Towards the conceptualisation of an expert system for evaluating R&D projects. Overview of relevant literature.

Francesco Molinari
Department of Informatics, University of Rijeka, Croatia
mail@francescomolinari.it

***Abstract* – Along the pathway towards defining the conceptual building blocks of my PhD thesis and of the resulting scientific and technological output, I hereby present an overview of academic and industrial literature in three thematic areas: Automated Essay Evaluation (AEE), Rhetorical Structure Theory (RST), and Bayesian Network Analysis (BNA). Due to the vast and very diverse nature of underlying knowledge, and the fact that no previous attempt has ever been made to consider these three areas jointly in support of a scientific or technological endeavor, the basic approach I have followed to collect, analyze and classify retrieved resources has been problem-based rather than aimed to systematize, and product feature-oriented rather than research-driven. This narrower view has resulted into a collection that is certainly partial and may also be lacking some relevant contribution to the three research questions that characterize my thesis.**

Keywords – Automated Essay Evaluation, Rhetorical Structure Theory, Bayesian Network Analysis.

I. INTRODUCTION

The basic idea of my PhD thesis is to explore the feasibility of an expert system supporting the makers – and possibly the reviewers – of R&D and innovation projects in the (self-)assessment of the overall quality of a candidature. By quality I mean fulfilment of the Call's evaluation criteria up to the level required to win the selection. The system is conceived to resemble state-of-the-art Automated Essay Evaluation software [1, 41] – now quite popular especially in the US for the assessment of individual learning outcomes – with some crucial semantic analysis capabilities added. This constitutes an advance over start-of-the-art as most existing tools do not display these capabilities to the required extent. I suggest adopting Rhetorical Structure Theory [24] to define, both theoretically and in a machine-readable fashion, the fuzzy "middleware knowledge" staying between the official template of an R&D and innovation proposal and the logical structure it will be evaluated against. Finally, such logical structure is to be modelled using the methods and tools of Bayesian Network Analysis [15], with the aim of predicting the variations in expected scores resulting from improved clarity in communicating the project's contents and demonstrating compliance with the Call.

This paper presents an overview of relevant literature in the three thematic areas of fundamental interest for my PhD thesis, namely: Automated Essay Evaluation (henceforth: AEE), Rhetorical Structure Theory (henceforth: RST), and Bayesian Network Analysis (henceforth: BNA). These all pertain to the wide and fast-growing field of Text Mining and Natural Language Processing. Due to the vast and very diverse nature of the underlying literature, and to the fact that no previous attempt has been made to consider the three areas jointly as possible background of a scientific and technological endeavor, the approach I have followed to collect, analyze and classify retrieved resources has been problem-based rather than systematic, and product feature-oriented rather than research-driven. This means on the one hand, that the collection of publications presented herein does not cover the state-of-the-art in full, but has been clustered around the three basic research questions of my PhD thesis, that are:

A) to which existing category of software could my proposed expert system be said to belong,

B) how can the system in question be trained to build meaningful bridges between the table of contents of the official template for an R&D or innovation grant application and the logical structure of the evaluation criteria and process, and

C) how can such logical structure be modeled in such a way to formulate a prediction of the possible outcome of evaluation.

This also means that for the particular case of A) – but also and more generally for B) and C) – retrieved resources have prioritized according to their suitability to contribute to shaping the key features of the envisaged expert system, which inevitably led to discard a certain number of parallel research streams, not because of their merits, but only for being less relevant to the proposed research agenda.

This paper is structured as follows: section II presents the results of literature search in the three thematic areas of AEE, RST and BNA. Section III is a brief discussion and conclusion.

II. LITERATURE OVERVIEW

This section is structured in sequence according to the three thematic areas of basic interest for my PhD thesis. All errors are mine.

A. *Automated Essay Evaluation*

Also known as Automated Essay Scoring (AES) or Automated Essay Grading (AEG) or Automated Writing Evaluation (AWE), AEE consists in the use of specialized computer programs to assign grades to short essays written in an educational setting. Essays are short literary compositions on a particular subject, usually in prose and in English language, delivered by students to demonstrate their thematic knowledge and skills, such as synthesis and analysis. The AEE concept was developed in the mid-1960s by the American researcher and professor of education and psychology Ellis Batten Page [30]. However, it is only in the mid-1990s that the progress of Natural Language Processing (NLP) techniques has encouraged the diffusion of a good number of software packages, some of which have started to deal with other languages than English, such as Chinese, Finnish, French, German and Japanese. The following, non-exhaustive list comes from a merge of [1] and [41] state of the art collections, with emphasis given to existing market products or services (Table 1.).

TABLE I. STATE-OF-THE-ART AEE SOLUTIONS

Product name	*Product/Service URL*	*Literature source*
AutoMark	https://www.cambridgeassessment.org.uk/insights/keeping-artificial-intelligence-human-combining-the-power-of-ai-with-the-experience-of-examiners/	[27]
AutoScore	https://github.com/TysonStanley/autoscore	[36]
BETSY	http://edres.org/betsy	[34]
C-rater®	https://www.ets.org/research/topics/as_nlp/written_content/	[20]
CRASE®	https://www.act.org/content/act/en/products-and-services/act-consulting-services/assessment-tools.html#crase	[22]
E-rater®	https://www.ets.org/erater/about	[7]
Intelligent Essay Assessor	https://www.pearsonassessments.com/professional-assessments/products/programs/write-to-learn.html	[10]
IntelliMetric®	http://www.intellimetric.com/direct/	[35]
LEXILE®	https://lexile.com/	[37]
LightSide	https://bitbucket.org/lightsidelabs/lightside/src/default/	[26]
Markit®	https://ihsmarkit.com/products/lead-scoring-solutions.html	[40]
Project Essay Grader	https://pegwriting.com/	[31]
SAGrader	http://www.sagrader.com	[6]

As far as market services are concerned, it can be noticed from the above listing that some of the major education publishing and assessment companies have adopted one or more of these tools in their normal practice – usually in combination with human scoring. The common Machine Learning approach of these state-of-the-art solutions – for what can be inferred from usage, considering that their specific algorithms are usually not disclosed – is to initially "train" the system with a number of past, already scored essays and then repeat the exercise with a new essay, with or without the "supplementary" grading of a human being, bottom line serving as benchmark (Figure 1).

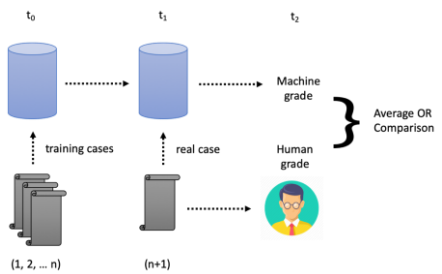

Figure 1. Stylized functioning of AEE systems.

In this sense, we can consider AEE "supervised learning" as a replica of Case Based Reasoning (CBR), another Artificial Intelligence method that gained wide popularity about 25 years ago [18, 14]. While CBR builds upon the principle that "similar problems must have similar solutions", which can also serve as an alternative decision support mechanism [12], AEE takes the goal of scoring a new essay as a prediction challenge, requiring to look back to already existing information to be properly executed. This is normally done by creating a mathematical model (algorithm) the relates the target features of the text to be graded with those of the pre-graded texts and their hand-given scores. Early attempts simply adopted linear regression models, while later tools have started using more sophisticated techniques such as Latent Semantic Analysis [3], Latent Dirichlet Allocation [5], Content Vector Analysis [2], or Bayesian inference [34].

Above and beyond their practical advantages, in terms of time, effort and cost savings, which can explain their success in the educational market, AEE systems have been questioned in their capacity to emulate human scoring and even the results of public verification trials are partly obscured by business confidentiality needs [see the Wikipedia report of the Hewlett Foundation sponsored 2012 competition, entitled Automated Student Assessment Prize, in [42].

However, the key argument in my opinion is not how good they (out)perform, but which features of a text these tools actually take into consideration. In particular, there seems to be consensus in both literature and practice on the three following statements:

(1) Reliability declines with the increasing complexity of a text;
(2) Reliability declines with the increasing length of a text;
(3) Handling of semantic aspects is not comparable with that of syntactic and grammar features.

The three statements together are relevant to define the perimeter of my PhD thesis, which concerns lengthy application forms (2), having the nature of scientific papers (1 and 3), the quality of which needs to be assessed against an unpredictable benchmark – being original R&D papers, there is no way to collect previous examples to compare the new ones which – making the CBR-like or supervised learning approach ineffective. This paves the way to an intermediate approach, which must be "unsupervised" in the sense of not requiring previously graded cases to compare, while at the same time relying with alternative ways of Machine Learning, focusing more on the (appraisal and evaluation of) meaning of long and structured texts – luckily structured according to predefined schemas, as it normally happens when dealing with academic papers or R&D and innovation grant applications.

B. *Rhetorical Structure Theory (RST)*

In the direction outlined above, an interesting research stream – also common to both AEE and CBR fields – is the use of ontologies, or at least, OWL representations of some "middleware knowledge", to help bridge the gap

between the structure and meaning of a complex and lengthy text and the ideal benchmark it has to be evaluated against.

As a first attempt at identifying that "middleware knowledge", Rhetorical Structure Theory (RST) can be to some extent helpful. The theory was originally formulated in 1988 by William Mann and Sandra Thompson of the University of Southern California's Information Sciences Institute (ISI) [24]. Daniel Marcu, also from ISI, demonstrated that practical discourse parsing and text summarization goals could be achieved using RST [25]. He also created with Lynn Carlson and other colleagues the RST Discourse Treebank [8], a corpus composed of 385 Wall Street Journal articles annotated according to RST principles. This is considered a reference corpus in the area, particularly because it includes a number of humanly-generated extracts and abstracts associated with the original documents, to help verify the coherence of attributions [see 43]. In 1997, Ana Cristina Bicharra Garcia and Clarisse Sieckenius de Souz used RST to complement an existing Design Rationale System called ADD+. Design Rationale Systems are explanations of why artefacts are designed the way they are, including all the relevant background knowledge and decisions amongst concurrent options that have led to that specific design output [33]. In 2009, Nancy Green adopted RST as the basis for the representation of biomedical text argumentation, or the correlations between discourse structure and meaning [13].

The principles of RST are twofold:

- Coherent texts consist of minimal units, which are linked to each other, recursively, through rhetorical relations (see below), and
- Coherent texts do not show gaps or non-sequiturs.

Rhetorical relations are connections between different parts of a text, which are postulated to be hierarchical, following different possible schemas, the common aspect of which is to provide a layered representation of how the discourse is structured. RST based analysis is carried out by reading a text and constructing a tree of relations, such as the one depicted in Figure 2, which de-structures the title and summary of a Scientific American article [source: 38].

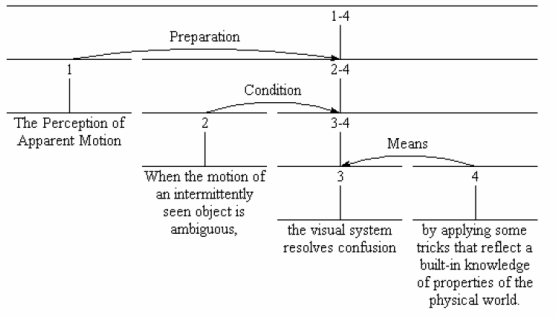

Figure 2. Destructuring of a small text using RST

The three rhetorical relations involved are "Preparation", "Condition" and "Means" – each defined in the classification provided by the RST creators. However, what is important to note is that many taxonomies have emerged in the state of the art, following the application domains of the theory itself (for a non-exhaustive overview see [39]). This implies that a promising research avenue for the purposes of my thesis is to develop an ad-hoc representation of scientific discourses that can be used for text summarization and analysis via some partly existing, partly newly developed rhetorical relations. Such approach might start from any collection of documents, including a corpus of a suitable number of DoAs (Descriptions of the Action) downloaded from the Internet websites of the awarded consortia. These could be manually annotated to form a similar dataset to the aforementioned Discourse Treebank. A tool for drawing RST schemas has been developed [by 28], which is accessible at [44]. Then based on that corpus, a constructivist approach might be adopted as in [23] to add a relational discourse structure annotation to the texts of new, yet to be evaluated, R&D or innovation proposals.

C. Bayesian Network Analysis (BNA)

The third building block of my PhD thesis is about modelling the logical structure of the "middleware knowledge" acting as a benchmark for the evaluation process as a Bayesian Network [32], that is a (sort of) Social Network Analysis (SNA) diagram defined by a set of random variables (nodes) and directed edges (arcs or arrows) connecting them in such a way to form a directed acyclic graph [15]. The key differences between SNA and BNA diagrams are threefold:

- Bayesian Networks are directed, while Social Networks are not. Indeed, SNA looks at the patterns and implications of connectedness among multiple entities (actors, nodes), including e.g. measuring the distance between every pair of them in terms of "steps" (edges, ties), and identifying which of the nodes in a network are comparatively more "central" (hubs) than others (peripheral or isolated) in terms of number or "density" of connections. On the other hand, BNA focuses on causality and dependency between variables, namely the graphical structure of its representations is similar to a cause-and-effect diagram and one of its key principles is the absence of cyclical routines or feedback loops. Some BNA software packages indicate the strength of the causality through the thickness of the arrows: the thicker the arrow, the stronger the dependency between those two variables.
- Bayesian Networks are probabilistic models, while Social Networks are deterministic. In BNA, every node is attributed a probability, a subjective and conditional probability (in terms of the Bayes theorem, from which the naming itself) to take on a certain value for each possible combination of values of its "parent" nodes. If

two specific nodes are not connected, this means they are statistically independent: a circumstance of little interest for SNA being so much focused on visualizing existing relations, very important for BNA as a way to simplify the underlying model representation.

- Bayesian Networks are predictive, while Social Networks are descriptive. For instance, BNA has been used to map the probabilistic relationships between diseases and symptoms, taking a symptom as an event that occurred and predicting the likelihood that any one of several possible known diseases was the contributing factor. A popular application of this method has been Hepar II, a sizeable Bayesian Network model for diagnosis of liver disorders, developed in collaboration with medical experts and parametrized using clinical information from an original database built in 1990 and maintained since then at the Gastroenterological Clinic of the Institute of Food and Feeding in Warsaw, Poland [29].

An interesting application of BNA is described by [11] as mixed-method approach (statistical as well as theory based) to analyze the behavioral impacts of public (financial) support to SMEs in the framework of EU cohesion policy. In short, and as Figure 3 exhibits, Bayesian Networks have been used to visualize the "theory of change" behind a certain public intervention. Then based on the results of a survey of beneficiary SMEs, each node has been associated with a distribution of frequency and the connections between nodes have been highlighted (the thickness of the edge increased), confirmed or neglected, up to the level of totally discarding the relevance of some variables, confined into

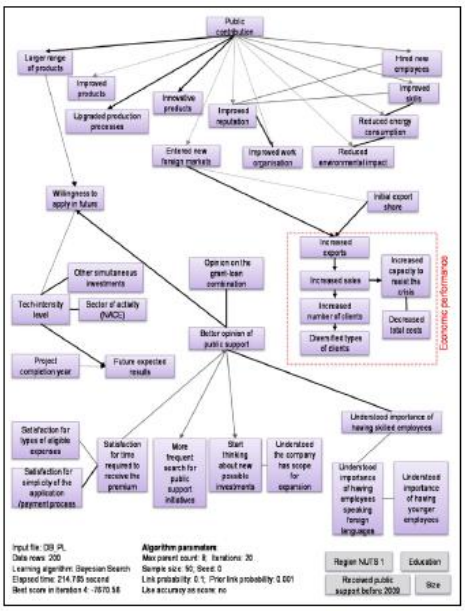

Figure 3. Policy visualisation using BNA

the bottom right part of the graph.

To create and update the Bayesian Networks, the authors used the proprietary modelling software GeNIe from BayesFusion, LLC formerly from the University of Pittsburgh and licensed for free to non-commercial users (see [45]). In my thesis, the same software package can be used to create and update the structure of the key aspects to be examined by the evaluators in charge of assessing an R&D and innovation project, and the features of which will be represented as RST schemas (see subsection B).

III. Discussion and Conclusion

The above overview of academic and industrial state-of-the-art in the three thematic areas of relevance for my PhD thesis – AEE, RST and BNA – may certainly be seen as "narrow minded", being linked, as was stated in the Introduction, to specific problems and product features that will be considered within the scope of my PhD thesis, rather than aiming to a full and complete description of the existing literature. This approach also brings with it the risk of overlooking some relevant contribution to the three research questions that characterize my thesis, a limitation that to the best possible extent, will be kept in mind and periodically re-examined during the course of the thesis preparation process.

References

[1] Ade-Ibijola, A. O., Wakama, I., & Amadi, J. C. (2012). An expert system for automated essay scoring (AES) in computing using shallow NLP techniques for inferencing. International Journal of Computer Applications, 51(10).

[2] Attali, Y. (2011). A Differential Word Use Measure for Content Analysis in Automated Essay Scoring, ETS Research Report Series 36.

[3] Bennett, R. E., & Ben-Simon, A. (2006). Toward theoretically meaningful automated essay scoring. National Institute for Testing & Evaluation.

[4] Bicharra Garcia, A. C., & De Souz, C. S. (1997). ADD+: Including rhetorical structures in active documents. Artificial Intelligence for Engineering, Design, Analysis and Manufacturing, 11(2), 109-124.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3 (4–5), 993–1022.

[6] Brent, E., Atkisson, C., & Green, N. (2010). Time-shifted collaboration: Creating teachable moments through automated grading. In: A. Juan, T. Daradournis, S. Caballe (Eds.), Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-learning Support, IGI Global, 55–73.

[7] Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® automated essay scoring system. In: M. D. Shermis, J. Burstein (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions. New York, NY: Routledge, 55–67.

[8] Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.), Current Directions in Discourse and Dialogue, Kluwer Academic Publishers, 85-112.

[9] Fazal, A., Dillon, T., & Chang, E. (2011). Noise reduction in essay datasets for automated essay grading. In: R. Meersman, T. Dillon, P. Herrero (Eds), On the Move to Meaningful Internet Systems: OTM 2011 Workshops. OTM 2011. Lecture Notes in Computer Science 7046, 484–493.

[10] Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay

assessor. In: M. D. Shermis, J. Burstein (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions, Routledge, New York, 68–88.

[11] Giffoni, F., Salini, S., & Sirtori, E. (2018). Evaluating business support measures: The Bayesian Network approach. Evaluation, 24 (2), 133-152.

[12] Gilboa, I., & Schmeidler, D. (2001). A theory of case-based decisions. Cambridge, UK: Cambridge University Press.

[13] Green, N. L. (2009). Representation of argumentation in text with Rhetorical Structure Theory. Argumentation, 24(2), 181–196.

[14] Hüllermeier, E. (2007). Case-based approximate reasoning (Vol. 44). Springer Science & Business Media.

[15] Jensen, F. V., & Nielsen, T.D. (2007). Bayesian Networks and Decision Graphs. 2nd ed. Information Science and Statistics. New York, NY: Springer.

[16] Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of Dimension Reduction Methods for Automated Essay Grading. Educational Technology & Society, 11 (3), 275–288.

[17] Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rose, C. P. (2015). Data mining and education. Wiley Interdisciplinary Reviews: Cognitive Science, 6(4), 333–353.

[18] Kolodner, J. L. (1993). Case-based Reasoning. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

[19] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25 (2-3), 259–284.

[20] Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. Computers and the Humanities 37, 389–405.

[21] Lebowitz, M. (1983). Memory-based parsing. Artificial Intelligence 21, 363-404.

[22] Lottridge, S. M., Schulz, E. M., Mitzel, H. C. (2013). Using automated scoring to monitor reader performance and detect reader drift in essay scoring. in: M. D. Shermis, J. Burstein (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions. New York, NY: Routledge, 233–250.

[23] Lüngen, H., Bärenfänger, M., Hilbert, M., Lobin, H., & Puskás, C. (2006). Text parsing of a complex genre.

[24] Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3), 243-281.

[25] Marcu, D. (2000). The theory and practice of discourse parsing and summarization. Cambridge, Mass.: MIT Press.

[26] Mayfield, E., & Rose, C. P. (2013). LightSide: Open source machine learning for text. In Handbook of Automated Essay Evaluation. London: Routledge, 146-157.

[27] Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. In: Proceedings of the 6th CAA Conference, Loughborough, UK: Loughborough University.

[28] O'Donnell, M. (2000). RSTTool 2.4 - A markup tool for Rhetorical Structure Theory. Proceedings of the International Natural Language Generation Conference (INLG'2000), 13-16 June, Mitzpe Ramon, Israel, 253 - 256.

[29] Oniśko, A., & Druzdzel, M. J. (2013). Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. Artificial intelligence in medicine, 57(3), 197-206.

[30] Page, E. B. (1966). The imminence of... grading essays by computer. The Phi Delta Kappan 47 (5), 238–243.

[31] Page, E. B. (1994). Computer grading of student prose using modern concepts and software. Journal of Experimental Education 62 (2), 127–142.

[32] Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

[33] Regli, W. C., Hu, X., Atwood, M., & Sun, W. (2000). A survey of Design Rationale Systems: Approaches, representation, capture and retrieval. Engineering with Computers, 16(3-4), 209–235.

[34] Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. The Journal of Technology, Learning and Assessment, 1 (2), 3–21.

[35] Schultz, M. T. (2013). The IntelliMetric automated essay scoring engine - A review and an application to Chinese essay scoring. In: M. D. Shermis, J. C. Burstein (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions, Routledge, New York, 89–98.

[36] Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays: Analysis. In: M. D. Shermis, J. Burstein (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions, Routledge, New York, 313–346.

[37] Smith, M. I., Schiano, A., & Lattanzio, E. (2014). Beyond the classroom. Knowledge Quest, 42 (3), 20–29.

[38] Taboada, M., & Mann, W. C. (2006a). Rhetorical Structure Theory: Looking back and moving ahead. Discourse Studies, 8(3), 423-459.

[39] Taboada, M., & Mann, W. C. (2006b). Applications of Rhetorical Structure Theory. Discourse Studies, 8(4), 567-588.

[40] Williams, R., & Dreher, H. (2004). Automatically grading essays with Markit®. Issues in Informing Science and Information Technology 1, 693–700.

[41] Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. Knowledge-Based Systems, 120, 118-132.

[42] https://en.wikipedia.org/wiki/Automated_essay_scoring

[43] https://www.isi.edu/~marcu/discourse/

[44] http://www.wagsoft.com/RSTTool/

[45] https://support.bayesfusion.com/docs/GeNIe.pdf