Human and Artificial Intelligence for the Society of the Future
European Distance and E-Learning Network (EDEN) Proceedings
2020 Annual Conference | Timisoara, 22-24 June, 2020
ISSN 2707-2819
doi: 10.38069/edenconf-2020-ac0008

ASSESSING CRITICAL THINKING IN OPEN-ENDED ANSWERS: AN AUTOMATIC APPROACH

Antonella Poce, Francesca Amenduni, Carlo De Medio, Alessandra Norgini,
Roma Tre University, Department of Educational Sciences, Italy

Abstract

The role of Higher Education (HE) is growingly acknowledged for the promotion of Critical Thinking (CT). Constructed-response tasks (CRT) are recognized to be necessary for the CT assessment, though they present problems related to scoring quality and cost (Ku, 2009). Researchers (Liu, Frankel, & Roohr, 2014) have proposed using automated scoring to address the above concerns. The present work is aimed at comparing the features of different Natural Language Processing (NLP) techniques adopted to improve the reliability of a prototype designed to automatically assess six sub-skills of CT in CRT: use of language, argumentation, relevance, importance, critical evaluation and novelty (Poce, 2017). We will present the first (1.0) and the second (2.0) version of the CT prototype and their respective reliability results. Our research question is the following: Which level of reliability are shown respectively by the 1.0 and 2.0 automatic CT assessment prototype compared to expert human evaluation? Data collection is realized in two moments, to measure respectively the CT prototype 1.0 and 2.0 reliability from a total of 264 participants and 592 open-ended answers. Two human assessors rated all of these responses on each of the subskills on a scale of 1-5. Similarly, NLP approaches are adopted to compute a feature on each dimension. Quadratic Weighted Kappa and Pearson product-moment correlation were used to evaluate the between-human agreement and human-NLP agreement. Preliminary findings based on the first data set suggest adequate level of between-human rating agreement and a lower level human-NLP agreement ($r > .43$ for the subscales of Relevance and Importance). We are continuing the analysis of the data collected in the 2nd step and expect to complete them in June 2020.

Introduction

Despite the scepticism toward the possibility to objectively assess Critical Thinking (CT), CT skills are considered a desirable learning outcome in all the level of education, included Higher Education (HE), according to economic (OECD, 2012) cultural (UNESCO, 2015)

and educational research-oriented organizations (IEA, 2018). In response to the Bologna Declaration of 1999 aimed at developing a comparable degree system among European countries, the Tuning Projects identified different general and subject specific skills to develop in HE students, included CT (Gilpin & Wagenaar, 2008). The AHELO project carried out by OECD (2012) also included CT as one of the general skills that should be assessed at an international level. Thus, reflecting upon CT assessment choices is necessary at least for two reasons: firstly, CT is considered a desirable learning outcome for European HE students and should be assessed and recognized in a comparable way, according the Bologna Strategy; secondly, research is necessary to understand which teaching strategy can foster CT skills in HE. As asserted by Rear in a recent review (2019), the assessment of CT has become a significant enterprise with a number of standardized test available. Assessment tests could be classified in different ways. Hyytinen, Nissinen, Ursin, Toom, and Lindblom-Ylänne, (2015) differentiated self-report from performance-based measurements. Moreover, the performance-based measurements can be classified into multiple choice tests (MCT) / questionnaires and constructed response tasks (CRT). Although there is evidence that by applying a well-designed MCT it is possible to measure higher order skills, MCT cannot assess student's skill to synthesise or generate own answers, necessary components of CT (Ennis, 1987; Facione, 1990). To address this limitation, new CT assessment incorporates both CRT and MCT. CRT are often open-ended tasks in which students need to analyse, evaluate and synthesise complex information as well as provide reasoned explanation. Although CRT are recognized to be necessary for the CT assessment, they present problems related to inter-rater reliability and high-cost of scoring (Ku, 2009). Automated scoring could be a viable solution to the above concerns (Liu, Frankel, & Roohr, 2014). Recent research describe the development and the validation of automatic tools for the assessment of CT sub-skills, such as reasoning (Mao et al., 2018) or argumentation (Song, Heilman, Klebanov, & Deane, 2014). Having said that, there are still open-challenges in terms of validity and reliability of the measures.

The present work is aimed at comparing the features of different NLP techniques adopted to improve the reliability of a prototype designed to automatically assess CT in CRT. In the present work, we will present the first (1.0) and the second (2.0) version of the prototype and their respective reliability results. Our research questions is the following:

- Which level of reliability are shown respectively by the 1.0 and 2.0 automatic CT assessment prototype compared to expert human evaluation?

This research is aimed at developing a prototype which can assess six indicators in open-ended answers: use of language, argumentation, relevance, importance, critical evaluation and novelty (Newmann, Webb, & Cochrane, 1997; Poce, 2017).

The first macro-indicator, namely *use of the language*, is useful to assess the language form of the text. The macro-indicator called *justification* evaluates students' ability to elaborate on their thesis and support their arguments throughout a discourse. *Relevance* is a macro-indicator that analyses consistency in the texts produced. For instance, it refers to the correct use of outlines and to the capability to accurately use given stimuli. The macro-indicator called *importance* evaluates the knowledge used in discourses. *Critical evaluation* assess personal and critical elaboration of the sources, data and background knowledge. Finally, *novelty* concerns the development of new ideas and solutions based on the initial hypothesis and personal thesis. Even though different tools have been developed to automatically assess one or more of these sub-skills, this prototype has been developed to assess them together, based on different NLP techniques and Open Source tools and databases.

The CT assessment prototype 1.0

The CT prototype 1.0 was designed to assess four areas out of six: use of the language, relevance, importance and novelty. This version of the prototype at the moment works only with English Language.

The system is composed by four main modules:

- **A security module**: the module has been implemented by using the Spring Framework (https://spring.io/projects/spring-framework), an open source application adopted to automatically configure security processes, such as authentication and authorization.
- **Question / answer manager**: through this module it is possible to insert the questions and the answers to assess. For each question, in addition to the title and the text of the question, users are also asked to include words representing the *concepts* and the *successors*. Concepts could be defined as the topics that should be covered in a correct and exhaustive answer. Successors represent, instead, deepening or related topics of the given concepts.
- **Human evaluation input module**: Through this module, expert assessors can manually evaluate the answers. For each answer, it is possible to associate one or more anonymous evaluation; these evaluations will be compared with the automatic evaluations to verify the reliability of the proposed approach.
- **CT automatic evaluator**. The last module is at the heart of the system. To evaluate the *Use of language*, the prototype calculates the number of misspellings and obtains the correct version of the text, using an external service, the JLanguageTool. The *Importance* is assessed by extracting the concepts contained in the text of the question and in the answer using the Tagme service and after that execute the

intersection between those sets of word. The *Relevance* and the *Novelty* are obtained by crossing the concepts extracted from the answers in the previous calculation and crossing respectively with the concepts and successors defined in the creation of the question. To improve the precision of the calculations, the prototype applies the n-gram calculation to the sets and recalculates the intersections.

The CT assessment prototype 2.0

The CT assessment porotype presents the same general infrastructure of the previous version. However, two main innovations have been introduced: (a) the attempt to include the automatic assessment of the *argumentation* and *critical evaluation* indicators, (b) the adaptation of the prototype to the Italian language.

To assess *Use of language*, the prototype calculates: (a) misspelling and grammatical errors, (b) frequency of words and (c) lexical diversity. *Argumentation* is assessed training the prototype at distinguishing discourse categories, checking: tense verbs; polarity, and arguing lexicon. Human judges could also annotate hundreds of essays, so that the machine is facilitated at recognizing the discourse structure typical of persuasive writing. *Relevance* is evaluated using Latent Semantic Analysis (LSA), a statistical model of language learning and representation, based on the idea that the semantic similarity of words is reflected by the way they co-occur in a text. *Importance* is obtained by means of Intelligent Essay Assessor (Landauer, Laham, & Foltz, 1999). IEA is based on LSA; it makes a comparison between the semantic content of previously scored essays to esteem the score which the essay under analysis is nearer to. Since we hypothesize that the better the *Critical evaluation* of the writer, the deeper the parse tree of his sentences and the larger his use of persuasive syntactic patterns (e.g. ADV + ADJ + CONJ + ADJ), the prototype uses The Italian NLP Tool to analyse the syntactic trees of the essays' sentences under study. *Novelty* is assessed through LSA and TF-IDF (Term Frequency-Inverse Document Frequency). LSA checks words which co-occur in a context in which they usually do not. TF-IDF calculates the weight of a word assigning the importance to that word based on the number of times it appears in a document and in similar documents of the same corpus: the smaller the weight, the more common the term; the higher the numerical weight value, the rarer the term.

Data collection and analysis

Data collection is realized in two moments, to measure respectively the CT prototype 1.0 and 2.0 reliability.

The first experimentation was aimed at collecting evidence on CT prototype 1.0 reliability. Data were collected with a group of 64 university international teachers after workshops carried out in the USA and Belgium. Participants were required to answer to different

kinds of CRT. Since the context of the workshop was international, participants were required to write their answers in English. The task requires to read an extract from the Galilei's book "Dialogue on the Two Chief World Systems" and then to write a paraphrase, a comment and a critical analysis (Paul & Elder, 2006). The second experimentation was aimed at collecting evidence on CT prototype 2.0 reliability. Data were collected with a group of 200 Italian university students at the beginning and at the end of an annual university course in Experimental Education. Participants were required to read an extract from the Galilei's book "Dialogue on the Two Chief World Systems" and then write a short essay (Poce, 2017). Thus, they produce a total of 400 hundred essays.

In both the experimentation, two human assessors rated all of these responses on each of the on a scale of 1-5. Similarly, one of the two versions of the CT prototype was adopted to compute a feature on each dimension. Quadratic Weighted Kappa and Pearson product-moment correlation is adopted to evaluate the agreement between the human raters' scores and between human raters and the two versions of the CT prototype, as a measure of reliability.

Preliminary results

The rubric for CT assessment shows good properties, with satisfactory reliability between two human raters (see Table 1)

Table 1: The agreement among human raters regarding the indicators use of language, relevance and importance respectively in the paraphrase and in the commentary.

Macro-indicator	H-H Correlation	H-H Quadratic Weighted Kappa
Paraphrase_Use of Language	0.911*	0.83*
Commentary_Use of Language	0.745*	0.618*
Paraphrase_Relevance	0.75*	0.682*
Commentary_Relevance	0.881**	0.811*
Paraphrase_Importance	1.000**	1.000*
Commentary_Importance	0.642	0.571

*sign. <0.05 **sign<0.001

The best correlation among human raters and CT prototype 1.0 were obtained for the macro-indicators *Relevance* ($r = 0.47$) in the commentary and *Importance,* both in the paraphrase ($r = 0.45$) and commentary (0.45). However, the overall reliability could be not considered satisfactory yet (Poce, Amenduni, De Medio & Re, 2019).

In Figure 1, it is shown that in paraphrase the prototype provides higher score than human raters for the macro-indicators *Use of Language* and *Relevance*. On the other hand, the average score for the indicator *Importance* is slightly higher for human raters than in the prototype. In the commentary, there is a general trend of the prototype to provide lower scores comparing to the human raters. However, it is possible to see that the differences

between the average scores for the *Use of Language* scores and *Novelty* in the commentary is quite low.

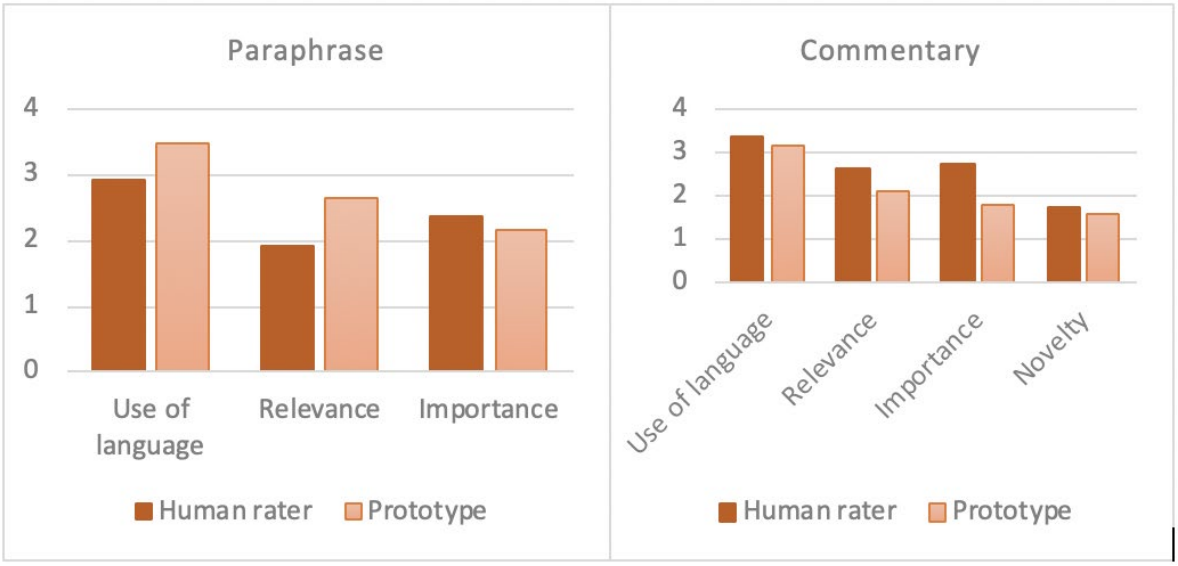

Figure 1. A comparison of CT scores calculated by a human rater and the prototype in paraphrase and commentary

Discussion and conclusive remarks

In line with previous research (Liu et al., 2014), human raters tended to assign higher scores than our CT assessment prototype 1.0 in the commentary. On the other hand, in the paraphrase the prototype assigned higher scores than human raters on the macro-indicators *Relevance* and *Importance.* This result could be explained because the prototype is designed to infer concepts from the questions and answers texts. In the paraphrase, the participants are required to report all the text's topics. In this condition, the prototype easily identifies all the concepts, without the need of further analysis. For these reasons, in paraphrase exercise the macro-indicators *Relevance* and *Importance* could obtain higher scores than the other macro-indicators and, more in general, than commentary or argumentation texts. This data leads us to think that it may be necessary to apply changes to the evaluation of the macro-indicators based on the type of stimulus given to the participants (paraphrase, argumentation, commentary, poetry).

We are continuing the analysis of the data collected in the 2nd step and expect to complete in June. Though early findings of this study suggest that the NLP approach appears to have a lower level of rating quality than human raters, more research seems necessary to explore features and possibilities to improve such rating quality in the future.

References

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Series of books in psychology. Teaching thinking skills: Theory and practice* (p. 9–26). W H Freeman/Times Books/ Henry Holt & Co.

Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (The Delphi Report).

Gilpin, A., & Wagenaar, R. (2008). *Approaches to Teaching, Learning and Assessment in Competence based Degree Programmes. Tuning Educational Structures in Europe.* Universities' Contribution to the Bologna Process. An Introduction, 91-118.

Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation, 44*, 1-8.

IEA. (2018). *IEA International Computer and Information Literacy Study 2018 assessment framework.* Springer Switzerland.

Landauer, T. K, Laham, D., & Foltz, P. W (1999) The Intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer – Enhanced Learning.*

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series, 2014*(1), 1-23.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*, 19–28. doi:10.1111/emip.12028

Mao, L. Liu, O., L., Roohr, K., Belur, V., Mulholland, M., Lee, H., & Pallant, A. (2018): Validation of Automated Scoring for a Formative Assessment that Employs Scientific Argumentation. *Educational Assessment, 23*(2), 121-138. doi: 10.1080/10627197.2018.1427570

OECD. (2012). *Assessment of Higher Education Learning Outcomes – Feasibility Study Report.* Volume 1 – Design and Implementation. Retrieved from http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume1.pdf

Paul, R.W., & Elder, L. (2006). *Critical Thinking Reading & Writing Test.* Tomales, CA: The Foundation for Critical Thinking.

Poce, A. (2017). *Verba Sequentur. Pensiero e scrittura per uno sviluppo critico delle competenze nella scuola secondaria.* Milano: Franco Angeli.

Poce, A., Amenduni, F., De Medio, C., & Re, M. R. (2019). Road to Critical Thinking automatic assessment: a pilot study. *Form@ re-Open Journal per la formazione in rete, 19*(3), 60-72.

Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. *Assessment & Evaluation in Higher Education, 44*(5), 664-675.

Ryser, G. R., Beeler, J. E., & McKenzie, C. M. (1995). Effects of a Computer-Supported Intentional Learning Environment (CSILE) on students' self-concept, self-regulatory behavior, and critical thinking ability. *Journal of Educational Computing Research, 13*(4), 375-385.

Saadé, R. G., Morin, D., & Thomas, J. D. (2012). Critical thinking in E-learning environments. *Computers in Human Behavior, 28*(5), 1608-1617.

Song, Y., Heilman, M., Klebanov, B. B., & Deane, P. (2014, June). Applying argumentation schemes for essay scoring. *Proceedings of the First Workshop on Argumentation Mining*, 69-78.

UNESCO. (2015). *The future of learning 2: what kind of learning for the 21st century skills?* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000242996

About the Authors

A. Poce coordinated the research presented in this paper. Research group is composed by the authors of the contribution that was edited in the following order: A. Poce (Introduction, Discussion and Conclusive Remarks) F. Amenduni (Data collection and analysis; primary results) C. De Medio (The CT assessment prototype 1.0). A. Norgini (The CT assessment prototype 2.0)