Form@re
Open Journal per la formazione in rete

articoli

Road to Critical Thinking automatic assessment: a pilot study

Verso la valutazione automatica del pensiero critico: uno studio pilota

Antonella Poce[a], Francesca Amenduni[b], Carlo De Medio[c], Maria Rosaria Re[d,1]

[a] Università degli Studi Roma Tre, antonella.poce@uniroma3.it
[b] Università degli Studi Roma Tre, francesca.amenduni@uniroma3.it
[c] Università degli Studi Roma Tre, carlo.demedio@uniroma3.it
[d] Università degli Studi Roma Tre, mariarosaria.re@uniroma3.it

Abstract

The growing attention on Critical Thinking as an essential driver for progress and knowledge growth has brought to the development of different kinds of assessment tools. Essays and open-ended questions are recognized to be pivotal in critical thinking assessment. However, they present problems related to inter-rater reliability and high-cost of scoring. Automated scoring could be a viable solution to the above concerns. In this paper, we introduce a prototype for critical thinking automatic assessment based on Natural Language Process techniques and preliminary accuracy evidence regarding its use. Data were collected from 48 university teachers after two workshops carried out both in the United States and in Italy aimed at developing critical thinking Skills in participants.

Keywords: critical thinking; automatic assessment; open-ended measures.

Sintesi

La crescente attenzione nei confronti del Pensiero Critico come competenza cruciale per l'innovazione e lo sviluppo di conoscenze hanno portato allo sviluppo di un notevole numero di strumenti di valutazione. Sebbene gli strumenti a stimolo aperto, come il saggio breve e le domande aperte, siano ritenuti essenziali per la valutazione del pensiero critico, essi presentano problemi relativi all'attendibilità inter-giudice e all'alto costo di valutazione. L'assegnazione automatica di punteggi potrebbe rappresentare una possibile soluzione a tali problemi. Un prototipo per la valutazione automatica del pensiero critico, basato su tecniche di elaborazione del linguaggio naturale ed evidenze preliminari sull'accuratezza dello strumento verranno presentati all'interno del presente articolo. In seguito a due workshop condotti rispettivamente negli Stati Uniti e in Italia, che miravano alla promozione del pensiero critico, i dati sono stati raccolti all'interno di un gruppo composto da 48 professori universitari.

Parole chiave: pensiero critico; valutazione automatica; risposte aperte.

[1] 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: Poce par. 1, 2, 3.1, 3.3, 5; Amenduni par. 4.2, 4.3; De Medio par. 3.2; Re par. 4.1.

© 2019 Author(s). Open Access article distributed under CC BY-NC-ND 4.0

1. Introduction

Educational policy makers identify Critical Thinking as an essential driver for progress and knowledge growth in any field and in the broad society. The World Economic Forum (2016) stated that critical thinking would be the second most important skill in 2020 for economic growth. Moreover, UNESCO (United Nations Educational, Scientific and Cultural Organization) includes critical thinking as one of the 21st century core skills (Scott, 2015). Having said that, critical thinking is still a disputed concept with many different definitions that come from different fields such as philosophy, educational sciences and cognitive sciences (Johnson & Hamby, 2015; Moore, 2013). Especially in the philosophical field, traditional perspectives describe critical thinking as an individual process on a reasoning task. For example, in the *Delphi Report* by the American Philosophical Association (Facione, 1990a), critical thinking is conceived as a set of cognitive skills such as interpretation, analysis, evaluation, argumentation, inference and meta-reflection. Traditional definitions failed to consider critical thinking as a crucial skill to navigate into social situations. Indeed, recent perspectives highlight the role of critical thinking in various types of information exchange and symbolic interaction (Byrnes & Dunbar, 2014). The basic idea is that much of our knowledge of the world comes from others, rather than being the result of primary self-experience, and requires an analysis and critical evaluation of sources, internal coherence and relation to other sources of information. According to Kuhn (2019), critical thinking is a dialogic practice people engage in and commit to, initially interactively and then in an interiorized, implicit form with the other. In advancing arguments, well-practiced thinkers anticipate their defeasibility as a consequence of others' objections, in addition to envisioning their own potential rebuttals. As a consequence, language skills are important precursors for critical thinking Development. By regulating thoughts through internal speech and navigating social situations through external speech, language helps people process information at increasingly sophisticated levels over time, providing foundation to be engaged in critical thinking (Kuhn, 1991). The emphasis on language-based activities as precursors for critical thinking has an impact both on pedagogical practices and critical thinking assessment.

According to recent perspectives on critical thinking definition, in this paper a new critical thinking assessment method is presented. The research group from the Centre for Museum Studies (CDM, http://centrodidatticamuseale.it/en/), has been developing and validating a prototype for the automatic assessment of critical thinking in open-ended questions through Natural Language Processing techniques based on a rubric previously adopted for qualitative content analysis (Poce, 2017). The purpose of this study was to collect preliminary validity evidence regarding the use of the above-mentioned critical thinking assessment tool.

2. Critical thinking Assessment

The general lack of agreement on critical thinking definition led to the production of different assessment methods. Indeed, the conceptualization and the assessment of critical thinking are interdependent issues that must be discussed together: the definition of critical thinking determines how to best measure it. The most common tools fall into four categories (Ku, 2009; Liu, Frankel, & Roohr, 2014):

1. multiple choice questionnaires (Facione, 1990b; Watson & Glaser, 1980);
2. open-ended questions (Ennis & Weir, 1985);

3. self-reporting measures (Facione, Facione, & Sanchez, 1994);
4. mixed methods (Halpern, 2007).

The critical thinking assessment has become a significant challenge, with a number of standardised tests available (Rear, 2019) that include mainly multiple-choice tests and self-reporting measures. Moreover, critical thinking is often assessed through researcher or teacher-made tests (Tiruneh, Verburgh, & Elen, 2014).

Although multiple choice tests could guarantee a higher reliability rate, they pose problems in terms of validity (Poce, 2017). More specifically, a satisfactory performance in a prompted thinking context cannot be generalized to contexts where prompts are not given (Ku, 2009). On the other hand, essays and open-ended measures pose problems related with inter-rater reliability and high-cost of scoring. Automated scoring could be a viable solution to these concerns (Liu, Frankel, & Roohr, 2014). There are automated scoring tools designed to score both short-answers to open questions and essays. In short-answer items, automated scoring mainly evaluates the content of the responses (e.g. accuracy of knowledge); on the other hand, in essay questions only the writing quality of the responses is assessed (e.g. grammar, coherence and argumentation). For instance, Gordon, Prakken, and Walton (2007) proposed a functional model for the evaluation of arguments in dialogical and argumentative contexts. Wegerif and his colleagues (2010) described a computational model to identify moments within e-discussion in which students adopted critical and creative thinking. In order to adopt this automatic method for critical thinking assessment, the accuracy of automated scores need to be examined to make sure they achieve an acceptable level of agreement with human scores. However, only few studies have validated automatic scoring test for critical thinking Assessment (Mao et al., 2018). In the context of automatic assessment and classification, validation implies the analysis of accuracy's levels in order to establish the reliability of the method under investigation (Grimmer & Stewart, 2013).

Liu, Frankel, & Roohr, (2014) evaluated the performance of an automatic scoring system on four short-answer items used in middle school science classes. The results showed that human raters in most cases tended to assign higher scores than automatic assessment. Mao and colleagues (2018) found that automated scores showed satisfactory agreement with human scores, but small discrepancies still existed. From our perspective, more research is needed in terms of development and validation of automatic tools for critical thinking assessment (Lewis Sevcikova, 2018).

Starting from these assumptions, the CDM research group elaborated a prototype for critical thinking assessment in open-ended questions and short essay based on six different indicators (use of the language, justification, relevance, importance, critical evaluation and novelty) (Poce, 2017). The prototype has been adopted to develop an automatic critical thinking assessment tool composed by the same six indicators and aimed at overcoming the problems related to critical thinking assessment in open-ended questions. Human evaluators together with the automatic assessment tool and the comparison between the two kind of assessment results are adopted to verify the level of reliability of the test and to obtain useful data for the implementation of the designed critical thinking prototype.

The purpose of this study was to collect preliminary validity evidence regarding the use of our critical thinking assessment method. Thus, we tried to answer the following research questions: which reliability levels are shown respectively through the human and the automatic assessment processes?

3 Methods and analysis

3.1. Study design and data collection

In order to collect preliminary validity evidence regarding our prototype for automated critical thinking assessment, data were collected during two workshops carried out respectively in the United States and in Italy. The activities were designed according to a general structure inspired by the *Crithinkedu* project[2] training course (Dominguez, 2018):

1. in the United States, the workshop took place in the setting of the 6th Annual Conference "Defining Critical in the 21st Century?"[3] at Berkeley College, NYC. The conference was devoted to critical thinking in Higher Education and university teachers participated to improve their critical thinking teaching and professional practices;
2. the workshop conducted in Italy took place in the framework of the Roma TRE University "Inclusive Memory"[4] project. Local university teachers, from different fields, were involved in learning activities aimed at developing their critical thinking knowledge, skills and dispositions and to design inclusive learning paths to be used in their own courses.

Data were collected after the two workshops through an online questionnaire developed and adapted in the above-mentioned *Erasmus + Crithinkedu* project. We received 22 answers from the Italian group and 26 answers from the US group. The questionnaire included both open-ended and multiple-choice questions. The tools presented closed questions regarding the following topics:

- personal details and information;
- departments and subject field (STEM, humanities, social sciences);
- kind of skills and dispositions they meant to develop within their classes.

At the end of the questionnaire, the following open questions were inserted:

- (Q1) mention max. 3 activities that you would adopt in your teaching to promote critical thinking. Please, also mention why you decided to include those activities in your course;
- (Q2) in what way do you think the planned activities would affect participants' critical thinking?
- (Q3) In what way could participants' critical thinking development contribute to the achievement of other learning objectives?

In order to detect critical thinking levels shown by university teachers, we analysed the answers to the open questions mentioned above by comparing human assessment with the one carried out by our prototype for the automatic assessment of critical thinking.

---

[2] The project was meant to enhance critical thinking teaching and learning in Higher Education http://crithinkedu.utad.pt/it/cosa-e-crithinkedu/

[3] https://ccrwt.weebly.com/2018-ccrwt.html

[4] The project is aimed to support inclusiveness of minorities and disadvantaged groups through the fruition of cultural heritage in museums and through the development of the 4Cs (Collaboration, Creativity, Communication and critical thinking).

3.2. A prototype for the automatic assessment of critical thinking

The prototype is based on a rubric developed in previous researches (Poce, 2017) and is mainly based on the model by Newman, Webb, and Cochrane (1995). The rubric (Figure 1) is based on six macro-indicators: *use of language, justification, relevance, importance, critical evaluation* and *novelty*. All the macro-indicators are scored on a five-point scale.

Macro-indicators	Indicators	Descriptors
Use of the language	Language ability (punctuation, spelling, morphosyntax, lexicon)	5. rich and original 4. appropriate 3. mainly correct 2. not precise 1. not correct and improper
Justification Argumentation	Elaboration ability (thesis definition and elements of reasoning)	5. rich and articulate 4. clear and ordered 3. too synthetic 2. quite consistent 1. inconsistent
Relevance	Consistency (the topic under issue is mentioned)	5. complete, deep and original 4. complete and correct 3. generic 2. partial 1. out of line
Importance	Knowledge of the topic (main issues related to the topic are mentioned)	5. deep and critical 4. complete 3. appropriate 2. superficial 1. not sufficient
Critical evaluation	Personal and critical elaboration of sources and background	5. critical and well sounded 4. wide and adequate 3. essential and simple 2. partial 1. contradictory
Novelty	New information, ideas and solutions are added to discuss the issues raised in the questions	5. widely, critically and originally 4. in detail 3. correctly 2. simply and or partially 1. no new information and solutions are added

Figure 1. Rubric to assess critical thinking in essays and open-ended questions.

At the moment, the prototype has been designed to assess four macro-indicators out of six: use of language, relevance, importance, and novelty. The compound system is composed by four main modules that allow to perform all the operations necessary to obtain the experimental results. Figure 2 describes the four modules of the system.

- Security module. An open source Security Framework application has been implemented to automatically set security processes, such as authentication and authorization. Every operation within the system is logged anonymously in order not to affect the interactions with the system. The module allows online registration via email and provides a secure login form to access the services offered;

- Question/Answer input manager. The module manages the insertion of the questions and answers to be evaluated. For each question, in addition to the title, the text of the question and a *golden answer* are to be inserted. At the moment the use of the golden answer is still in the experimental phase; the goal is to automatically infer the concepts of importance and successors automatically from this answer. Presently, we are working to enable the system identifying these two sets of concepts. Users are also asked to include words representing the *concepts* and the *successors* respectively for the evaluation of importance and novelty. 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. Concepts and successors will be used by the automatic response analysis module to evaluate the four critical thinking indicators. It is possible to insert more questions or answers at the same time using the import function from google forms and uploading the generated xml file. The module interacts with Hibernate, a framework for the automatic management of entities in the local database where all the questions and answers are saved;
- Human evaluation input module. Through this module, field experts can manually evaluate the indicators for the answers entered. It is possible to select any question on the system and this latter generates all the answers to be evaluated. The user can then decide whether to evaluate the answer or assign it to another teacher. For each question, it is possible to associate one or more anonymous evaluation; these evaluations will be compared with the automatic evaluations to verify the effectiveness of the proposed approach;
- critical thinking automatic evaluator. This module is the heart of the system which uses two external tools to perform the automatic evaluation of the four indicators presented;
  - *Use of Language.* The system uses an external tool, the JLanguageTool (https://languagetool.org) a java module to query the https://languagetool.org API, which allows you to send texts and receive information on grammatical errors within just a few milliseconds. It also allows you to receive a version of the text with the most probable revisions. This correct version of the text is fundamental for a more advanced analysis since an incorrect text introduces noise that lowers the performance of the whole system. The value of the indicator is given by normalizing the number of errors considering the number of words contained in the answer;
  - *Relevance.* The indicator is assessed carrying out an analysis of the *concepts*. The text is processed by a Part of Speech Tagger, a software that extracts entities such as nouns and verbs from any kinds of text. After a stemming process that reduce the words to their root, an algorithm is applied on this set of nouns by generating n-grams with a length from one to three.

    Taken a text the set of 1-grams is composed of all the single words taken in order as they appear in the text, while the 2-grams are the set of all words taken in pairs and thus the 3-grams;

    For example, take the sentence:

    “All mice love cheese” we can create the three sets in the following way:

    1. grams: “all”, “mice”, “love”, “the”, “cheese”;
    2. grams: “all mice”, “mice love”, “love the”, “the cheese”;
    3. grams: “all mice love”, “mice love the”, “love the cheese”.

The sets generated this way are compared with the concepts defined as necessary for a good answer by the human rater.

The number of the intersection between the n-grams and the concepts will give the relevance of the answer;

- *Importance*. The system exploits an open source knowledge base (Wikipedia (https://en.wikipedia.org) and Wiki Data (https://www.wikidata.org). Initially, the text of the answer is sent to an online tagging service through entities pages, which are the Wikipedia pages associated with the concepts extracted from the tagging service from the application. The service returns a set of entities pages associated with a given text, in our case the text of the answer. Afterwards, each defined concept is automatically linked to its page. All the outgoing links of this page are considered. The importance indicator is given by the number of known pages that the tagging service system detects respectively from the answers given by the participants and from the concepts defined by the assessor/researcher;
- *Novelty*. The indicator is assessed by carrying out an analysis of the *successors*. As for the relevance indicator, all the nouns and n-grams are extracted from the answers' texts. The frequency of intersections between n-grams and successors results in the *novelty* dimension of the answer. To manage the issue of different languages, in order to obtain more accurate values, we used a module (made of different language specific models) able to detect the inflected forms of word.

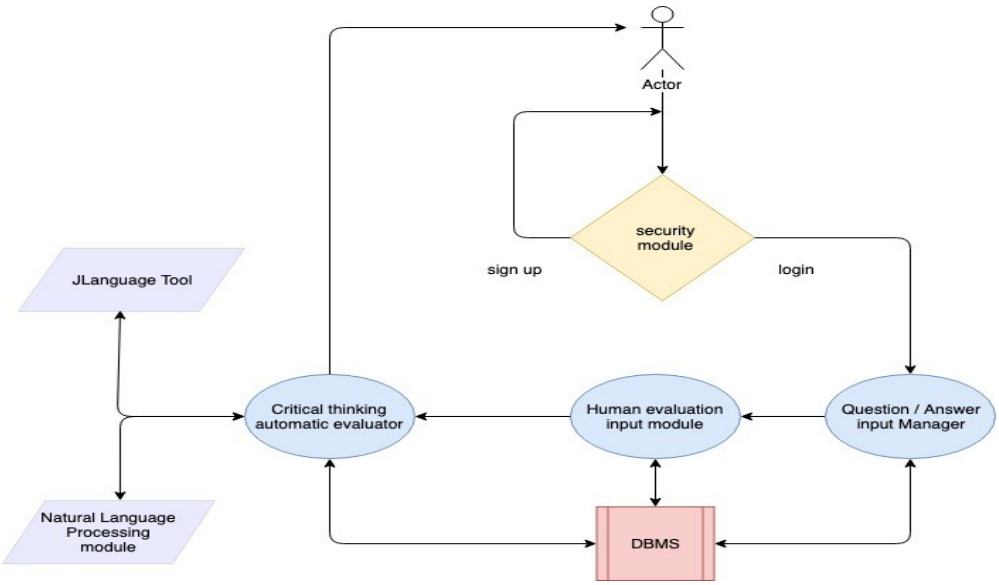

Figure 2. The four modules of the system.

3.3. Data analysis

The first step to build automated scoring models is to score responses by human raters. The 3 answers for 48 university teachers (22 participating in the Italian workshop and 26 in the American one) were scored by two trained human raters with prior experience with the critical thinking scoring rubric. Then, we calculated the following indicators: the Pearson product-moment correlation and accuracy to evaluate the agreement between the two

raters' scores and between human rater and automatic assessment (Fleiss & Cohen, 1973). Pearson correlation and accuracy are two criteria that can be used to evaluate consistency between two raters. For this first evaluation of the prototype, we analysed only the US group because our prototype is supported by an English open source knowledge base. In the future we are going to extend the approach to different languages.

4. Results

4.1. Level of critical thinking shown by university teachers

Most of the teachers (Figure 3) are based in the field of humanities (43%), social sciences and education (33%). A minor percentage comes from business and political studies (12%), STEM (8%) and health science (4%).

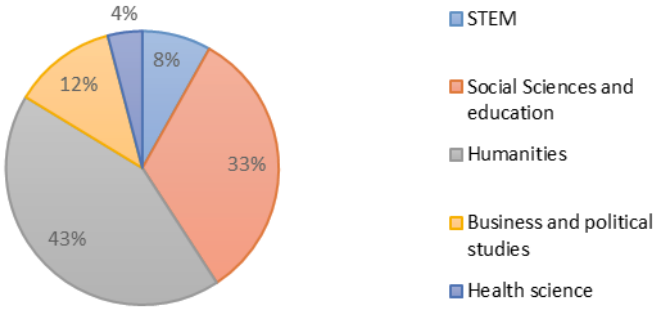

Figure 3. Discipline sectors of the teachers involved in the analysis.

The two groups of teachers from Italy and USA have achieved similar total scores and had similar performance in the three open questions (Figure 4). For all the questions the average score is higher than 17 on a maximum of 30. In addition, the average total score is higher than 53 on a maximum of 90 for both the groups.

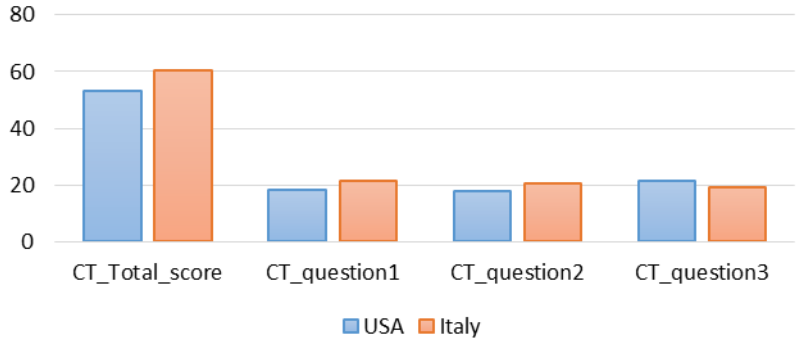

Figure 4. Critical thinking level in American and Italian group.

We also compared the performance of university teachers who come from different fields (Figure 5). Health science group of teachers obtained the highest total average (67 on a maximum of 90) whilst social sciences and education groups obtained the lowest average score (42.5 on a maximum of 90). However, differences among groups were not statistically significant.

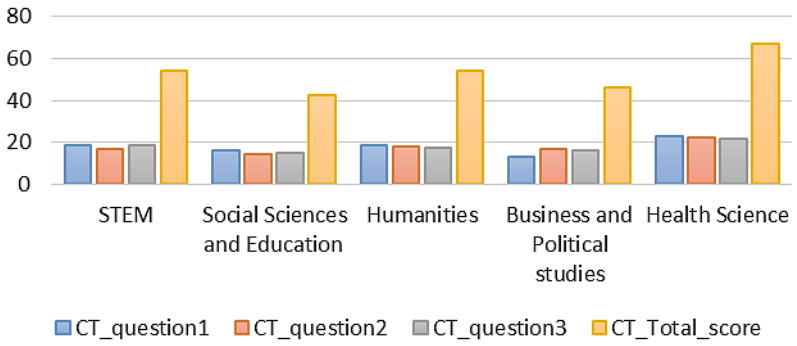

Figure 5. Critical thinking level for different discipline sector.

4.2. Level of reliability of the human and the automatic assessment

The agreement between scores from two human raters is shown below (Figure 6). Almost all the items show a satisfactory correlation (i.e., $r > 0.69$) between the scores from the two human raters. The results suggested a good reliability of the assessment method when it is performed by human raters. On the other hand, there were not significant correlations among each human rater and the automatic assessment prototype.

Item	H-H Correlation	Sign
critical thinking – Question 1	0.785	0.000
critical thinking – Question 2	0.690	0.000
critical thinking – Question 3	0.744	0.000
critical thinking – Total	0.866	0.000
Use of Language	0.749	0.000
Relevance	0.873	0.000
Importance	0.807	0.000
Novelty	0.725	0.000

Figure 6. Pearson's correlation among two human raters.

Manual evaluation is slightly higher than evaluation calculated by the prototype (Figure 7). At the moment, the prototype agreed with the domain expert in 30% of cases. Analysing only a sub-sample of the dataset, the one with the best answers (more complete and longer in terms of words), the value of agreement rises to almost 34%. The best automatic evaluations were obtained for the *Use of Language* and *Importance* indicators with accuracy values of 67% and 39% respectively.

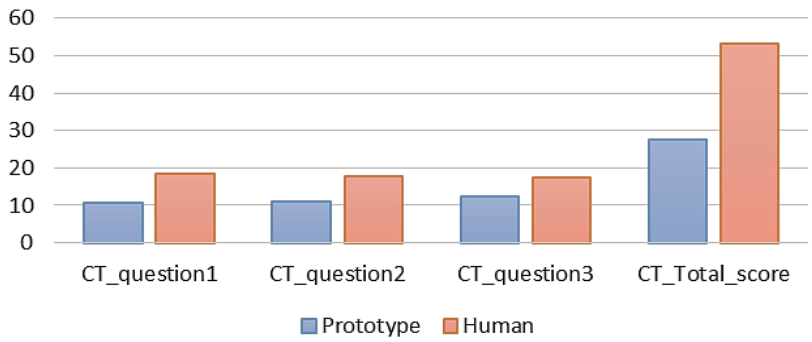

Figure 7. Critical thinking level assessed manually and automatically.

4.3. Macro-indicators properties in the human and automatic assessment

The macro-indicators total correlation in the human assessment are presented below (Figure 8). All the macro-indicators in the three questions showed high correlation with total critical thinking score and therefore they all had good discriminating power.

Macro-indicator	Macro-indicators-Total Correlation	Sign.
UOL_Q1	0.860**	0.000
ARG_Q1	0.887**	0.000
REL_Q1	0.878**	0.000
IMP_Q1	0.873**	0.000
CE_Q1	0.903**	0.000
NOV_Q1	0.845**	0.000
UOL_Q2	0.791**	0.000
ARG_Q2	0.833**	0.000
REL_Q2	0.845**	0.000
IMP_Q2	0.836**	0.000
CE_Q2	0.855**	0.000
NOV_Q2	0.707**	0.000
UOL_Q3	0.842**	0.000
ARG_Q3	0,922**	0,000
REL_Q3	0,9**	0,000
IMP_Q3	0,906**	0,000
CE_Q3	0,899**	0,000
NOV_Q3	0,872**	0,000

Figure 8. Human assessment Macro-indicators correlation. UOL = Use of Language; ARG = Argumentation; REL = Relevance; IMP = Importance; CE = Critical Evaluation; NOV = Novelty. sign < 0.05*; sign <0.01**.

We wanted to compare the macro-indicators properties of the human and automatic assessment (Figure 9). The following table presents the prototype's macro-indicators total correlation. In four cases out of twelve, the macro-indicators showed from moderate to high correlation with total score. However, in other cases the correlation between macro-indicators and total scores was not significant.

Item No.	Macro-indicators	Macro-indicators-Total Correlation	Sign.
1	UOL_Q1	0.575**	0.003
3	IMP_Q1	0.495*	0.14
5	UOL_Q2	0.595*	0.03
10	REL_Q3	0.478*	0.018

Figure 9. Macro-indicators total correlation in the automatic assessment. UOL = Use of Language; REL = Relevance; IMP = Importance; sign < 0.05*; sign < 0.01**

5. Discussion and conclusions

Taking into consideration the starting research questions, for the sake of the present contribution, some final remarks can be made. First of all, data collected and presented in

this paper are limited to a pilot activity with a small number of participants (48 in total), so any generalization is not possible.

In the sample analysed, university teachers' critical thinking performance results are satisfactory in both the groups considered. The rubric for critical thinking assessment shows good properties, with satisfactory reliability between two human raters ($r > 0.69$). However, the results of the prototype validation are not satisfactory yet for an effective classification considering the application of the study to the domain (only three question), but they allow to identify different kinds of improvement of the approach. An analysis of the negatively classified instances highlights some evidence: the process of defining the associated concepts to the *Importance* indicator must be very specific, otherwise the system cannot evaluate the indicator correctly because general concepts lead the system out of topic in the analysis. Moreover, it has been found that the more general the question is, the more the system performance worsens in calculating the relativity of the answer, due to the number of concepts found in the open knowledge base not related to the question.

Moreover, in the *novelty* calculation it can be interesting to apply weights to the successors: in the case in which one or more of the successors have been considered by all the participants the weight will be reduced. On the contrary, for the successors treated only by a small part of users we will associate an augmented weight. As shown by Liu, Brew, Blackmore, Gerard, Madhok, and Linn (2014), human raters tend to assign higher scores than automatic assessment tools. In the future, in order to increase the accuracy of the scoring it may be interesting to analyse a semantic database for a better contextualization of the questions and answers considered; specifically, it could be interesting to extract the set of associated concepts and travel the tree of the open source knowledge base categories.

The attempt to automatize critical thinking assessment through open-ended questions is at its beginning but it proves to be a useful support to human evaluation. The use of Language analysis procedures seems to be a possible direction according to the first results collected in the study herewith presented. The research group feels therefore encouraged to follow up the research described above, through further experimentation, working also on different macro-indicators from the Newman et al. (1995) adapted model used so far. In following studies, it is foreseen to work more deeply on the choice of the questions to be analysed and evaluated by the automatic system, taken into consideration that golden answer, concepts and successors are crucial to determine the tool validity. In particular, it will be mandatory to explore the linguistic different dimensions of the texts used to assess critical thinking to improve both the system employability and its reliability value.

In future studies, we are going to expand the textual corpus because our prototype achieved slightly better performances with longer and more elaborated open-answers. We will conduct further validation studies with a larger sample and with different kinds of questions.

In addition, we will clarify which mental operations are necessary to detect critical thinking in open-answers through task analysis and we will use the information to improve the design and the functioning of our prototype. Finally, we'll try to analyse the semantic relationships among concepts exploiting other open source knowledge bases.

Reference list

Byrnes, J. P., & Dunbar, K. N. (2014). The nature and development of critical-analytic thinking. *Educational Psychology Review*, *26*(4), 477–493.

CDM. Centro di Didattica Museale. http://centrodidatticamuseale.it/en/ (ver. 10.12.2019).

CRITHINKEDU. *critical thinking Across the European Higher Education Curricula*. http://crithinkedu.utad.pt/en/crithinkedu/ (ver. 10.12.2019).

Dominguez, C. (2018). A European review on *critical thinking* educational practices in Higher Education Institutions. Vila Real: UTAD.

Ennis, R. H., & Weir, E. (1985). *The Ennis–Weir critical thinking essay test*. Pacific Grove, CA: Midwest Publications.

Facione, P. (1990a). *Executive summary of 'The Delphi Report'*. Millbrae, CA: The California Academic Press.

Facione, P. A. (1990b). *The California critical thinking Skills Test.* Millbrae, CA: California Academic Press.

Facione, N. C., Facione, P. A., & Sanchez, C. A. (1994). critical thinking disposition as a measure of competent clinical judgment: The development of the California critical thinking Disposition Inventory. *Journal of Nursing Education*, *33*(8), 345–350.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, *171*(10-15), 875–896.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Halpern, D. F. (2007). *Halpern critical thinking assessment using everyday situations: Background and scoring standards*. Claremont, CA: Claremont McKenna College.

JLanguageTool. *LanguageTool proofreading software*. https://languagetool.org/ (ver. 10.12.2019)

Johnson, R. H., & Hamby, B. (2015). A meta-level approach to the problem of defining 'critical thinking'. *Argumentation*, *29*(4), 417–430.

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking skills and creativity*, *4*(1), 70–76.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Kuhn, D. (2019). Critical thinking as Discourse. *Human Development*, *62*(3), 146–164.

Lewis Sevcikova, B. (2018). Human versus automated essay scoring: A critical review. *Arab World English Journal, 9*(2), 157–174.

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.

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.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138.

Moore, T. (2013). Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education*, *38*(4), 506–522.

Newman, D. R., Webb B., & Cochrane C. (1995). A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology, 3*(2), 56–77.

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

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

Scott, C. L. (2015). The futures of learning 3: What kind of pedagogies for the 21st century. *Education Research and Foresight*, *15*, 1–21.

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, *4*(1), 1–17.

Watson, G., & Glaser, E. M. (1980). *Watson–Glaser critical thinking appraisal.* Cleveland, OH: Psychological Corporation.

Wegerif, R., McLaren, B. M., Chamrada, M., Scheuer, O., Mansour, N., Mikšátko, J., & Williams, M. (2010). Exploring creative thinking in graphically mediated synchronous dialogues. *Computers & Education*, *54*(3), 613–621.

World Economic Forum (19 January 2016). *The 10 skills you need to thrive in the Fourth Industrial Revolution*. https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/ (ver. 10.12.2019).