Extended Multi-document Cohesion Network Analysis Centered on Comprehension Prediction

Bogdan Nicula[1], Cecile A. Perret[2], Mihai Dascalu[1,3(✉)], and Danielle S. McNamara[2]

[1] University Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania
bogdan.nicula@cti.pub.ro, mihai.dascalu@cs.pub.ro
[2] Department of Psychology, Arizona State University, PO Box 871104, Tempe, AZ 85287, USA
{cperret,dsmcnama}@asu.edu
[3] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania

Abstract. Theories of discourse argue that comprehension depends on the coherence of the learner's mental representation. Our aim is to create a reliable automated representation to estimate readers' level of comprehension based on different productions, namely self-explanations and answers to open-ended questions. Previous work relied on Cohesion Network Analysis to model a cohesion graph composed of semantic links between multiple reference texts and student productions. From this graph, a set of features was derived and used to build machine learning models to predict student comprehension scores. In this paper, we build on top of the previous study by: a) extending the CNA graph by adding new semantic links targeting specific sentences that should have been captured within the learner's productions, and b) cleaning the self-explanations by eliminating frozen expression, as well as entries which seemed nearly identical to the source text. The results are in line with the conclusions of the previous study regarding the importance of both self-explanations and question answers in predicting the students' reading comprehension level. They also outline the limitations of our feature generation approach, in which no substantial improvements were detected, despite adding more fine-grained features.

Keywords: Multi-document comprehension modeling · Cohesion Network Analysis · Natural Language Processing

1 Introduction

Reading comprehension is a complex task composed of numerous steps, phases, and parallel processes. It involves extracting ideas from a text at multiple levels, including individual sentences, paragraphs as macro-constituents, and even entire documents when multiple texts are considered. Concurrently, a coherent mental representation of

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 228–233, 2020.
https://doi.org/10.1007/978-3-030-52240-7_42

the text is established through connections between various text-based information, as well as with prior knowledge. One key aspect of a reader's mental representation is its coherence, or interconnectedness [1]. Our objective in this project is to develop automated measures of the coherence of readers' mental representation both during and after reading to provide dynamic indicators of readers' level of comprehension.

In our work, we analyze semantic distances (considered a good estimator for coherence) between a set of documents and productions generated by learners under two conditions: a) self-explanations (SEs), generated at specific target sentences while reading the reference documents, and b) open-ended comprehension questions (QAs) that relate to one or more documents. Our aim is to predict multi-document comprehension based on semantic features denoting the links between the reference documents and the student productions. Similar approaches were previously attempted for single text comprehension [2, 3], as well as multiple document scenarios [4].

Cohesion Network Analysis (CNA) [3] was applied in a study by Nicula, Perret, Dascalu and McNamara [5] in a multiple document setting to model the coherence of learner productions, and predict their comprehension level. CNA relies on Natural Language Processing [6] techniques to model discourse in terms of semantic links. CNA is inspired by and transcends Social Network Analysis [7] by considering semantic relatedness between text segments. Its core purpose is to represent cohesion as a graph composed of multiple types of links reflecting semantic distances between elements of different granularity levels (i.e., n-gram sequence, sentence, paragraph, or texts). Several semantic models (such as: LSA [8], Wu-Palmer semantic distance in WordNet [9], word2vec [10] or GloVe [11]) can be used to compute these distances, all of them being available within the ReaderBench framework [12]. For the current study, the CNA graph modeled how information from the reference texts was extracted and structured by readers, while analyzing the links between their productions and the source texts.

Three enhancements were considered while relating to the initial study performed by Nicula, Perret, Dascalu and McNamara [5]. First, we examined the effects of adding features targeting the relation between SEs and specific reference sentences from the target text sequence. This was done in order to better assess whether students' SEs related to relevant information from the prior text. Second, we performed a thorough SE cleaning to check for copy and paste, as well as specific frozen expressions, to provide feedback. Third, a more rigorous and in-depth analysis was performed by calculating the regressions for multiple iterations in an attempt to obtain more informative results less prone to possible outliers.

2 Method

2.1 Corpus

The same corpus in [5] was used, consisting of self-explanations and answers to open-ended questions from 146 students on 4 texts, discussing the same topic. Readers are prompted to write an SE to a sentence at several intervals throughout each text to help them generate inferences within a text. In contrast, the QAs have a target text, but,

depending on the question type, they may require linking information from the other texts as well. The students' answers to the 12 questions (3 per text) were graded, resulting in a comprehension score with values ranging from 0 to 12. The students also produced 30 self-explanations on specific target sentences distributed throughout the texts, but these self-explanations were not individually scored.

2.2 Feature Extraction and Selection

A set of features was generated based on the students' responses (i.e., SEs and QAs) reflecting the overlap between the information covered by each response and the information available in the target text. The SE features contain information regarding the semantic similarity between each SE and the four reference texts, the sequences of text targeted by the SEs, and the paragraphs targeted by the SEs. In the case of links between SEs and paragraphs, the extracted features represent aggregate statistics such as the mean, maximum, or standard deviation of the semantic similarity scores corresponding to the links from one SE to all the paragraphs in the targeted text. The information extracted per SE is then aggregated per student by computing the mean, maximum, or standard deviation of these values for all the SEs generated by that student. This results in 272 SE-related features per student.

Compared to previous work, efforts were made to clean up the SEs by eliminating information that is not relevant to our task and by removing SEs that copy-pasted information from the original texts. An approach based on pattern matching with regular expressions was employed to eliminate redundant, uninformative content. In terms of eliminating self-explanations that seemed to be copied, an approach using both n-grams and bag-of-words was applied, eliminating entries that had a high overlap with the source texts. The QA features in the original paper contained information regarding the semantic similarity between the QAs and the 4 texts, and the paragraphs targeted by the QAs. As part of this work, extra information has been added to the model described by [5] in the form of specifying the exact sentences and self-explanations to which a question refers. The semantic distance between the questions and the specified sentences/self-explanations was computed using the same approach. This increased the number of QA-related features from 90 to 330. The extended set of features was passed through the same 2-stage filtering pipeline, which eliminates features with high intra-correlation and features with low correlation to the reading comprehension score. A grid search approach was used to find the most predictive combination of thresholds for the 2 filtering stages. A set of reasonable values were selected for each of the 2 thresholds, and all combinations were tested to determine the best combinations.

3 Results

The 5-fold cross-validation experiments were run 10 times with different random seeds to have more robust results, while the mean and best results were recorded. In this setup, results were slightly different from the ones reported in the original paper, but the conclusions mentioned there still hold using only the original features. When adding the two enhancements (i.e., cleaning of SEs and the extra information regarding links

between QAs, SEs, and specific targeted sentences), the best results were slightly below those obtained in the original work; however, the results for all the models except the linear regression improved, implying that threshold selection should be improved. After the extended set of 602 features was generated on the cleaned SEs, the two thresholds for the 2-stage feature filtering were sought using grid search. Depending on the threshold parameters, the filtered set of features varied between 12 and 55 features, but the best performance in all of these experiments was still 2% worse than the results obtained with the original set of features, on the original task (Table 1).

Table 1. Results obtained with features from the 602-feature extended set.

Experimental setup	# SE features	# QA features	Best average performance (MAE)	Best performing model
Original set + intra-corr. < .90 + comp. r > .40	7	13	**1.305**	Linear regression
Extended set + intra-corr. < .90 + comp. r > .40	10	46	1.329	Support Vector Regression
Extended set + intra-corr. < .90 + comp. r > .50	1	19	1.424	Extra trees
Extended set + intra-corr. < .85 + comp. r > .40	8	33	1.319	Support Vector Regression
Extended set + intra-corr. < .95 + comp. r > .40	12	72	1.338	Support Vector Regression
Extended set + intra-corr. < .95 + comp. r > .50	1	32	1.389	Linear regression

* *intra-corr* =intra-correlation above threshold; *comp. r* = reading comprehension score

4 Conclusions

This study confirms some of the conclusions from the original paper [5], namely that the usage of both QA and SE features yields better predictions, while the step of filtering features by intra-correlation helps improve performance. Nevertheless, it seems that that the additional information (i.e., specifically targeting the sentences that should have been referred to by both SEs and questions) is not extremely helpful in the final prediction. A possible explanation resides in the manner in which we extract the semantic data at sentence-level (i.e., average word2vec representations of all words [13]) – which may be too rudimentary.

Nevertheless, we must consider the limitations of this study. Extensions to additional datasets are required to validate and generalize our findings by building machine learning models that take into account more features, without overfitting. This need for larger datasets will also enable a better discrimination as a function of performance. In addition, we will also consider linguistic features (i.e., textual complexity indices), which, in general, are less predictive, but more generalizable.

Despite these limitations, the ultimate value of this extended analysis resides in its potential to provide stealth assessments and scaffolding to students who have not

understood the targeted documents. Feedback can be provided either after self-explaining or after the questions and can include additional interventions – such as functionalities to go back and redo a task, or hints, with the aim to provide better answers (reflecting more coherent understanding of the text). The proposed models also deliver more rapid student assessments that provide valuable insights on understanding performance by estimating how well students are capable of conceptualizing and linking ideas from the initial documents.

Acknowledgments. This research was partially supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0689/"Lib2Life - Revitalizing Libraries and Cultural Heritage through Advanced Technologies" within PNCDI III, the Institute of Education Sciences (R305A180144, R305A180261 and R305A190063), and the Office of Naval Research (N00014-17-1-2300). The opinions expressed are those of the authors and do not represent views of the IES or ONR.

References

1. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-metrix: analysis of text on cohesion and language. Behav. Res. Meth. Instrum. Comput. **36**(2), 193–202 (2004)
2. Allen, L.K., Jacovina, M., McNamara, D.S.: Cohesive features of deep text comprehension processes. In: 38th Annual Meeting of the Cognitive Science Society, pp. 2681–2686. Cognitive Science Society, Philadelphia, PA (2016)
3. Dascalu, M., McNamara, D.S., Trausan-Matu, S., Allen, L.K.: Cohesion network analysis of CSCL participation. Behav. Res. Methods **50**(2), 604–619 (2018)
4. Hastings, P., Hughes, S., Magliano, J.P., Goldman, S.R., Lawless, K.: Assessing the use of multiple sources in student essays. Behav. Res. Methods **44**(3), 622–633 (2012)
5. Nicula, B., Perret, C.A., Dascalu, M., McNamara, D.S.: *Predicting Multi-document Comprehension:* cohesion network analysis. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 358–369. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_30
6. Jurafsky, D., Martin, J.H.: An introduction to Natural Language Processing. Computational LINGUISTICS, and Speech Recognition. Pearson Prentice Hall, London (2009)
7. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
8. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Disc. Processes **25**(2/3), 259–284 (1998)
9. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, ACL 1994, pp. 133–138. ACL, New Mexico (1994)
10. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196. JMLR, Beijing (2014)
11. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: The 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), vol. 14. ACL, Doha, Qatar (2014)

12. Dascalu, M., Crossley, S., McNamara, D.S., Dessus, P., Trausan-Matu, S.: Please readerbench this text: a multi-dimensional textual complexity assessment framework. In: Craig, S. (ed.) Tutoring and Intelligent Tutoring Systems, pp. 251–271. Nova Science Publishers Inc., Hauppauge (2018)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representation in vector space. In: Workshop at ICLR, Scottsdale, AZ (2013)