Automated Scoring of Self-explanations Using Recurrent Neural Networks

Marilena Panaite[1], Stefan Ruseti[1], Mihai Dascalu[1,2], Renu Balyan[3], Danielle S. McNamara[3], and Stefan Trausan-Matu[1,2]

[1]University “Politehnica” of Bucharest

[2]Academy of Romanian Scientists

[3]Arizona State University

Automated Scoring of Self-explanations Using Recurrent Neural Networks

Abstract

Intelligence Tutoring Systems (ITSs) focus on promoting knowledge acquisition, while providing relevant feedback during students' practice. Self-explanation practice is an effective method used to help students understand complex texts by leveraging comprehension. Our aim is to introduce a deep learning neural model for automatically scoring student self-explanations that are targeted at specific sentences. The first stage of the processing pipeline performs an initial text cleaning and applies a set of predefined rules established by human experts in order to identify specific cases (e.g., students who do not understand the text, or students who simply copy and paste their self-explanations from the given input text). The second step uses a Recurrent Neural Network with pre-trained Glove word embeddings to predict self-explanation scores on a scale of 1 to 3. In contrast to previous SVM models trained on the same dataset of 4109 self-explanations, we obtain a significant increase of accuracy from 59% to 73%. Moreover, the new pipeline can be integrated in learning scenarios requiring near real-time responses from the ITS, thus addressing a major limitation in terms of processing speed exhibited by the previous approach.

Keywords: Natural Language Processing, Comprehensive tutoring system, Self-explanations, Recurrent Neural Network

Automated Scoring of Self-explanations Using Recurrent Neural Networks

Marilena Panaite[1], Stefan Ruseti[1], Mihai Dascalu[1,2](✉), Renu Balyan[3], Danielle S. McNamara[3], and Stefan Trausan-Matu[1,2]

[1] Faculty of Automatic Control and Computers,
University "Politehnica" of Bucharest, 313 Splaiul Independenței,
60042 Bucharest, Romania
marilena.panaite@gmail.com, {stefan.ruseti,
mihai.dascalu, stefan.trausan}@cs.pub.ro

[2] Academy of Romanian Scientists, Splaiul Independenţei 54,
050094 Bucharest, Romania

[3] Institute for the Science of Teaching and Learning, Arizona State University,
PO Box 872111, Tempe, AZ 85287, USA
{renu.balyan, dsmcnama}@asu.edu

Abstract. Intelligent Tutoring Systems (ITSs) focus on promoting knowledge acquisition, while providing relevant feedback during students' practice. Self-explanation practice is an effective method used to help students understand complex texts by leveraging comprehension. Our aim is to introduce a deep learning neural model for automatically scoring student self-explanations that are targeted at specific sentences. The first stage of the processing pipeline performs an initial text cleaning and applies a set of predefined rules established by human experts in order to identify specific cases (e.g., students who do not understand the text, or students who simply copy and paste their self-explanations from the given input text). The second step uses a Recurrent Neural Network with pre-trained Glove word embeddings to predict self-explanation scores on a scale of 1 to 3. In contrast to previous SVM models trained on the same dataset of 4109 self-explanations, we obtain a significant increase of accuracy from 59% to 73%. Moreover, the new pipeline can be integrated in learning scenarios requiring near real-time responses from the ITS, thus addressing a major limitation in terms of processing speed exhibited by the previous approach.

Keywords: Natural Language Processing · Comprehensive tutoring system · Self-explanations · Recurrent Neural Network

1 Introduction

Learning involves integration of new information into prior knowledge [1]. In this case, reading a text is not a guaranty that students have acknowledged new presented terms and that they have made connections with prior learned terms. Self-explanation facilitates this process and improves comprehension by encouraging students to engage in both metacognition and inference generation [1]. However, providing individual

M. Scheffel et al. (Eds.): EC-TEL 2019, LNCS 11722, pp. 659–663, 2019.
https://doi.org/10.1007/978-3-030-29736-7_61

feedback to each student self-explanation is cumbersome and cannot be easily scaled by tutors without the help of Intelligent Tutoring Systems. In this regard, automated systems that provide help for scoring students' self-explanations can speed up the process.

The aim of this study is to train a Recurrent Neural Network (RNN) [2] to improve the automated score prediction of the student's self-explanations. The trained RNN model is then integrated into the new workflow in the state-of-the-art tutoring system - Interactive Strategy Training for Active Reading and Thinking (iSTART) [3] which is a web-based ITS created to improve adolescent students' comprehension of complex scientific texts. In order to help them, iSTART utilizes non-game and game-based generative practice, in which learners produce their own self-explanations, and game-based identification practice, in which learners attempt to identify which strategy is being used in certain self-explanations.

With the help of the iSTART practice, the full pipeline of the Intelligent Tutoring System uses advanced Natural Language Processing (NLP) techniques for extracting the main features of the automatically scored explanations. In this regard, the ReaderBench framework [4] offers a variety of NLP techniques, all grounded in Cohesion Network Analysis. Our updated pipeline computes textual complexity indices for the input target texts and corresponding explanations that are further used in the rule system, whereas an RNN model is used to automatically assess the quality of a student's explanation in term of metacognition and capacity to infer new knowledge.

2 Integrated Workflow with RNN Model

The corpus used to train the model contains 4,109 self-explanations from 277 high-school students on two science texts, namely "Heart Diseases" ($\sim$300 words) and "Red Blood Cell" ($\sim$280 words). Each text contains nine target sentences. To assess the performance of students, two experts evaluated the student's self-explanations, assigning scores from 0 (poor) to 3 (great). The human experts performed two rounds of scoring 60% of the entire dataset and achieved a high interrater reliability (Kappa = .81).

Our approach provides an integrated workflow that can compute an automated score for each self-explanation with relevant feedback. The first step performs cleaning the input data using a spell-check algorithm and the NLP processing pipeline from ReaderBench [4]. The second step of the pipeline consists of predefined rules that identify poorly written self-explanations. For example, a rule checks whether a self-explanation contains more than 75% frozen expressions by comparing the text with a predefined set of regular expressions. Moreover, copy and paste from the target sentence is also checked. We also identify the new concepts introduced by the student in the response, by checking words that are neither synonyms nor identical lemmas of the words present in the target sentence. If there are no new concepts introduced, students receive a score of 1 (fair) with corresponding feedback (e.g., "Can you add more information to explain what the text means?"). After making specific inferences using the rule-based system, the remaining self-explanations are assessed using an automated scoring system.

For the current approach we rely on an RNN model [2] wherein (a) Gated Recurrent Units (GRUs) [5] are used to represent the target sentences, and (b) Bidirectional Gated Recurrent Units (BiGRU) represent self-explanation obtained from student responses. After the encoding phase, each output matrix is reduced to a fixed size by retaining only the most meaningful features using an attention mechanism. The network uses max-pooling for obtaining the maximum of each sentence encoding matrix. Further, the outputs are concatenated, and a dropout regularization mechanism is used in order to avoid overfitting. A hidden layer of size 50 with sigmoid activation function is then added. The last step uses a softmax layer and computes the probabilities to classify the self-explanations into one of the three classes that reflect the quality of the self-explanation (Fig. 1).

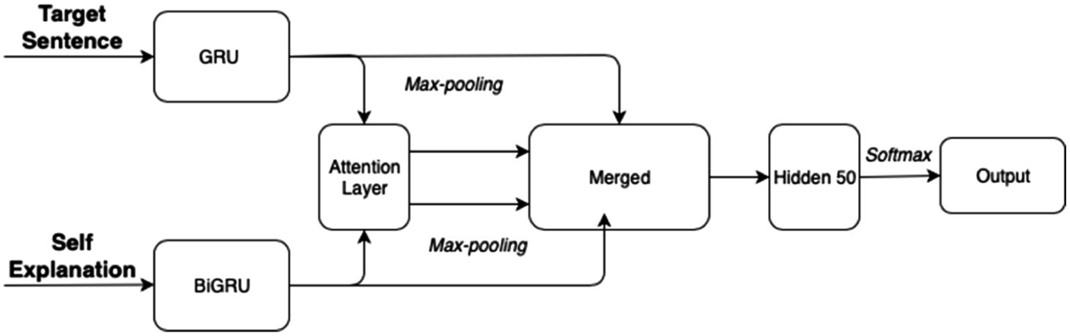

Fig. 1. Architecture of RNN model used for automated scoring.

In the final version of the model, sentences (target and self-explanations) were encoded using Glove-300d word embeddings [6] which provided the best overall results, but we experimented with other embeddings (e.g., Glove-100d, FastText) and obtained an accuracy that was 1–2% lower.

3 Results and Discussion

The current paper introduces major improvements to the automated scoring system for self-explanations by considering deep learning models in place of SVMs. The rule system for predicting poor (0) and fair (1) scores was retained from the previous version of the automated pipeline, while the SVM model trained with the textual complexity indices from ReaderBench was replaced with the RNN trained using Glove-300 for encoding the target sentences and the self-explanations of the students. In the initial SVM experiments, the best accuracy obtained was 59%. Using the same pre-checked rule-based system, the pre-trained RNN model was used to score the self-explanations; the best accuracy obtained in this case was 73.6%.

Another indication of the accuracy of a model is adjacent accuracy, which is assessed by calculating the proportion of automated scores that differ by no more than 1 from the expert scores. The best adjacent accuracy was 97% for the SVM [7] indicating that, although the accuracy was 59%, the automated scoring model was close to the expert scores. The same metric was computed for the updated pipeline that uses the RNN trained model; the model achieved an adjacent accuracy of 93.87%, which was slightly lower than the previous SVM model.

Moreover, the new approach eliminated the need for computing linguistic features and introduced the RNN model which only needs to encode the input data using Glove-300-word embeddings. Our results indicate that the use of deep learning models with specific NLP techniques can improve the performance of the overall system and perform better in terms of time and accuracy than classic machine learning models [8].

The current study is principally limited by the dataset, which includes only two target texts. Hence, one consideration concerns generalization of the model to other texts and populations. Our limitation stems from the resources necessary to have self-explanations (reliably) scored by experts. We will continue to explore more affordable options such as crowdsourcing for both explanation and their respective scores.

Another cautionary note stems from low accuracy achieved for self-explanations that received a score of 3 by experts. The model performance is higher than that reported by the previous SVM model, but the new model still struggles to make inferences similar to those generated by humans when judging explanations that go well beyond the text. One way to improve the score would be to include specific features for detecting scores noted as 3 by including the scores and the relevant embeddings from previous self-explanations of a student for the same text.

Acknowledgments. This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-III 72PCCDI/2018, ROBIN – "Roboții și Societatea: Sisteme Cognitive pentru Roboți Personali și Vehicule Autonome", the Department of Education, Institute of Education Sciences - Grant R305A130124 and R305A190063, and the Department of Defense, Office of Naval Research - Grants N00014140343 and N000141712300.

References

1. Chi, M.T., De Leeuw, N., Chiu, M.-H., LaVancher, C.: Eliciting self-explanations improves understanding. Cogn. Sci. **18**(3), 439–477 (1994)
2. Sundermeyer, M., Ney, H., Schlüter, R.: From feedforward to recurrent LSTM neural networks for language modeling. IEEE Trans. Audio Speech Lang. Process. **23**(3), 517–529 (2015)
3. McNamara, D.S., O'Reilly, T.P., Rowe, M., Boonthum, C., Levinstein, I.B.: iSTART: a web-based tutor that teaches self-explanation and metacognitive reading strategies. In: McNamara, D.S. (ed.) Reading comprehension strategies: Theories, interventions, and technologies, pp. 397–420. Erlbaum, Mahwah, NJ (2007)
4. Dascalu, M., McNamara, D.S., Trausan-Matu, S., Allen, L.K.: Cohesion network analysis of CSCL participation. Behav. Res. Methods, 1–16 (2017)
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
6. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

7. Panaite, M., et al.: Bring it on! challenges encountered while building a comprehensive tutoring system using ReaderBench. In: International Conference on AI in Ed., pp. 409–419. Springer (2018)
8. Balyan, R., McCarthy, K.S., McNamara, D.S.: Comparing machine learning classification approaches for predicting expository text difficulty. In: The Thirty-First International Flairs Conference (FLAIRS 31), pp. 421–426. AAAI, Melbourne, FL (2018)