Comparing the Accuracy of Automatic Scoring Solutions for a Text Comprehension Diagramming Intervention

Héctor J. Pijeira-Díaz
Maastricht University
h.pijeiradiaz@maastrichtuniversity.nl

Jasper Snel
Brightlands Institute for Smart Society Heerlen
j.snel@maastrichtuniversity.nl

Sophia Braumann
Utrecht University
s.e.e.braumann@uu.nl

Janneke van de Pol
Utrecht University
J.E.vandePol@uu.nl

Tamara van Gog
Utrecht University
T.vanGog@uu.nl

Anique de Bruin
Maastricht University
anique.debruin@maastrichtuniversity.nl

ABSTRACT: Students typically have great difficulty monitoring their comprehension of textual materials. Completing diagrams about causal relations in expository texts has been a successful intervention to enhance the accuracy of students' reading comprehension judgments (i.e., monitoring accuracy), although there is still room for improvement. Such judgments play a role in crucial self-regulated learning decisions that students make such as allocating time and effort, selecting content for restudy, and/or consulting additional sources. The automated scoring of students' diagram content can provide a basis for strengthening the diagramming intervention with individual and simultaneous feedback to a high number of students. Leveraging an existing human-coded (correct and incorrect) dataset of 6000+ diagram answers (completed in Dutch by 700+ secondary students), we compared different automatic scoring solutions in terms of classification accuracy. Four computational linguistic models for Dutch were identified and tested in combination with four popular machine learning classification algorithms. The best solution reached 81% accuracy (i.e., four out of five answers matched the human coding). Depending on the accuracy required for different applications, these results could be used for fully- or semi-automated scorings of students' answers to generative activities used in reading comprehension interventions.

Keywords: computational linguistics, automatic scoring, reading comprehension, diagramming, monitoring judgments

1 INTRODUCTION

As students progress through the educational system, they switch from an early goal of learning to read, to increasingly reading to learn. In the process of reading to acquire knowledge, students monitor their comprehension, which is an important determinant of their study decisions (Thiede et al., 2009). For example, they spend more time restudying those texts that they think they understand less well. Study decisions, in turn, are likely to influence their exam scores.

The accuracy of students' monitoring of their reading comprehension is determined by relating how students think they will score on a test about a certain text, to their actual test scores. Unfortunately, meta-analytic results show an average accuracy below 0.30 (Goodman and Kruskal's gamma) (Prinz et al., 2020). Consequently, educational researchers have developed interventions (e.g., generating keywords, summaries, or diagrams) to enhance students' monitoring accuracy. In particular, completing pre-structured diagrams about causal relations in expository texts have proved effective in raising monitoring accuracy over 0.55 (van de Pol et al., 2019). Nonetheless, the numbers indicate that, despite the progress, there is room for improvement.

Concurrently, advances in computational linguistics offer promising opportunities for the automatic scoring of the diagrams produced by the students in those interventions. The interest in automation is motivated by the real-time, individual feedback possibilities it affords towards increasing monitoring accuracy. Leveraging an existing dataset of 6000+ diagram answers in Dutch, including human scores of whether answers are correct or not, we evaluated the accuracy of automatic scoring as compared to human scoring. The dataset answers were produced by 700+ Dutch-speaking secondary education students in a series of diagram completion interventions to enhance monitoring accuracy (van de Pol et al., 2019; van Loon et al., 2014). The automatic scoring capitalizes on computational linguistics models for text representation (Mikolov et al., 2013) in combination with machine learning classification algorithms. Our aims are therefore to identify available computational linguistics models for representing Dutch text and to compare their performance in terms of classification accuracy. Four such models and four classifiers are tested in this study.

2 METHODOLOGY

First, a literature review was conducted to identify the latest developments in computational linguistic models for text representation in Dutch. Then, four popular classification algorithms in machine learning, namely 1) logistic regression, 2) support vector machines (SVM), 3) random forests, and 4) neural networks; were used in combination with the linguistic models.

The content of each diagram answer in the existing dataset was human coded as correct or incorrect with satisfactory interrater reliability. For the automatic scoring, since mathematical computer algorithms operate with numbers instead of texts, the first step was to represent the text of each diagram answer as a multidimensional numerical vector (300 dimensions) using the identified computational linguistic models. Then, the classification algorithms were trained with 90% of the data, leaving the remaining 10% for evaluation of the classification performance.

3 RESULTS

Four available state-of-the-art computational linguistics models were identified, namely 1) “spaCy medium”, 2) “spaCy large”, 3) “FastText”, and 4) “ConceptNet Numberbatch” (ConNum). Table 1 shows the classification accuracies of the automatic scoring of the texts obtained for all the combinations of the four Dutch models and the four classification algorithms. The size of the model is also included as an indication of their complexity. The best accuracy (81%) was obtained for the “spaCy medium” model in combination with a neural network classifier. Thus, the best performing combination for automatic scoring gave, in slightly over 4 out of 5 cases, the same score as a human did. Remarkably, it also means that the simplest model in terms of size (i.e., “spaCy medium”) offered the best automatic scoring accuracy.

Table 1. Automatic scoring accuracy by classification algorithm and Dutch language model.

Dutch Model	Size	Classification Algorithm			
		Logistic Regression	SVM	Random Forests	Neural Networks
FastText	7GB	79%	77%	70%	81%
spaCy medium	44MB	80%	79%	70%	80%
spaCy large	545MB	77%	79%	70%	70%
ConNum	223MB	76%	80%	71%	70%

4 EDUCATIONAL SIGNIFICANCE

The educational significance of this work is threefold. First, our automated scoring solution enables the development and testing of automated individualized feedback interventions to further improve students’ monitoring accuracy. Second, it can alleviate teachers’ workload in scoring students’ text comprehension diagrams. Third, it can be integrated into educational technology applications such as intelligent tutoring systems, especially those focusing on assisting reading comprehension.

REFERENCES

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations - Workshop Track Proceedings*.

Prinz, A., Golke, S., & Wittwer, J. (2020). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-analytic insights. *Educational Psychology Review*, *32*(4), 917–949. https://doi.org/10.1007/s10648-020-09558-6

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). Routledge. https://doi.org/10.4324/9780203876428

van de Pol, J., de Bruin, A. B. H., van Loon, M. H., & van Gog, T. (2019). Students’ and teachers’ monitoring and regulation of students’ text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology*, *56*, 236–249. https://doi.org/10.1016/j.cedpsych.2019.02.001

van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica*, *151*(September), 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007