An Automatic Long Answer Evaluation system based on n-gram and Euclidian Distance

KANCHAN NAITHANI ,
Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand, India 248002,

Abstract—
Assessing the value of one's answers constitutes one of the most essential aspects of learning. In modern day and age of digital technology, there have been produced a great deal of systems that are able to automatically handle evaluation. Long answers and brief answers are the two primary varieties of this type of response. Some of the existing scoring systems that are used on lengthy answers have demonstrated outcomes that are around average when it comes to putting a score on the student's answer. The information retrieval approach is employed in such systems to determine the degree of similarity between the answers provided by the students and the answers provided by the references; nevertheless, such scoring methods do not yet produce the best possible results. Each answer in a short answer just contains a select few keywords. The evaluation of such brief responses, which only contain a small number of keywords, calls for a unique approach, particularly when it comes to the weighing procedure. Cosine, dice, Euclidian distance, and overlap are the four metrics that have been analysed in this particular piece of research. To begin, the texts need to be transformed into some form of numerical expression. In order to accomplish this goal, the text was first broken up into n-grams at the word level, and then a bag containing all of these n-grams was constructed. After the frequencies of the n-grams have been computed, the frequency matrix of the dataset can be constructed. In order to produce a bag of n-grams, numerous filters, such as stemming algorithms, stop word removers, number and comma removers, etc., are utilised. The entire experimental inquiry was carried out with a dataset consisting of plagiarised long replies from a corpus.

Keywords— Automated Essay Evaluation, Automatic Scoring, Automatic Grading, String Similarity, Content-Based Similarity, Natural Language Processing
DOI Number: 10.48047/nq.2019.17.03.2012 **NeuroQuantology 2019; 17(03):128-135**

I. INTRODUCTION

In the current learning process, the bulk of tests for assessing the success of learning have been accomplished using basic composition questions. They include questions such as "true" or "false," "multiple choice," and "straightforward word" questions. In any case, in spite of the fact that it is helpful, the examination that utilises straightforward compose inquiries has a notoriety for being unseemly when it comes to evaluating the achievement of learning about just the procedure information, information useful, genuine reasoning capacity, and so on. It is generally accepted that illuminating evaluation techniques are essential testing devices for assessing academic accomplishment, the absorption of ideas, and the capacity to review, despite the fact that they are expensive and time-consuming to evaluate physically. Regardless of the fact that they make use of grading rubrics, the process of manually scoring the graphic responses is not only time-consuming and laborious, but it also presents challenges in terms of maintaining a consistent scoring system. The costs of grading will be significantly reduced if it can be proved that automated grading can match or exceed the consistent quality of human graders.

There are a variety of inquiry compositions and grading methods that can be utilised to foster the evaluation of learning outcomes through the use of

tests and examinations. Questions that require standard linguistic responses, such as brief answers or essays, are examples of the specific inquiry composes that may be defined as anything from straightforward multiple-choice questions to more in-depth types of queries. The contrast between multiple-choice and short-answer questions, for instance, is anything but difficult to comprehend. In any case, the contrast between other types of question formats, such as short-answer questions and essays, can wind up being muddled. Along these same lines, we say that a question qualifies as a short response question if it may be seen as meeting no less than five specific requirements. Instead of requiring the answer to be perceived from within the inquiry itself, the first step of the inquiry should involve the requirement of a reaction that analyses previous learning. Second, the inquiry needs to have a requirement that the response be made in regular speech. Finally, the length of the response ought to be somewhere in the range of one phrase to one passage. Fourth, the evaluation of the responses ought to place more of an emphasis on the content than on the author's writing style. Finally, using a target question strategy is the best way to limit the amount of openness and transparency in open-finished versus closed-finished responses.

In terms of the approaches to grading, there are certain questions that are physically more difficult to examine than others. When innovation is associated with automatic grading, there is actually a large amount of diversity possible. Given that an understanding of the normal dialect is necessary, scoring responses to short answer questions given in the common dialect can be considered to be significantly more challenging. Study in rating common dialect responses with computational methods has a history that stretches back to the early work of Page. This research has been going on for quite some time. Around that time period, the automatic grading of distinctive responses in many dialects had developed into a significant subject of study. In addition, the methods have diversified in accordance with the nature of the question, such as providing brief responses as opposed to writing essays. Because of this, we have made the conscious decision to devote the entirety of this post to discussing the process of analysing brief answers that have been automatically generated.

Automated evaluation as a rule, automated grading of typical dialect reactions, and ASAG in particular all come with a number of benefits that can be utilised to one's advantage. They are organised around the concepts of giving evaluations and input, as well as adequacy. When it comes to summative evaluation, the requirements of large class numbers and evaluation practises call for competent and financially wise arrangements. Moreover, people make mistakes when they are grading, and consistency is necessary when inter rater knowledge is wrong, which may arise from weariness, predisposition, other asking influences. Another advantage is that the capability of automatic grading in and of itself may progress the formalisation of evaluation criteria when something else is not accomplished. This is an advantage even when it is not performed. In the past, test takers were required to wait around for the human marking to finish grading their work, but now, with automatic grading systems, test takers don't have to wait around at all. This is something that should be taken into consideration. Regarding performance evaluation, automatic grading is gaining a lot of interest for more broad applications such as e-learning and intelligent coaching frameworks. Finally, in terms of practicability, automatic grading is getting to be extremely competitive with manual grading for both the automatic grading of short answer questions and automatic grading of essays.

There are still questions that need to be answered regarding the development of intrigue. The ongoing investigation concurrently focuses on both the nature of the scores and the level of confidence. Undoubtedly, several of the places of interest that were discussed before do not come free of the problems that they sometimes bring. For instance, the manual labour required to create an automated arrangement frequently requires a considerable amount of development time; the normality additional benefit can be a responsibility for poorest communities of a model when the poor parts make consistent wrong decisions; and caution has to be exercised to ensure that examples in framework conduct are still not gamed during evaluation by using language that is not natural.

In this research, a strategy is proposed for finding resemblance between texts by combining not only a quantitative estimation but also text grouping and visualisation. The goal of this approach is to speed up the process of finding similarities between texts. The

discovery of textual similarities is accomplished by first slicing the input text into word-level n-grams and then analyzing the results with a self-organizing map (SOM) and four numerical measurements. The text analysis that uses a bag of words is ineffective due to it being difficult to determine how similar two texts are by studying the frequency of single terms. This makes it impossible to determine how similar the bag of words analysis is.

II. An Overview of Automatic Assessment

There has been a significant amount of writing done on the Automated assessment framework, and more recently, notably in the last ten years, there have been several productions. We believe that it is necessary to completely identify the type of question that we are managing in order to proceed with our work towards a particular end goal. In light of this, the purpose of this section is to provide evidence that short response questions may be distinguished from other types of questions through the use of automated review.

The Educational Testing Service (ETS) is among the most important organisations in the field of computer-based testing and computerised evaluation. Their website features a classification system that outlines their investigations into automated scoring and the management of common dialects, such as composing material (such as short responses), written work quality (such as essays), science, and discourse. Those of [5] and [6] are incorporated into more advanced typologies, which provide a great deal more detail. On the other hand, [7] provides a progression that gives a gathering to dynamic and detached questions, as well as a sub-gathering of dynamic questions that require replies in the form of numbers or substance. In order to summarise these recently completed bodies of work, Figure 1 organises the characteristics into three different "swim paths": "profundity of learning," "question class," and "question composition." Although the figure is not intended to be exhaustive, its purpose is to merely demonstrate adequate and typical examples in order to differentiate ASAG queries from those of other organisations. Now, we are doing an audit of the three different swim pathways that highlight the different aspects of ASAG that are shown in Figure 1.

A. Depth of Learning

The primary degree of connection relates to the breadth of knowledge that may be gained from "acknowledgment" and "review" inquiries, which is a phrasing that is supported by the writing [6]. Yet, another distinction is whether the questions are passive or active, as was mentioned earlier [7]. In most cases, the replies need just organise or recognise a few basic pieces of information in order to complete the acknowledgment addresses. On the other hand, review questions have the benefit of compelling responders to come up with one-of-a-kind responses that are articulated in their very own distinctive way. When it comes to the mode of instruction, Bloom's scientific classification of learning objectives [8] places a greater emphasis on the significance of review strategies. In this context, acknowledgment questions can be interpreted as having to do with low-level verifiable learning. In the most fundamental sense, review questions are not as susceptible to test taking systems and speculation [9] as acknowledgment questions are.

Automatic assessment is a problem that has been worked out for the comprehension problems that include acknowledgment, as the answer is always among a set of possibilities. This is highlighted in the "Acknowledgment" section of Fig. 1, which can be seen here. Because of this reason as well as the others mentioned above, the focus of the energy that goes into automatic grading is on review queries. This category includes the types of questions known as short answer questions.

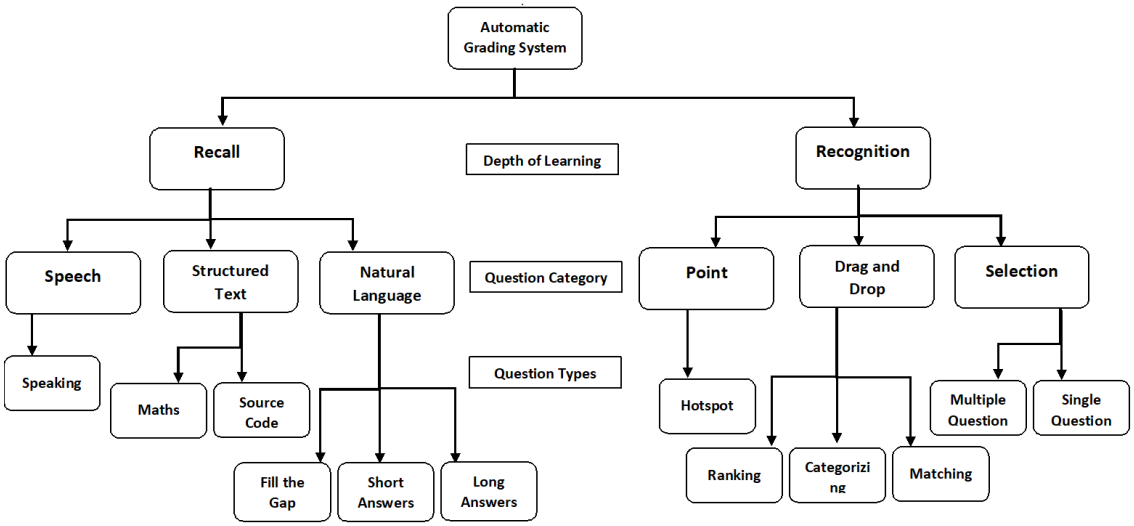

Figure 1. A Hierarchical Representation Of Common Question Types Where Automatic Grading Methods Can Be Applied

B. Question Category

We will just consider the bottom half of the second level of association, which is the review, as being relevant to this subject. The second level of association contains several main groups for different kinds of inquiries. The first of these is the appropriate rating for "brief answers," which is the "common dialect" rating. This provides an explanation for the absence of questions concerning mathematical documentation in our writing review: mathematical documentation can be regarded as ordered content, rather than as common dialect. The specialised analysis of programming language as organised content has attracted consideration in areas such as copyright violation location [10] and commencement [11]. This serves as an additional example of organised content. Finally, this leaves us with discourse: some of the topics can be considered with standard dialect after translation [12], however the concepts of elocution and pronunciation are interesting as well. Due to the fact that our passion for review addresses only reaches as far as those that can be displayed in a content-based configuration, we ignore certain graphical inquiry types that come from Fig. 1.

C. Question Type

In the third degree of association, we provide a listing of several specific inquiry composes. We have to differentiate between fill-in-the-hole questions, essay questions, and short response questions when it comes to the regular dialect question types. When additional terminology is used, such as "free-content answer" [13] and "constructed reaction" [14], the distinction between these types can be difficult to pin down, particularly when comparing short replies to essays. Length, centre, and transparency are the three fundamental measurements that we use to recognise typical query composes in normal dialect. Table 1 condenses all of these measurements, which are being looked at right now.

TABLE I. PROPERTIES DIFFRENTIATING TYPES OF NATURAL LANGUAGE QUESTIONS

Question Type			
Property	***Fill-the-gap***	***Short Answer***	***Essay***
Length	One Word To A Few Words	One Phrase To One Paragraph	Two Paragraphs To Several Pages.
Focus	Words.	Content.	Style.
Openness	Fixed.	Closed.	Open.

The length of the answer is the primary crucial measurement that is used to separate the standard dialect question composes. In both the short answer format and the essay format, the responses need to be sufficiently long in order to ensure that a diverse range of unique wordings and replies can be conveyed. This does not hold true for fill-in-the-blank questions due to the fact that the arrangements already contain close to two words. In the case of brief answers, the range for what constitutes "long" should be anywhere from about one expression (a few words) to one section so as to maintain coherence with the existing language. According to the examples that we find, the length of short answers might range from "expressions to three

to four sentences" [15] or "a couple of words to around 100 words" [16]. This results in essays being classified as having at least two sections and up to a few pages each.

The grading procedure centres on the second major measurement as its primary point of emphasis. In this regard, ASAG frameworks have a tendency to focus more on the subject matter, whereas automatic essay grading (AEG) frameworks [17] have a tendency to focus more on the writing style. This view is supported by two ETS frameworks that are each referred to as c-rater and erater respectively. These frameworks are examples of ASAG and AEG frameworks. In particular, [18] state that the purpose of the c-rater is to "delineate responses onto the specialists' models with the end goal of deciding their accuracy or sufficiency," whereas the framework of the e-rater "depends on the a bland model of composing that is attached to any provoke that belongs to an evaluation."

To put it another way, the authors of [19] assert that AEG frameworks "centre around measurements that comprehensively associate with composing style, enlarged with overall proportions of vocabulary utilisation," whereas ASAG methodologies are "worried about checking for content to the exclusion of everything else." One further thing to look at, however, is the substance in comparison to pronunciation and familiarity [20]. It is possible to make an exception for frameworks that advertise their ability to grade both essays and short answers. When it comes to fill-in-the-blank inquiries, we simply mean to convey that certain words are being emphasised. The degree to which the investigation can be seen as being open and honest is the third and most important criterion. In particular, ASAG frameworks anticipate responses to inquiries that are either goal- or close-focused. On the other hand, AEG frameworks anticipate responses to questions with abstract or open-ended conclusions. To phrase it another way, the distinction lies between situations and evaluations as opposed to certainties and proclamations. In the case of fill-in-the-blank inquiries, we claim that the reactions are settled because there is essentially no room for communicating any form of interest.

III. PROPOSED SYSTEM

The engineering of the suggested framework is depicted in figure 1, which may be found here. Under the methodology that we have proposed, the customer will just enter one record for the plagiarism check. First, a pre-handling process is carried out on the document, during which unnecessary white space, unusual characters, and so on are removed. Subsequently, a stopword expulsion process is carried out, during which stopwords such as an, a, the, numbers in records, and other stopwords are eliminated. After that, a procedure called stemming preparation is carried out, in which the ing, ed, and so on letters of each catchphrase are removed. In the latter part of the information report, only watchwords pertaining to word references have been retained.

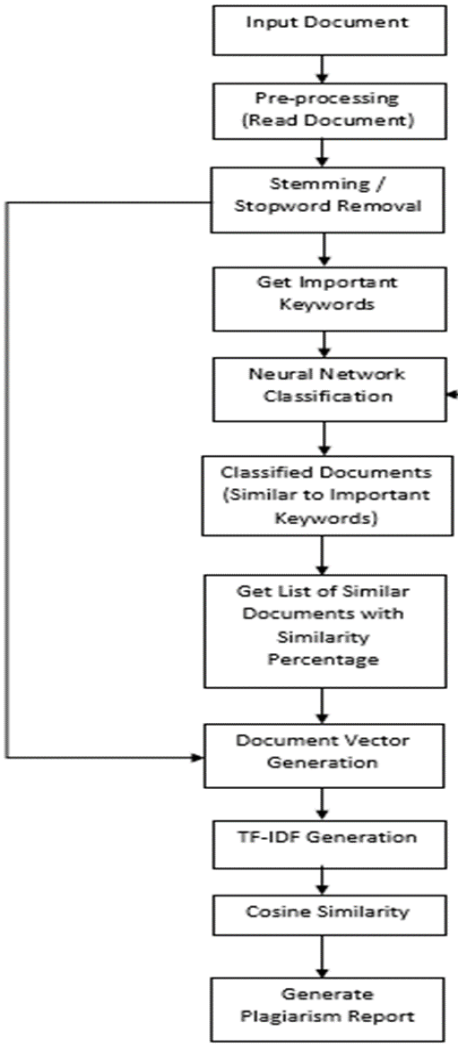

Figure 1. System Architecture

Following the acquisition of the catchphrase for the word reference from the library, vital watchwords were categorised (catchphrases having check more noteworthy than edge k). After that, the best k keyword

set is input into a neural system classifier, which divides newly-saved reports in the database into one of two categories: those that contain top k catchphrases (state class 1) and those that do not contain top k watchwords (state class 2). (state class 0). After that, we move on to managing things with archives that contain the top k catchphrases, which fall under class 1.

After this step is completed, the record vector of the input archive and the class 1 report are generated. When this step is complete, the TF-IDF of all records is generated, and finally, the cosine closeness between the information archive and class 1 reports is calculated. If it is discovered that there is a similarity between the information archive and another report, the input record will be marked as a plagiarism record, and the similarity rate will be calculated.

IV. IMPLEMENTATION

Algorithm: TF-IDF

In order to calculate the TF-IDF vector for a single record, such as a website page, we start by taking that record and running the TF-IDF score calculation for each intriguing word. We will do the following for each extraordinary word or phrase:

TF: Term Frequency

The word Frequency (TF) refers to a percentage that indicates how frequently a particular term appears in a particular report. We keep track of it by using the following
calculation:

$$\mathrm{TF}(\text{term}, \text{document}) = \frac{\text{Number of times the term appears in doc}}{\text{Total number of words in a doc}}$$

Take note that the TF score increases in proportion to the frequency with which the term appears in the record. If the customer's search query is "doughnuts" and the word "doughnuts" appears frequently on the website page, it is safe to assume that the page is about doughnuts and is something the customer will want to read. This is a positive sign.

IDF: Inverse Document Frequency

The Inverse Document Frequency (IDF) is a metric that determines how frequently a particular term appears throughout all of the documents that use this recipe.

$$\mathrm{IDF}(\text{term}) = \begin{cases} 0 & \text{if term doesn't appear in any doc} \\ \ln\left(\frac{\text{Total number of docs}}{\text{Number of docs containing the term}}\right) & \text{Otherwise} \end{cases}$$

Please be aware that in the event that you register the TF for a particular period, we will obtain an alternative TF for each report. On the other hand, when it comes to registering IDF, we do so by looking at each record individually rather than focusing on a single one.

You should also take note of the fact that the word's IDF score will be lower if it appears in a large number of documents. This is due to the fact that we do not require really fundamental terms like "the" or "of" to actually be identifying components of any record.

We choose to use the log because, from all appearances, it performs quite well in practise: based on our measurements, a relatively small number of words will be considered typical for the vast majority of records, and we will need to penalize those words more.

Algorithm- Neural Network

Network Contribution Phase:

A given passage is cut up into individual sentences. The clauses were then extracted from each sentence individually. After that, the phrases are extracted for each clause after the punctuation has been removed. After that, the POS and semantic role of each word in each phrase are extracted and labelled.

Network Learning Phase:

In order to avoid an excessively rapid increase in the energy associated with phrases that are used frequently, the starting weight of each connection is defined as

$$W_{ij} = \frac{w_{base}}{f_{ij}}$$

w_{base} is the value of the initial weight.

f_{ij} is the no. of times neuron i & j are connected for the input passage in the network construction phase.

Information Recall Phase:

Following learning, each neuron has connections with some other neurons that are stronger than its connections with the other neurons. When one of the word neurons is active, other word neurons, phrases, clauses, sentences, and concepts that are related to that word are also activated.

Similarity:

The experimental research was carried out in three stages so that the similarities between the investigated datasets could be discovered. At the first stage, the process of producing a bag of n-grams was broken down and examined. The original research indicates that the maximum number of words that can be included in an n-gram for this dataset is five; otherwise, some data is lost as a result of short texts. Moreover, in

order to produce the bag of n-grams. The emphasis is placed when the numbers of words in the n-grams are equal to three to five, therefore a total of fifteen different permutations were investigated. Figure 3 illustrates the size of the bag in terms of n-grams.

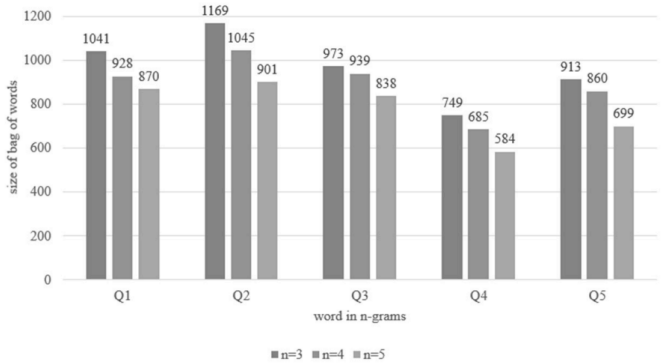

Figure 3. Comparison of bag of Words Found in Long answer Dataset

V. CONCLUSION

At the turn of the millennium, there has been rapid advancement in research concerning the automatic grading of regular dialect problems, and Automatic Grading System is not an exception to this trend. The descriptive answer question notate is one of the numerous types of questions that requires noteworthy understanding of material to review learning with the assumption of complimentary articulation. Because of this, the descriptive answer discussion compose is a tests to check assignment to review automatically. This paragraph makes reference to the stages of concept mapping, data extraction, approaches based on corpora, machine learning, and evaluation. We have noticed that the pattern in the eras that came before is moving more towards strategies that are based on administration. These strategies either analyse the responses in sections by using idea mapping techniques or in their entirety by using data extraction methods. We came to this conclusion as a result of our findings. After that, we came to the conclusion that the pattern was shifting more towards factual strategies, in which the highlights are produced with the assistance of corpus-based procedures or NLP techniques that are employed as a primary component of a machine learning framework. In conclusion, we discovered that the most recent trend is towards evaluation, where competitions and openly accessible informational sets are at last enabling significant comparisons between different methods.

REFERENCES

[1] M Ali Fauzi, Djoko Cahyo Utomo, Fakultas Ilmu Komputer, Budi Darma Setiawan (2017). Automatic Essay Scoring System Using N-Gram And Cosine Similarity For Gamification Based Elearning. In ICAIP 2017 Association for Computing Machinery. ACM ISBN 978-1-4503-5295-6/17/08.

[2] Eko Sakti Pramukantoro, M. Ali Fauzi (2016). Comparative Analysis Of String Similarity And Corpus-Based Similarity For Automatic Essay Scoring System On E-Learning Gamification. In ICACSIS 2016 Proceeding of IEEE.

[3] Kaja Zupanc and Zoran Bosnic (2015). Advances in the Field of Automated Essay Evaluation. In Informatica 2015.

[4] Kaja Zupanc and Zoran Bosnic (2018).Increasing accuracy of automated essay grading by grouping similar graders. In WIMS ’18, June 25–27, 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-5489-9/18/06.

[5] Bejar, I.I. (2011). A validity-based approach to quality control and assurance of automated scoring. Assessment in Education: Principles, Policy & Practice, 18(3), 319–341.

[6] Gy'orgy, A., & Vajda, I. (2007). Intelligent mathematics assessment in eMax. In Proceedings of the 8th Africon conference (pp. 1–6). Windhoek: IEEE.

[7] Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. Applied Measurement in Education, 15(4), 337–362.

[8] Krathwohl, D.R. (2002). A revision of bloom’s taxonomy: an overview. Theory into Practice, 41(4), 212–219.

[9] Conole, G., &Warburton, B. (2005). A review of computer-assisted assessment. Journal of the Association for Learning Technology, 13(1), 17–31.

[10] Burrows, S., Tahaghoghi, S.M.M., Zobel, J. (2007). Efficient plagiarism detection for large code repositories. Software: Practice and Experience, 37(2), 151–175.

[11] Burrows, S., Uitdenbogerd, A.L., Turpin, A. (2014). Comparing techniques for authorship attribution of source code. Software: Practice and Experience, 44(1), 1–32.

[12] Wang, X., Evanini, K., Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. In L. Vanderwende, H. Daum´e, K. Kirchhoff (Eds.), Proceedings of the

conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 81–819). Atlanta: Association for Computational Linguistics.

[13] Sargeant, J., McGee Wood, M., Anderson, S.M. (2004). A human-computer collaborative approach to the marking of free text answers. In Proceedings of the 8th computer assisted assessment conference (pp. 361–370). Loughborough: Loughborough University.

[14] Bennett, R.E. (2011). Automated scoring of constructed-response literacy and mathematics items. White paper, Educational Testing Service. Princeton.

[15] Siddiqi, R., Harrison, C.J., Siddiqi, R. (2010). Improving teaching and learning through automated shortanswer marking. IEEE Transactions on Learning Technologies, 3(3), 237–249.

[16] Sukkarieh, J.Z., & Stoyanchev, S. (2009). Automating model building in c-rater. In C. Callison-Burch & F.M. Zanzotto (Eds.), Proceedings of the 1st ACL/IJCNLP workshop on applied textual inference, TextInfer '09 (pp. 61–69). Suntec: Association for Computational Linguistics.

[17] Shermis, M.D., & Burstein, J. (2013). Handbook of automated essay evaluation: current applications and new directions, 1st edn. New York: Routledge New York City.

[18] Attali, Y., & Burstein, J. (2006). Automated essay Scoring with e-rater V.2. The Journal of Technology, Learning, and Assessment, 4(3), 1–31.

[19] Jordan, S., & Mitchell, T. (2009). e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. British Journal of Educational Technology, 40(2), 371–385.

[20] Williamson, D.M., Xi, X., Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13.