Towards Context-Aware Automated Writing Evaluation Systems

Pierre-André Patout
LCLD, Université Libre de Bruxelles
Belgium
ppatout@ulb.ac.be

Maxime Cordy
SnT, University of Luxembourg
Luxembourg
maxime.cordy@uni.lu

ABSTRACT

Writing is a crucial skill in our society, which is regularly exerted by students across all disciplines. Automated essay scoring and automatic writing evaluation systems can support professors in the evaluation of written texts and, conversely, help students improving their writing. However, most of those systems fail to consider the context of the writing, such as the targeted audience and the genre. In this paper, we depict our vision towards new-generation AES systems that could evaluate written products while considering their specific context. In education, such tools could support students not only in adapting their written product to their particular context, but also in identifying points for improvement and situational settings where their writing is less proficient.

CCS CONCEPTS

• **Social and professional topics → Student assessment**; • **Computing methodologies → Natural language processing**.

KEYWORDS

automated writing evaluation, context awareness, education

ACM Reference Format:
Pierre-André Patout and Maxime Cordy. 2019. Towards Context-Aware Automated Writing Evaluation Systems. In *Proceedings of the 1st ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence (EASEAI '19), August 26, 2019, Tallinn, Estonia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340435.3342722

1 INTRODUCTION

Writing is a crucial skill in our society. Particularly in education, the student writing performance is a good predictor for academic and professional achievement [1, 2]. In higher education, students need to take notes, have to make written homework and take exams, which are often written.

For the past decades, automated essay scoring (AES) systems and the automatic writing evaluation (AWE) systems have been developed to assess students' written products. These systems are designed as supports for professors during corrections. They allow them to save time and to lighten their cognitive endeavour. They are even capable of estimating text complexity, scoring text quality and, ultimately, grading students on the basis of their written text. Conversely, students can rely on such systems to benefit from early scoring and feedback on how to improve their writing, typically via the detection of errors and weaknesses in their essay. Most of the algorithms behind AES and AWE systems lean on shallow features (grammar and spelling). However, as mentioned by [14], the feedback based on mechanic traits does not help improving writing performance. This raises the need for advanced features that can better capture the quality of the text, its cohesion and coherence.

Yet, even though some systems are able to learn how to discriminate good and poor quality essays on the basis of such deeper linguistic features [11, 21], the perceived coherence and quality are always a mental construct of the reader/rater [6, 10, 15, 17, 20]. It is, therefore, necessary to consider the final written product as the outcome of the writing process during which the writer has considered the reader (e.g. his/her degree of knowledge about the topic) and the reading context (e.g. the genre of the text). As stated by [3, 9], the context is inseparable from the text. An essay is written for a particular audience, in a particular genre, for a particular goal.

To our knowledge, nowadays, no AES system takes into account the situational settings, the audience and the context in the written production analysis. On the basis of this observation, in this paper, we envision a new generation of AES systems that can learn how to evaluate the quality of an essay for a particular audience and within a given context. Such a tool will be able to process texts on several linguistic, within textual and contextual features along with human holistic and analytic text quality scores. Ultimately, our system will also provide feedback to the writer, not only on how to improve his/her essay but also by generalizing on his/her common errors to identify recurrent weaknesses. Although the scope of our envisioned tool goes beyond education, we see an immediate and essential benefit of it in this particular domain: to support students during their writing, allowing them to adapt their product to the particular context and to identify the situational settings where their writing is less proficient, as well as the specific points for improvement.

In what follows, We depict our vision behind the research that would be performed to build this system. We present the state of the art on writing evaluation (Section 2), our envisioned approach (Section 3) and we conclude with some perspectives (Section 4).

2 RELATED WORK

Most of the work on writing evaluation usually measure cohesive features: use of connectives, use of pronouns, noun overlap, argument overlap, and so on. This is the case of Coh-Metrix [1, 6, 16]. Coh-Metrix evaluates to what extent the essay has referential cohesion, syntactic complexity and lexical sophistication among other features such as the number of words, sentences, paragraphs, the

EASEAI '19, August 26, 2019, Tallinn, Estonia

ACM ISBN 978-1-4503-6852-0/19/08…$15.00
https://doi.org/10.1145/3340435.3342722

Flesch reading ease, latent semantic analysis. In order to know if these features could distinguish between high- and low-quality texts, they use holistic assessments and rubrics ratings by expert raters. The researchers found that the Coh-Metrix features were either negatively correlated with human assessment of the text quality or not at all related. They also showed that the human ratings of coherence are good predictors of the holistic scores. Crossley and McNamara [1] assume that some important factors for the text comprehension process are involved in the essay quality assessment. Indeed, multiple authors [10, 13, 15] demonstrated that the reader's prior knowledge impacts on the comprehension and hence on the text quality estimation. However, the unique rubric which deals with the reader, "Reader Orientation", actually is introduced as "Overall coherence and ease of understanding". In other words, it is focused on how the writer guides the reader through the text. Actually, this score indirectly considers the reader's point of view through a writer ability.

Amorim et al. [4] raise awareness about the readers' biases. In order to examine the extent to which these biases affect the ratings and thus affect the algorithms based on these ratings, they investigate a corpus of scored texts which are commented by raters. They showed that the biases actually affect the human ratings and the algorithms which are based on it. The researchers argue that putting the biased text aside from the training set make the AES systems more efficient. However, this is no more a random approach and it deletes a part of variability. In the aim to make assessments more objective, they only make it more neutral, sterilized. The human raters' individual differences exist. There is a need to take them into account and not to avoid or ignore them. By contrast, Crossley et al. [5] chose to consider the individual differences and insert them in their analyses. In order to increase the accuracy of the AES systems at the level of the matches with human scores of text quality, they use not only linguistic features but also students' attributes such as demographic information, reading comprehension skills, vocabulary knowledge, prior knowledge, reading apprehension, and so on. The results actually showed significant improvement in the human quality text assessments' predictions. These findings provide encouraging advances for taking into account non-linguistic features.

In an indirect way, Zesch et al. [21] showed the importance of considering these kinds of features. Indeed, aiming to find how to transfer an algorithm trained on a certain data set to another task, they showed that the tool trained only with task-independent features performed better but, at the same time, lost a great part of is explanatory power. In our opinion, these results corroborate several linguistic and psycholinguistic research works which demonstrated that the text genre and registers differ concerning their features and that it is needed to add them to the analyzed features. Besides, the study of Zesch et al. [21] differs from the previously mentioned studies in another way. While the other studies used systems which extracted some measures from texts which are used in correlation analyses and regression models (that is, AWEs), Zesch et al. [21] used an automatic essay grading system. This kind of systems usually uses handcrafting linguistic features which are related to a theoretical ground. These features help to extract measures which will be computed to match with human scores of the quality of essay, text coherence or multi-trait [12, 14]. These systems aim to automatically score essays and provide feedback to students. Other tools do not rely on feature engineering. The algorithm learns the features automatically from the data extracted from the essay and the relations with the human ratings [8, 19]. However, in many cases, because of the need of inter-rater agreement, the raters are always experts and are always over-trained on specific rubrics. It is a methodological choice that allows having a good rate of agreement. Notwithstanding this ease, it deletes the inherent variability we mentioned above. The fact that the raters are trained to score in the same way leads to a high agreement does not imply that the assessment criteria are correct, relevant and understood, much less that they would be identically used by other language experts (teachers, speech, language therapists . . .).

Our envisioned system will accept essays and context-based information as inputs. Since the reader/rater elaborates the coherence and constructs the text quality, the context-based information will take into account the supposed prior knowledge of the audience about a topic (high, medium, low), the level of expertise of the audience, the rhetoric constraints (text genre), and the communication goals. Regarding the essay, our system will extract measures related to linguistic features such as cohesive devices and features of coherence at various levels (word, sentence, proposition, paragraph, text). Moreover, it will learn to automatically extract these features and it will match them to the human scores from holistic and analytic ratings of coherence and quality of the text.

3 ENVISIONED APPROACH

Our envisioned system takes as inputs (i) any essay of any length and (ii) information about the final reading/rating context as inputs. This information is four-fold: (a) the extent of presupposed knowledge of the audience about the topic (e.g. high – medium – low); (b) the register (scientific, formal, sustained, educated, colloquial, informal); (c) text genre (narrative, persuasive, expository, descriptive); (d) the presupposed extent of vocabulary knowledge (e.g. high – medium – low) of the audience. Based on the processed text and context-based information, the system is able to give an overall quality score and feedback about potential improvements.

To reach this goal and a high degree of accuracy, there are four main steps to go through, following a typical machine learning pipeline. First of all, we have to create a corpus of texts acting as a training set for our system. The second step consists of selecting a sufficiently representative sample of the human population, in terms of individual features and their occurrence. The third step is the assessment of the texts by individuals of different backgrounds. Then, during the final step, we build a learning model that will be trained to rate texts and recognize what makes a good quality one.

3.1 Building a Corpus of Essays

The first step is arguably the most critical one. We have to control several linguistic and textual features in order to create an extremely varied and representative corpus of texts. To our knowledge, we are the first to aspire to work on a machine-learning-based AWE system at such a scale (i.e. a varied corpus evaluated by a varied audience). We consider five categories of features.

The vocabulary-oriented category determines the lexical richness, the lexical diversity and the content of the text. To this end, we rely on the type-token ratio (lexical diversity: the proportion of unique words), the hapax-token ratio (richness: the proportion of unique word forms), the percentage of content words – words that can be fully understood by themselves – and function words – words that are understood when occurring with other words – (content: proportion of words which convey information). For example, consider the following two sentences: (1) *John and Betty are strolling*; (2) *When I was a kid, I went to the sea and the mountains and I had fun every time*. In (1) there are five different words and five unique word forms, for a total of five words. Both diversity and richness are thus equal to 100%. In (2), there are nineteen words, out of which fifteen are different and twelve are unique. Thus, diversity is equal to 15/19 (79%), while richness is equal to 12/19 (63%). Regarding content and function words, in (1) there are three content words (*John, Betty*, and *strolling*) (60%) and two function words (*and, are*) (40%). In (2), *kid, went, sea, mountains, fun*, and *time* are the six content words (6/19, 32%), the other thirteen are function words (13/19, 68%).

The syntactic category includes: (i) the percentage of simple sentences (a single conjugated verb), of complex sentences (two or more conjugated verbs) and of non-verbal sentences (no conjugated verb) as a proxy for text complexity; (ii) the Flesch Reading Ease (proxy of text difficulty); (iii) the percentage of different verb tenses and modes (to evaluate the sequence of tenses), (iv) the percentage of syntactic structures (complexity: canonical order, inversion, extraction and dislocation for the theme's or rheme's emphasis . . .), and (v) the percentage of syntactic parallelisms (rhetorical figure that can be related to the ease of reading). Indeed, the more verbs a sentence contains, the more complex it is, regardless of the paratactic (without conjunction), subordinate or coordinate (with conjunction) nature of the juxtaposition. Knowing the proportion of complex, simple or non-verbal sentences can inform about the overall syntactic complexity. Besides, the Flesch Reading Ease is the most used index of text difficulty based on linguistic features (number of syllables, words and sentences). It is computed according to the following regression formula: 206.835 – (1.015 x Average Sentence Length) – (84.6 x Average Syllables per Word). The average sentence length refers to the average number of words per sentence. Finally, regarding the syntactic forms, the more different kinds of structure, the more complex the text. On the contrary, successively inserting new information with the same syntactic pattern can help to read, process, tie and integrate it.

The semantic category focuses on the presence and the progression of the main topics of the text. These features allow determining if the text forms a whole content unity, without digression, and which follows the need of adding new information while recalling the already-read or known one. For this, we rely on the proportion of propositions contained in the text. Since a proposition can be understood as a change (of state, of place, etc.), it brings up new information. Therefore, it is indirectly related to and underlies the topic progression. We can expect that more propositions correlate with more themes and/or information about the theme. Additionally, we also rely on Latent Semantic Analysis (LSA), along with a thematic progression recognition algorithm (inspired by [18]) and an isotopy evaluation based on noun overlap, argument overlap, and seme overlap. Semes are the fundamental units of semantic. They refer to the semantic traits associated with a word. For example, the word "Human" can be described through three semes: "animal", "sentient" and "rational". Moreover, to consider the metaphoric use of language, the overlap computation will also be based on the imaged meaning of nouns. The LSA and the different kinds of overlap are computed on adjacent pairs of sentences, as well as through all the sentences of a paragraph and between paragraphs. This allows for examining the semantic similarity and the topic progression at a micro-level (adjacent sentences), a macro-level (between all paragraph's sentences) and an overall level (between paragraphs).

Then, the referential category assesses the givenness of the information (the information already read by or known to the reader) and the way to remind information by calculating the percentage of use of the same item, synonyms, pronouns -– of different kinds -– according to one referent. Like in the semantic category, the computation is made at the micro, macro and overall levels. It consists of calculating the proportion of the different kinds of anaphora which are used to refer to one referent. There are different "degrees" of anaphora. The more explicit form is the use of the same word or noun phrase. It makes the text easier for poorly-skilled readers. A bit less explicit, other kinds use synonyms or periphrasis. Then, there is the use of pronouns. Pronouns are more synthetic forms and can lead to ambiguities, but also are more helpful to the more skilled reader. Finally, the "zero degree" of anaphora is the ellipsis, but we do not consider it.

Finally, the relational category deals with the connectives and the relations they make between two text segments. Evaluating the number of connectives of each kind can make it possible to estimate if the text is (well-)organized at an overall level (whole text), as well as at macro and micro levels. Most of these measures are computed at different scopes: word, proposition, sentence, paragraph and overall text. Usually, researchers base their computation of connectives on the types of coherence relations they are supposed to convey. Notably, these relations can be evaluated as positive or negative (e.g., [16]). However, it is the human mind that defines this valence. Moreover, there is no univocal relationship between connectives and coherence relations. Therefore, we will use a database of connectives which rely on the logical articulations of the text structure [7]. This taxonomy accepts that connectives can have several uses. Their typology counts fifty categories distributed in five main types: (1) space; (2) time (chronology, timespan); (3) facts (their prerequisites, details on prerequisites, modalities of the doing, purposes, the relations between facts); (4) discourse; (5) the writer/speaker's intervention.

3.2 Sampling the Audience Populations

The second and third steps can take place at the same time. The second step focuses on the individual data collection while the third one is on the human text evaluation. For both, it is conceivable to use tools and platforms which allow an extensive collection of various types of data (e.g., Psytoolbox and Amazon's Mechanical Turk). Relatively to the second step, the participants are asked to (anonymously) give personal information about, e.g., their age, mother tongue, knowledge and expertise of second languages, level of education, history of language, visual, auditory and neurologic impairments (corrected or not, if applicable), occupation, fields of expertise, degree of knowledge about the topics which texts deal with (high, medium, low), feeling about reading (like vs don't like; average of time spent reading books, magazines, newspapers, scientific papers, emails, social networks, etc.) and the feeling about writing (like vs don't like; average of time spent writing essays, emails, on social networks ...). Furthermore, their lexical and spelling, as well as their text comprehension performances and their vocabulary knowledge, are assessed through a few psycholinguistic tasks (lexical decision and spelling decision tasks and a reading comprehension task).

3.3 Evaluation of the Essays by the Audience

The third step holds in two stages with different samples of participants. During the first stage, the participants first provide personal information (as in the second step). Then, texts are randomly displayed one at a time. Participants are asked to read each text and give an overall quality text score (compulsory) and a comment about text quality (optional). The second stage follows the same design with other participants except that they do not give an overall quality text score. Instead, they assess texts on fifteen criteria (analytic scores) about text coherence, text cohesion and text quality, each on a five-point scale (compulsory). They can add some comment for each criterion and for the whole text (optional).

This data collection step is particularly challenging to set because of the number of texts to read and the required number of participants. Indeed, on the one hand, it is recommended to provide more than one text of each condition. That is, for a set of controlled features, we need to create several different texts. On the other hand, considering the number of texts to read, there is a risk that some participants quit before reading/rating all the texts. It can be attenuated by the randomized displaying, but we still have to collect a great number of answers for getting clear and accurate further measures as well as a large effect size.

3.4 Building the Model

For the last step, we rely on typical machine learning techniques. The algorithm will learn (a) how to match text and context-based features with human holistic scores, (b) how to match human holistic scores with human analytic scores, (c) how to match text and context-based features with human analytic scores, and (d) how to match all these features and scores.

According to this, by providing him with a text and guidelines about the final audience (readers/raters), our system will be able to score the overall quality of the text. Moreover, we expect that it will also provide feedback to improve the essay. This can be achieved, e.g., by determine which text features have to be modified/improved and to inform on which general criteria have to be checked (on the basis of the fifteen mentioned analytic criteria), or by explaining how the essay differs from the most of the processed written essays which received a high quality score in the targeted context.

4 CONCLUSION

The major novelty of our envisioned system lies in its ability to take into account the audience, or rather the inherent variety of the audience. Until now, essay assessment by AES or AWE are focused on the writer and how he/she has to improve his/her essay to reach the quality requirements of over-trained expert raters. Our system will allow the writer to improve his/her writing skills not only in a general way but also according to several contexts, which are determined by a specific register, a specific audience expecting a specific text genre, etc.

There are three main difficulties to which we will be confronted. First, there is the corpus creation step because the essays have to correspond to several sets of assembled criteria. Then, the human evaluation step implies having a lot of complete surveys, as mentioned above. Finally, we have to elaborate a machine learning algorithm which has to handle a great amount of data of several kinds (quantitative, qualitative, nominal and ordinal variables). Nevertheless, our efforts will result in an adaptable and helpful system which, we believe, could pave the way for new research.

REFERENCES

[1] Scott A Crossley, Danielle McNamara, and Com . 2010. Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.*

[2] Scott A Crossley and Danielle McNamara. 2011. Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence. *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

[3] Jean-Michel Adam. 2011. *La linguistique textuelle.* A. Colin Eds.

[4] Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated Essay Scoring in the Presence of Biased Ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* Association for Computational Linguistics, New Orleans, Louisiana.

[5] Scott Crossley, Laura K. Allen, Erica L. Snow, and Danielle S. McNamara. 2015. Pssst... Textual Features... There is More to Automatic Essay Scoring Than Just You!. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15).* 203–207.

[6] Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48, 4 (2016), 1227–1237.

[7] A.E. Dalcq. 1999. *Mettre de l'ordre dans ses idées: classification des articulations logiques pour structurer son texte.* Duculot.

[8] Fei Dong and Yue Zhang. 2016. Automatic Features for Essay Scoring - An Empirical Study. In *EMNLP.*

[9] Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English.* Longman, London.

[10] Judith Kamalski, Ted Sanders, and L.R. Lentz. 2008. Coherence Marking, Prior Knowledge, and Comprehension of Informative and Persuasive Texts: Sorting Things Out. *Discourse Processes* 45 (07 2008).

[11] Mirella Lapata and Regina Barzilay. 2005. Automatic Evaluation of Text Coherence: Models and Representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05).* 1085–1090.

[12] Yong-Won Lee, Claudia Gentile, and R Kantor. 2010. Toward Automated Multitrait Scoring of Essays: Investigating Links among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics - APPL LINGUIST* 31 (07 2010), 391–417.

[13] Danielle McNamara. 2001. Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian journal of experimental psychology = Revue canadienne de psychologie expérimentale* 55 (03 2001), 51–62.

[14] Danielle McNamara, Scott A. Crossley, Rod Roscoe, Laura K. Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing* 23 (1 1 2015), 35–59.

[15] Danielle McNamara and Walter Kintsch. 1996. Learning From Texts: Effects of Prior Knowledge and Text Coherence. *Discourse Processes* 22 (10 1996), 247–288.

[16] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix.* Cambridge University Press, New York, NY, USA.

[17] Fayol Michel. 2000. Des idées au texte, psychologie cognitive de la production verbale, orale et écrite. *Pratiques* 105 (01 2000), 247–250.

[18] Marie-Francine Moens. 2008. Using patterns of thematic progression for building a table of contents of a text. *Natural Language Engineering* 14 (04 2008), 145–172.

[19] Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of EMNLP.* 1882–1891.

[20] Elena Tribushinina, Elena Dubinkina, and Ted Sanders. 2015. Can connective use differentiate between children with and without specific language impairment? *First Language* 35, 1 (2015), 3–26.

[21] Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. [n. d.]. Task-Independent Features for Automated Essay Grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications.* 224–232.