.65, 0.65, and 0.65 respectively.

In our second analysis we implemented a second linear ranking function that estimates the likelihood of each sentence being a hypothesis statement based upon the aforementioned features. The weights in this model were trained using leave-one-out cross-validation. For each essay we then selected the sentence or sentences with the highest likelihood of being a hypothesis statement. For this analysis we compared predictions based on the static features alone, similarity alone, and the combined set. These algorithms were trained via leave-one-out cross-validation and were designed to select the best sentence on a per-paper basis. As in the classification study we found that the combined model outperformed the static and similarity models with precision scores of 73%, 66%, and 55% respectively.

3. CONCLUSIONS

We found that combined models which used both the static static features and the similarity measures were better at classifying hypothesis statements and ranking candidate statements within a written essay than either the static or similarity features alone. These results lead us to conclude that it is possible to use this similarity information to link the hypothesis nodes and hypothesis statements most of the time. In future work we plan to test automatic alignment of other diagram components and to investigate other linking mechanisms. We believe that a hybrid model which incorporates information from multiple nodes can be more robust than any individual comparison.

Acknowledgments

This work was supported by NSF award 1122504, "DIP: Teaching Writing and Argumentation with AI-Supported Diagramming and Peer Review," Kevin D. Ashley PI with Chris Schunn and Diane Litman, co-PIs.

4. REFERENCES

[1] C. Fellbaum. *WordNet: An electronic lexical database*. MIT Press, 1998.

[2] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In Fellbaum [1], pages 265–283.

[3] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, 10, 1966.

[4] Collin F. Lynch. The Diagnosticity of Argument Diagrams, 2014. (defended January 30th 2014).

[5] Collin F. Lynch and Kevin D. Ashley. Empirically valid rules for ill-defined domains. In John Stamper and Zachary Pardos, editors, *Proceedings of The 7^{th} International Conference on Educational Data Mining (EDM 2014)*. International Educational Datamining Society IEDMS, 2014. (In Press).

[6] Collin F. Lynch, Kevin D. Ashley, and Min Chi. Can diagrams predict essays? In Stefan Trausen-Matu and Kristy Boyer, editors, *Intelligent Tutoring Systems, 12th International Conference, ITS 2014, Honolulu, Hawai'i, USA*, Lecture Notes in Computer Science. Springer, 2014. (In Press).

[7] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1–6, 2006.

[8] Niels Pinkwart, Kevin D. Ashley, Collin F. Lynch, and Vincent Aleven. Evaluating an intelligent tutoring system for making legal arguments with hypotheticals. *International Journal of Artificial Intelligence in Education*, 19(4):401–424, 2009.

[9] J. W. Ratcliff and David Metzener. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 7(46), 1988.

[10] Wikipedia. Jaro-winkler distance — wikipedia, the free encyclopedia, 2014. [Online; accessed 28-March-2014].

[11] Wikipedia. Levenshtein distance — wikipedia, the free encyclopedia, 2014. [Online; accessed 1-March-2014].

[12] W. E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, pages 354–359, 1990.

[13] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico*, 1994.