Automated language essay scoring systems: A Literature Review

Mohamed Abdellatif Hussein Corresp., 1 , **Hesham Ahmed Hassan** 2 , **Mohamed Nassef** 2

1 Information and Operations, National Center for Examination and Educational Evaluation, Cairo, Egypt

2 Faculty of Computers and Information, Computer Science Department, Cairo University, Cairo, Egypt

Corresponding Author: Mohamed Abdellatif Hussein
Email address: teeefa@nceee.edu.eg

Background. Writing composition is a significant factor for measuring test-takers' ability in any language exam. However, the assessment (scoring) of these writing compositions or essays is a very challenging process in terms of reliability and time. The need for objective and quick scores has raised the need for a computer system that can automatically grade essay questions targeting specific prompt. Automated Essay Scoring (AES) systems are used to overcome the challenges of scoring writing tasks by using Natural Language Processing and Machine Learning techniques. The purpose of this paper is to review the literature for the AES systems used for grading the essay questions. **Methodology.** We have reviewed the existing literature using Google Scholar, EBSCO and ERIC to search the terms "AES", "Automated Essay Scoring", "Automated Essay Grading", or "Automatic Essay", and two categories have been identified: handcrafted features and automatic featuring AES systems. The systems of the first category are closely bonded to the quality of the designed features. On the other hand, the systems of the other category are based on the automatic learning of the features and relations between an essay and its score without any handcrafted features. We reviewed the systems of the two categories in terms of system primary focus, technique(s) used in the system, training data (y/n), instructional application (feedback system), and the correlation between e-scores and human scores. The paper is composed of three main sections. Firstly, we present a structured literature review of the available Handcrafted Features AES systems. Secondly, we present a structured literature review of the available Automatic Featuring AES systems. Finally, we draw a set of discussions and conclusions. **Results.** AES models have been found to utilize a broad range of manually-tuned shallow and deep linguistic features. AES systems have many strengths in reducing labour-intensive marking activities, ensuring a consistent application of marking criteria, and facilitating equity in scoring. Although many techniques have been implemented to improve the AES systems, three primary challenges have been concluded: they lack the sense of the rater as a person, they can be tricked into

assigning a lower or higher score to an essay than it deserved or not, and they cannot assess the creativity of the ideas and propositions and evaluating their practicality. Many techniques have been used to address the first two challenges only.

Automated language essay scoring systems: A Literature Review

Mohamed Abdullatif Hussein[1], Hesham Ahmed Hassan[2], Mohamed Nassef[2]

[1] Information and Operations, National Center for Examination and Educational Evaluation, Cairo, Cairo, Egypt
[2] Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt

Corresponding Author:
Mohamed Hussein[1]
84E Hadayk Ahram, Haram, Giza, 12556, Egypt
Email address: teeefa@nceee.edu.eg

Abstract

Background. Writing composition is a significant factor for measuring test-takers' ability in any language exam. However, the assessment (scoring) of these writing compositions or essays is a very challenging process in terms of reliability and time. The need for objective and quick scores has raised the need for a computer system that can automatically grade essay questions targeting specific prompt. Automated Essay Scoring (AES) systems are used to overcome the challenges of scoring writing tasks by using Natural Language Processing and Machine Learning techniques. The purpose of this paper is to review the literature for the AES systems used for grading the essay questions. **Methodology.** We have reviewed the existing literature using Google Scholar, EBSCO and ERIC to search the terms "AES", "Automated Essay Scoring", "Automated Essay Grading", or "Automatic Essay", and two categories have been identified: handcrafted features and automatic featuring AES systems. The systems of the first category are closely bonded to the quality of the designed features. On the other hand, the systems of the other category are based on the automatic learning of the features and relations between an essay and its score without any handcrafted features. We reviewed the systems of the two categories in terms of system primary focus, technique(s) used in the system, training data (y/n), instructional application (feedback system), and the correlation between e-scores and human scores. The paper is composed of three main sections. Firstly, we present a structured literature review of the available Handcrafted Features AES systems. Secondly, we present a structured literature review of the available Automatic Featuring AES systems. Finally, we draw a set of discussions and conclusions. **Results.** AES models have been found to utilize a broad range of manually-tuned shallow and deep linguistic features. AES systems have many strengths in reducing labor-intensive marking activities, ensuring a consistent application of scoring criteria, and ensuring the objectivity of scoring. Although many techniques have been implemented to improve the AES systems, three primary challenges have been concluded: they lack the sense of the rater as a person, they can be deceived into giving a lower or higher score to an essay than it deserved or not, and they cannot assess the creativity of the ideas and propositions and evaluating their practicality. Many techniques have been used to address the first two challenges only.

Introduction

Test items (questions) are usually classified into two types: objective or selective-response (SR), and subjective or constructed-response (CR). The SR items, such as true/false, matching or multiple-choice, are much easier than the CR items in terms of marking objectively (Isaacs, 2013). The SR questions are commonly used for gathering information about knowledge, facts, higher-order thinking, and problem-solving skills. However, considerable skill is required to develop test items that measure analysis, evaluation, and other higher cognitive skills (Stecher et al., 1997).

The CR items, sometimes called open-ended, consist of two sub-types: short-response and extended-response answers (Nitko & Brookhart, 2007). The extended-response, such as essays, problem-based examinations, and scenarios, are like short-response items, except that they extend the demands made on test-takers to include more complex situations, more difficult reasoning, and higher levels of understanding that are based on real-life situations requiring test-takers to apply their knowledge and skills to new settings or situations (Isaacs, 2013).

In language tests, test-takers are usually required to write an essay about a given topic, and human-raters score these essays based on specific scoring rubrics or schemes. It occurs that the score of an essay scored by different human-raters vary substantially because human scoring is subjective (Peng, Ke, & Xu, 2012). As the process of human scoring takes much time, effort, and are not always as objective as required, there is a need for an automated essay scoring system that reduces cost, time and determines an accurate and reliable score.

The Automated Essay Scoring (AES) systems usually utilize Natural Language Processing and Machine Learning techniques to automatically rate essays written for a target prompt (Dikli, 2006). Many AES systems have been developed over the past decades. They focus on the automatic analysis of the quality of the writings and assignation of a score to a text. Typically, the AES models exploit a wide range of manually-tuned shallow and deep linguistic features (Farag, Yannakoudakis, & Briscoe, 2018). Recent advances in Deep Learning have shown that neural approaches applied to the AES systems accomplished state-of-the-art results (Page, 2003; Valenti, Neri, & Cucchiarelli, 2017) with the additional benefit of using features that are learned automatically from the data.

Survey methodology

The purpose of this paper is to review the literature for the AES systems that specifically score the extended-response items in language writing exams. Using Google Scholar, EBSCO and ERIC, we searched the terms “AES”, “Automated Essay Scoring”, “Automated Essay Grading”, or “Automatic Essay”. The AES systems that score objective or short-response items are excluded from the current research.

The most common models found for the AES systems are Natural Language Processing (NLP), Bayesian text classification, Latent Semantic Analysis (LSA), and Neural Networks. We have categorized the reviewed AES systems into two main categories: The first category is based on handcrafted discrete features bounded to specific domains. The second category is based on automatic feature extraction. For instance, the Artificial Neural Network (ANN)-based approaches are capable of automatically inducing dense syntactic and semantic features from a text.

The literature of the two categories have been structurally reviewed and evaluated in regards of some factors: system primary focus, technique(s) used in the system, training data (y/n), instructional application (feedback system), and the correlation between e-scores and human scores.

Handcrafted Features AES Systems

Project Essay Grader™ (PEG)

Ellis Page developed the PEG in 1966. The PEG is considered the earliest AES system that has been built in this field. It utilizes correlation coefficients to predict the intrinsic quality of the text. It uses the terms “trins” and “proxes” to assign a score. Where “trins” refers to the intrinsic variables like diction, fluency, punctuation, and grammar. In the other hand, “proxes” refers to the correlation of the intrinsic variables such as average length of words in a text and text length. (Dikli, 2006; Valenti et al., 2017).

The PEG uses a simple scoring methodology that consists of two stages. The first one is the training stage and the second one is the scoring stage. PEG has been trained on a sample of essays from 100 to 400 essays. In the scoring stage, proxes are identified for each essay, and are inserted into the prediction equation. To end, a score is determined by estimating coefficients (β weights) from the training stage (Dikli, 2006).

Some issues have been marked as a criticism for the PEG such as disregarding the semantic side of essays, focusing on the surface structures, and not working effectively with the case of receiving student responses directly (which might ignore writing errors). The PEG has a modified version released in 1990, which focuses on grammar checking with a correlation between human assessors and the system ($r=0.87$) (Dikli, 2006; Page, 1994; Refaat, Ewees, & Eisa, 2012).

Measurement Inc. acquired the rights of PEG in 2002 and continued to develop it. The modified PEG analyzes the training essays and calculates more than 500 features that reflect the intrinsic characteristics of writing, such as fluency, diction, grammar, and construction. Once the features have been calculated, the PEG uses them to build statistical and linguistic models for the accurate prediction of essay scores ("Home | Measurement Incorporated," n.d.).

Intelligent Essay Assessor™ (IEA)

The IEA was developed by Landauer et al. in 1997. The IEA uses a statistical combination of several measures to produce an overall score. It relies on using the Latent Semantic Analysis (LSA); a machine-learning model of human understanding of the text that depends on the training and calibration methods of the model and the ways it is used tutorially (Dikli, 2006; Foltz, Gilliam, & Kendall, 2003; Refaat et al., 2012).

The IEA can handle students' innovative answers by using a mix of scored essays and the domain content text in the training stage. It also spots plagiarism and provides feedback (Dikli, 2006; Landauer, 2004). It uses a procedure in assigning scores in a process that begins with comparing each essay to every other one in a set. LSA examines the extremely similar essays. Irrespective of the replacement of paraphrasing, synonym, or reorganization of sentences, the two essays will be alike LSA. Plagiarism is an essential feature to overcome academic dishonesty, which is difficult to be detected by human-raters, especially in the case of grading a large number of essays (Dikli, 2006; Landauer, 2004). (Figure 1) represents the IEA architecture (Landauer, 2004).

The IEA requires smaller numbers of pre-scored essays for training. On the contrary of other AES systems, IEA requires only 100 pre-scored training essays per each prompt vs. 300-500 on other systems (Dikli, 2006).

Landauer et al. in 2003 used IEA to score more than 800 students' answers in middle school. The results showed a 0.90 correlation value between IEA and the human-raters. He explained the high correlation value due to several reasons such as the human-raters could not compare each essay to each other for the 800 students while IEA can do so (Dikli, 2006; Landauer, 2004).

E-rater®

Educational Testing Services (ETS) developed E-rater in 1998 to estimate the quality of essays in various assessments. It relies on using a combination of statistical and NLP techniques to extract the linguistic features (such as grammar, usage, mechanics, development) from text to start processing, then compares scores with human graded essays (Attali & Burstein, 2014; Dikli, 2006; Ramineni & Williamson, 2018).

The E-rater system is upgraded annually. The current version uses 11 features divided into two areas: The first one is the writing quality (grammar, usage, mechanics, style, organization, development, word choice, average word length, proper prepositions, and collocation usage) and the second one is content or use of prompt-specific vocabulary (Ramineni & Williamson, 2018).

The E-rater scoring model consists of two stages. The first stage is the model of the training stage, and the other one is the model of the evaluation stage. Human scores are used for training and evaluating the E-rater scoring models. The quality of the E-rater models and its effective functioning in an operational environment depend on the nature and quality of the training and evaluation data (Williamson, Xi, & Breyer, 2012). The correlation between human assessors and the system ranged from 0.87 to 0.94 (Refaat et al., 2012).

Criterion℠

Criterion is a web-based scoring and feedback system based on ETS text analysis tools: E-rater® and Critique. As a text analysis tool, Critique integrates a collection of modules that detect faults in usage, grammar, and mechanics, and recognizes discourse and undesirable style elements in writing. It provides immediate holistic scores as well (Crozier & Kennedy, 1994; Dikli, 2006).

Criterion similarly gives personalized diagnostic feedback reports based on the types of assessment instructors give when they comment on students' writings. This component of the Criterion is called an advisory component. It is added to the score, but it does not control the score [18]. The types of feedback the advisory component may provide are like the following:

- The text is too brief (a student may write more).
- The essay text does not look like other essays on the topic (the essay is off-topic).
- The essay text is overly repetitive (student may use more synonyms).(Crozier & Kennedy, 1994)

IntelliMetric™

Vantage Learning developed the IntelliMetric systems in 1998. It is considered as the first AES system that relies on Artificial Intelligence (AI) to simulate manual scoring process carried out by human-raters under the traditions of cognitive processing, computational linguistics, and classification (Dikli, 2006; Refaat et al., 2012).

IntelliMetric relies on using a combination of Artificial Intelligence (AI), Natural Language Processing (NLP) techniques, and statistical techniques. It used CogniSearch and Quantum Reasoning technologies that were designed to enable IntelliMetric to understand the natural language to support essay scoring (Dikli, 2006).

IntelliMetric uses three steps to score essays as follow:

a) First, the training step that provides the system with known scores essays.

b) Second, the validation step examines the scoring model against a smaller set of known scores essays.
c) Finally, applying new essays with unknown scores. (Learning, 2000, 2003; Shermis & Barrera, 2002)

IntelliMetric identifies the text related characteristics as larger categories called Latent Semantic Dimensions (LSD). (Figure. 2) represents the IntelliMetric features model.

IntelliMetric scores essays in several languages (English, French, German, Arabic, Hebrew, Portuguese, Spanish, Dutch, Italian, and Japanese) (Elliot, 2003). According to Rudner, Garcia, and Welch (L. M. Rudner, Garcia, & Welch, 2006), the correlations average between IntelliMetric and human-raters was 0.83 (Refaat et al., 2012).

MY Access!

MY Access is a web-based writing assessment system based on the IntelliMetric AES system. The primary aim of this system is to provide immediate scoring and diagnostic feedback for the students' writings in order to motivate them to improve their writing proficiency on the topic (Dikli, 2006).

The MY Access system contains more than 200 prompts that assist in an immediate analysis of the essay. It can provide personalized Spanish and Chinese feedback on several genres of writing such as narrative, persuasive, and informative essays. Also, it provides multilevel feedback – developing, proficient, and advanced – as well (Dikli, 2006; Learning, 2003).

Bayesian Essay Test Scoring System™ (BETSY)

The BETSY classifies the text based on trained material and has been developed in 2002 by Lawrence Rudner at the College Park of the University of Maryland with funds from the U.S. Department of Education (Valenti et al., 2017). It has been designed to automate essay scoring, but can be applied to any text classification task (Taylor, 2005).

The BETSY needs to be trained on a huge number (1000 texts) of human classified essays to learn how to classify new essays. The goal of the system is to determine the most likely classification of an essay to a set of groups (Pass-Fail) and (Advanced - Proficient - Basic - Below Basic) (Dikli, 2006; Valenti et al., 2017). It learns how to classify a new document through the following steps:

The first-word training step is concerned with the training of words, evaluating database statistics, eliminating infrequent words, and determining stop words.

The second-word pairs training step is concerned with the training of word-pairs, evaluating database statistics, eliminating infrequent word-pairs, maybe scoring the training set, and trimming misclassified training sets.

Finally, BETSY can be applied to a set of experimental texts to identify the classification precision for several new texts or a single text. (Dikli, 2006)

The BETSY has achieved accuracy over 80%, when trained with 462 essays, and tested with 80 essays (L. M. Rudner & Liang, 2002).

Automatic Featuring AES Systems

Automatic Text Scoring Using Neural Networks

Alikaniotis, Yannakoudakis, and Rei introduced in 2016 a deep neural network model capable to learn features automatically to score essays. This model has introduced a novel method to identify the regions of the text that are more discriminative using: 1) a Score-Specific Word Embedding (SSWE) for represent words and 2) a two-layer Bidirectional Long-Short-Term Memory (LSTM) network to learn essay representations. (Alikaniotis, Yannakoudakis, & Rei, 2016; Taghipour & Ng, 2016).

Alikaniotis and his colleagues have extended the *C&W Embeddings* model into the *Augmented C&W* model to capture, not only the local linguistic environment of each word, but also how each word subsidizes to the overall score of an essay. In order to capture *SSWEs*, a further linear unit has been added in the output layer of the previous model that performs linear regression, predicting the essay score (Alikaniotis et al., 2016). (Figure 3) shows the architectures of two models, A) Original C&W model and B) Augmented C&W model. (Figure 4) shows the example of A) standard neural embeddings to B) *SSWE* word embeddings.

The SSWEs obtained by their model used to derive continuous representations for each essay. Each essay is identified as a sequence of tokens. The uni- and bi-directional LSTMs have efficiently used for embedding long sequences (Alikaniotis et al., 2016).

They used the Kaggle dataset (which used in ASAP competition). It consists of 12.976 150-to-550 word-essays, each was double makred (Cohen's = 0.86). The essays presented eight different prompts, each with distinct marking criteria and score range.

Results showed that the SSWE and the LSTM approach, without any prior knowledge of the language grammar or the text domain, was able to mark the essays in a very human-like way, beating other state-of-the-art systems. Furthermore, while tuning the models' hyperparameters on a separate validation set (Alikaniotis et al., 2016), they did not perform any further preprocessing of the text other than simple tokenization. Also, it outperforms the traditional SVM model by combining the SSWE and LSTM. On the contrary, LSTM alone did not give significant more accuracies compared to the SVM.

According to Alikaniotis, Yannakoudakis, and Rei (Alikaniotis et al., 2016), the combination of the SSWE with the two-layer bi-directional LSTM had the highest correlation value on the test set averaged 0.91 (Spearman) and 0.96 (Pearson).

A Neural Approach to Automated Essay Scoring

Taghipour and H. T. Ng developed in 2016 a Recurrent Neural Networks (RNNs) approach which automatically learn the relation between an essay and its grade. Since the system is based on the RNNs, so it can use the non-linear neural layers to identify the complex pattern in the data and learn it, and encode all the information required for essay evaluation and scoring (Taghipour & Ng, 2016).

The designed model architecture can be presented in five layers as follow:

a) Lookup Table Layer: The primary function is building d_{LT} dimensional space containing each word projection.

b) Convolution Layer: The primary function of this layer is to extract feature vectors from n-grams. It can possibly capture local contextual dependencies in writing and therefore enhance the performance of the system.
c) Recurrent Layer: The primary function of this layer is to process the input to generate a representation for the given essay.
d) Mean over Time: The main function of this layer is to aggregate the variable number of inputs into a fixed length vector.
e) Linear Layer with Sigmoid Activation: The primary function of this layer is to map the generated output vector from the mean-over-time layer to a scalar value. (Taghipour & Ng, 2016)

Taghipour and his colleagues employed in experiments the ASAP contest dataset organized by Kaggle. 60% of the data was a training set, 20% was a development set, and 20% was a testing set. They used Quadratic Weighted Kappa (QWK) as an evaluation metric. For evaluating the performance of the system, they compared it to an available opensource AES system called the 'Enhanced AI Scoring Engine' (EASE)[1]. To identify the best model, they performed several experiments like Convolutional vs. Recurrent Neural Network, basic RNN vs. Gated Recurrent Units (GRU) vs. LSTM, unidirectional vs. Bidirectional LSTM, and using with vs. without mean-over-time layer (Taghipour & Ng, 2016).

The results showed multiple observations according to (Taghipour & Ng, 2016), summarized as follow:

a) RNN failed to get accurate results as LSTM or GRU and the other models outperformed it. This was possibly due to the relatively long sequences of words in writing.
b) The neural network performance affected significantly with the absence of the mean over-time layer, as a result, it did not learn the task in an exceedingly proper manner.
c) The best model was the combination of ten instances of LSTM models with ten instances of CNN models. The new model outperformed by 5.6% the baseline EASE system and with averaged QWK value 0.76.

Automatic Features for Essay Scoring – An Empirical Study

Dong and Zhang provided in 2016 an empirical study to examine a neural network method to learn syntactic and semantic characteristics automatically for AES, without the need for external pre-processing. They built a hierarchical Convolutional Neural Network (CNN) structure with two levels in order to model sentences separately (Dasgupta, Naskar, Saha, & Dey, 2018; Dong & Zhang, 2016).

Dong and his colleague built a model with two parts, summarized as follow:

a) Word Representations: A word embedding is used but does not rely on POS-tagging or other pre-processing.
b) CNN Model: They took essay scoring as a regression task and employed a two-layer CNN model, in which one Convolutional layer is used to extract sentences representations, and the other is stacked on sentence vectors to learn essays representations.

[1] https://github.com/edx/ease

The dataset that they employed in experiments is that the ASAP contest dataset organized by Kaggle, the settings of data preparation followed the one that Phandi, Chai, and Ng used (Phandi, Chai, & Ng, 2015). For domain adaptation (cross-domain) experiments, they followed Phandi, Chai, and Ng (Phandi et al., 2015), by picking four pairs of essay prompts, namely, 1 → 2, 3→4, 5→6 and 7→8, where 1→2 denotes prompt one as source domain and prompt. They used quadratic weighted Kappa (QWK) as the evaluation metric.

In order to evaluate the performance of the system, they compared it to EASE system (an open source AES available for public) with its both models Bayesian Linear Ridge Regression (BLRR) and Support Vector Regression (SVR).

The Empirical results showed that the two-layer Convolutional Neural Network (CNN) outperformed other baselines (e.g., Bayesian Linear Ridge Regression) on both in-domain and domain adaptation experiments on the ASAP dataset So, the neural features learned by CNN were very effective in essay marking, handling more high-level and abstracting information compared to manual feature templates. In domain average, QWK value was 0.73 vs. 0.75 for human rater (Dong & Zhang, 2016).

Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring

In 2018, Dasgupta *et al.* proposed a Qualitatively enhanced Deep Convolution Recurrent Neural Network architecture to score essays automatically. The model consider both word- and sentence-level representations. Using a Hierarchical CNN connected with a Bidirectional LSTM model they were able to consider linguistic, psychological and cognitive feature embeddings within a text (Dasgupta et al., 2018).

The designed model architecture for the linguistically informed Convolution RNN can be presented in five layers as follow:

a) Generating Embeddings Layer: The primary function is constructing sentence vectors which previously trained. The sentence vectors extracted from every input essay are appended with the formed vector from the linguistic features determined for that sentence.
b) Convolution Layer: For a given sequence of vectors with K windows, this layer function is to apply linear transformation for all these K windows. This layer is fed by each of the generated word embeddings from the previous layer.
c) Long Short-Term Memory Layer: The main function of this layer is to examine the future and past sequence context by connecting Bidirectional LSTMs (Bi-LSTM) networks.
d) Activation layer: The main function of this layer is to obtain the intermediate hidden layers from the Bi-LSTM layer $h_1, h_2, \ldots, h_T$, and in order to calculate the weights of sentence contribution to the final essay's score (quality of essay), they used an attention pooling layer over the sentence representations.
e) The Sigmoid Activation Function Layer: The main function of this layer is to perform a linear transformation for the input vector that convert it to a scalar value (continuous). (Dasgupta et al., 2018)

(Figure 5) represents the proposed linguistically informed Convolution Recurrent Neural Network architecture.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27715v1 | CC BY 4.0 Open Access | rec: 9 May 2019, publ: 9 May 2019

Dasgupta and his colleagues employed in their experiments is that the ASAP[2] contest dataset organized by Kaggle; they have done 7 folds using cross validation technique to assess their models. Every fold is distributed as follow; training set which represent 80% of the data, development set represented by 10%, and the rest 10% as the test set. They used quadratic weighted Kappa (QWK) as the evaluation metric.

The results showed that, in terms of all these parameters, the Qualitatively Enhanced Deep Convolution LSTM (Qe-C-LSTM) system performed better than the existing, LSTM, Bi-LSTM and EASE models. It achieved a Pearson's and Spearman's correlation of 0.94 and 0.97 respectively as compared to that of 0.91 and 0.96 in (Alikaniotis et al., 2016). They also accomplished an RMSE score of 2.09. They computed a pairwise Cohen's k value of 0.97 as well (Dasgupta et al., 2018).

Summary and Discussion

Over the past four decades, there have been several studies that have examined the approaches of applying computer technologies on scoring essay questions. Recently, computer technologies, especially NLP and AI, have been able to assess the quality of writing using AES technology. Many tries have took place in developing AES systems in the past years (Dikli, 2006).

The AES systems do not assess the intrinsic qualities of an essay directly as human-raters do, but they utilize the correlation coefficients of the intrinsic qualities to predict the score to be assigned to an essay. The performance of these systems is evaluated based on the comparison of the scores assigned to a set of essays scored by expert humans.

The AES systems have many strengths mainly in reducing labor-intensive marking activities, overcoming time, cost, and improving the reliability of writing tasks. Besides, they ensure a consistent application of marking criteria, therefore facilitating equity in scoring. However, there is substantial manual effort involved in reaching these results on different domains, genres, prompts and so forth. Also, linguistic features intended to capture the aspects of writing to be assessed are hand-selected and tuned for specific domains. In order to perform well on different data, separate models with distinct feature sets are typically tuned (Burstein, 2003; Dikli, 2006; Hamp-Lyons, 2001; L. Rudner & Gagne, 2001; L. M. Rudner & Liang, 2002). Despite its weaknesses, the AES systems continue to attract the attention of public schools, universities, testing agencies, researchers and educators. (Dikli, 2006).

The AES systems described in this paper under the first category are based on handcrafted features and usually, rely on regression methods. It employs several methods to obtain the scores. While E-rater and IntelliMetric use the NLP techniques, the IEA system utilizes the LSA. Moreover, PEG utilizes proxy measures (proxes), and BETSY™ uses Bayesian procedures to evaluate the quality of a text.

While E-rater, IntelliMetric, and BETSY evaluate style and semantic content of essays, PEG is only evaluating style and ignoring the semantic aspect of essays. Furthermore, IEA is concerned with only semantic content. Unlike PEG, E-rater, IntelliMetric, and IEA need smaller numbers of

[2] https://www.kaggle.com/c/asap-aes/data

pre-scored essays for training in contrast with BETSY which needs a huge number of training pre-scored essays.
The systems in the first category have high correlations with human-raters. While PEG, E-rater, IEA, and BETSY evaluate only the English language essay responses, IntelliMetric evaluates essay responses in multiple languages.
On contrary of PEG, IEA, and BETSY, E-rater, and IntelliMetric have instructional or immediate feedback applications (i.e., Criterion and MY Access!). The instructional-based AES systems have worked hard to provide formative assessments by allowing students to save their writing drafts on the system. Thus, students can review their writings as of the formative feedback received from either the system or the teacher. The recent version of MY Access! (6.0) provides online portfolios and peer review.
The drawbacks of this category can be summarized as a) the feature engineering, which can be time-consuming, since features need to be carefully handcrafted and selected to fit the appropriate model and b) they are sparse and instantiated by discrete pattern-matching.
The AES systems described in this paper under the second category are usually based on neural networks. Neural Networking approaches, especially Deep Learning techniques, have been shown to be capable of inducing dense syntactic and semantic features automatically, and apply them to text analysis and classification problems including AES systems (Alikaniotis et al., 2016; Dong & Zhang, 2016; Taghipour & Ng, 2016), and give better results in regards to the statistical models used in the handcrafted features (Dong & Zhang, 2016).
Recent advances in Deep Learning have shown that neural approaches to AES achieve state-of-the-art results (Alikaniotis et al., 2016; Taghipour & Ng, 2016) with the additional advantage of utilizing features that are automatically learned from the data. In order to facilitate interpretability of neural models, a number of visualizations techniques have been proposed to identify textual (superficial) features that contribute to model performance [7].
While Alikaniotis and his colleagues (2016) employed a two-layer Bidirectional LSTM combined with the SSWE for essay scoring tasks, Taghipour and Ng (2016) adopted the LSTM model and combined it with the CNN. Dong and Zhang (2016) developed a two-layer CNN, and Dasgupta and his colleagues (2018) proposed a Qualitatively Enhanced Deep Convolution LSTM. Unlike Alikaniotis and his colleagues (2016), Taghipour and Ng (2016), Dong and Zhang (2016), Dasgupta and his colleagues (2018) were interested in word-level and sentence-level representations as well as linguistic, cognitive and psychological feature embeddings. All linguistic and qualitative features were figured off-line and then entered in the Deep Learning architecture.
Although the Deep Learning-based approaches have achieved better performance than the previous approaches, the performance may not be better using the complex linguistic and cognitive characteristics, which are very important in modeling such essays. See (Table 1) for the comparison of AES systems.
In general, there are three primary challenges to AES systems. Firstly, they are not able to assess essays as human-raters do because they do what they have been programmed to do (Page, 2003).

They eliminate the human element in writing assessments and lack the sense of the rater as a person (Hamp-Lyons, 2001). This shortcoming was somehow overcome by obtaining high correlations between the computer and human-raters (Page, 2003) although this is still a challenge.

The second challenge is whether the computer can be fooled by students or not (Dikli, 2006). It is likely to "trick" the system by, e.g., writing a longer essay to obtain higher score (Kukich, 2000). Studies, such as the GRE study in 2001, examined whether a computer could be deceived and assign a lower or higher score to an essay than it should deserve or not, and results revealed that it might reward a poor essay (Dikli, 2006). The developers of AES systems have been utilizing algorithms to detect students who try to cheat.

Although the automatic learning AES systems depend on one of the most recent technologies, which is Neural Networks, the handcrafted AES systems transcend automatic learning systems in one important feature. Handcrafted systems are highly tight to the scoring rubrics that have been designed as a criterion for assessing a specific essay and human-raters use these rubrics to score essays a well. The objectivity of human-raters is measured by their commitment to the scoring rubrics. On the contrary, automatic learning systems extract the scoring criteria using machine learning and neural networks, which may include some factors that are not part of the scoring rubric such as raters' subjectivity (i.e., mode, nature of a rater's character, etc.) Considering this point, handcrafted AES systems may be considered as more objective and fairer to students from the viewpoint of educational assessment.

The third challenge to AES systems is measuring the creativity of human writing. Accessing the creativity of the ideas and propositions and evaluating their practicality are still a confronting challenge to both categories of AES systems and still need further research.

References

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). *Automatic Text Scoring Using Neural Networks*. https://doi.org/10.18653/v1/P16-1068

Attali, Y., & Burstein, J. (2014). Automated Essay Scoring With E-Rater® V.2.0. *ETS Research Report Series*, *2004*(2), i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Burstein, J. (2003). The e-rater Scoring Engine: Automated Essay Scoring with Natural Language Processing, 107–115. Retrieved from http://books.google.com/books?id=JIR6ihd7ZA4C&printsec=frontcover#v=onepage&q&f=false

Crozier, W. W., & Kennedy, G. J. A. (1994). Marine exploitation of Atlantic salmon (Salmo salar L.) from the River Bush, Northern Ireland. In *Fisheries Research* (Vol. 19, pp. 141–155). https://doi.org/10.1016/0165-7836(94)90020-5

Dasgupta, T., Naskar, A., Saha, R., & Dey, L. (2018). Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring. *Aclweb.Org*, 93–102. Retrieved from http://aclweb.org/anthology/W18-3713

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal Of Technology,*

Learning, and Assessment, *5*(1), 1–36.
Dong, F., & Zhang, Y. (2016). Automatic Features for Essay Scoring – An Empirical Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* (pp. 1072–1077). https://doi.org/10.18653/v1/d16-1115
Elliot, S. (2003). IntelliMetric: From here to validity. *Automated Essay Scoring: A Cross-Disciplinary Perspective*, 71–86. Retrieved from https://ci.nii.ac.jp/naid/10025900425/
Farag, Y., Yannakoudakis, H., & Briscoe, T. (2018). Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input, 263–271. https://doi.org/10.18653/v1/N18-1024
Foltz, P. W., Gilliam, S., & Kendall, S. (2003). Supporting Content-Based Feedback in On-Line Writing Evaluation with LSA. *Interactive Learning Environments*, *8*(2), 111–127. https://doi.org/10.1076/1049-4820(200008)8:2;1-b;ft111
Hamp-Lyons, L. (2001). Fourth generation writing assessement. *On Second Language Writing*, *117*, 117–128.
Home | Measurement Incorporated. (n.d.). Retrieved February 5, 2019, from http://www.measurementinc.com/
Isaacs, T. (2013). *Key Concepts in Educational Assessment [electronic resource]*. Sage.
Kukich, K. (2000). Beyond automated essay scoring, the debate on automated essay grading. *IEEE Intelligent Systems*, *15*(5), 22–27.
Landauer, T. K. (2004). Automatic Essay Assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 295–308. https://doi.org/10.1080/0969594032000148154
Learning, V. (2000). A true score study of IntelliMetric accuracy for holistic and dimensional scoring of college entry-level writing program (RB-407). *Newtown, PA: Vantage Learning*.
Learning, V. (2003). A true score study of 11th grade student writing responses using IntelliMetric Version 9.0 (RB-786). *Newtown, PA: Vantage Learning*, 1.
Nitko, A. J., & Brookhart, S. M. (2007). *Educational Assessment of Students, 5th edition*. Pearson Merrill Prentice Hall.
Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, *62*(2), 127–142. https://doi.org/10.1080/00220973.1994.9943835
Page, E. B. (2003). Project essay grade: PEG. *Automated Essay Scoring: A Cross-Disciplinary Perspective*.
Peng, X., Ke, D., & Xu, B. (2012). Automated Essay Scoring Based on Finite State Transducer: towards ASR Transcription of Oral English Speech. *Jeju, Republic of Korea*, (July), 50–59.
Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression. *Wiley*, 431–439. https://doi.org/10.18653/v1/d15-1049
Ramineni, C., & Williamson, D. (2018). Understanding Mean Score Differences Between the e-rater ® Automated Scoring Engine and Humans for Demographically Based Groups in the GRE ® General Test. *ETS Research Report Series*.

https://doi.org/10.1109/ISIE.1997.648935

Refaat, M. M., Ewees, A. A., & Eisa, M. M. (2012). Automated Assessment of Students ' Arabic Free-Text Answers. *International Journal of Intelligent Computing And Information Science*, *12*(1), 213–222. Retrieved from https://www.researchgate.net/profile/Ahmed_Ewees/publication/236019860_AUTOMATED_ASSESSMENT_OF_STUDENTS%27_ARABIC_FREE-TEXT_ANSWERS/links/5a9afae8aca2721e3f3017d4/AUTOMATED-ASSESSMENT-OF-STUDENTS-ARABIC-FREE-TEXT-ANSWERS.pdf

Rudner, L., & Gagne, P. (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. ERIC Digest.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An Evaluation of the IntelliMetricSM Essay Scoring System. *The Journal of Technology, Learning and Assessment*, *4*(4), 3–20. Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1651/1493

Rudner, L. M., & Liang, T. (2002). Automated Essay Scoring Using Bayes' Theorem. *The Journal of Technology, Learning, and Assessment*, *1*(2), 1–21. Retrieved from http://napoleon.bc.edu/ojs/index.php/jtla/article/view/1668

Shermis, M. D., & Barrera, F. D. (2002). Exit assessments evaluating writing ability through automated essay scoring. *Annual Meeting of the American Educational Research Association, New Orleans, LA*, 1–30. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED464950&site=ehost-live

Stecher, B. M., Rahn, M., Ruby, A., Alt, M., Robyn, A., & Ward, B. (1997). Types of Assessment. *Using Alternative Assessments in Vocational Education*.

Taghipour, K., & Ng, H. T. (2016). A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). https://doi.org/10.18653/v1/d16-1193

Taylor, A. R. (2005). *A Future in the Process of Arrival : Using Computer Technologies for the Assessment of Student Learning*. Retrieved from http://site.ebrary.com/lib/uwo/docDetail.action?docID=10276954

Valenti, S., Neri, F., & Cucchiarelli, A. (2017). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education: Research*, *2*, 319–330. https://doi.org/10.28945/331

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. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Table 1(on next page)

The comparison of AES systems

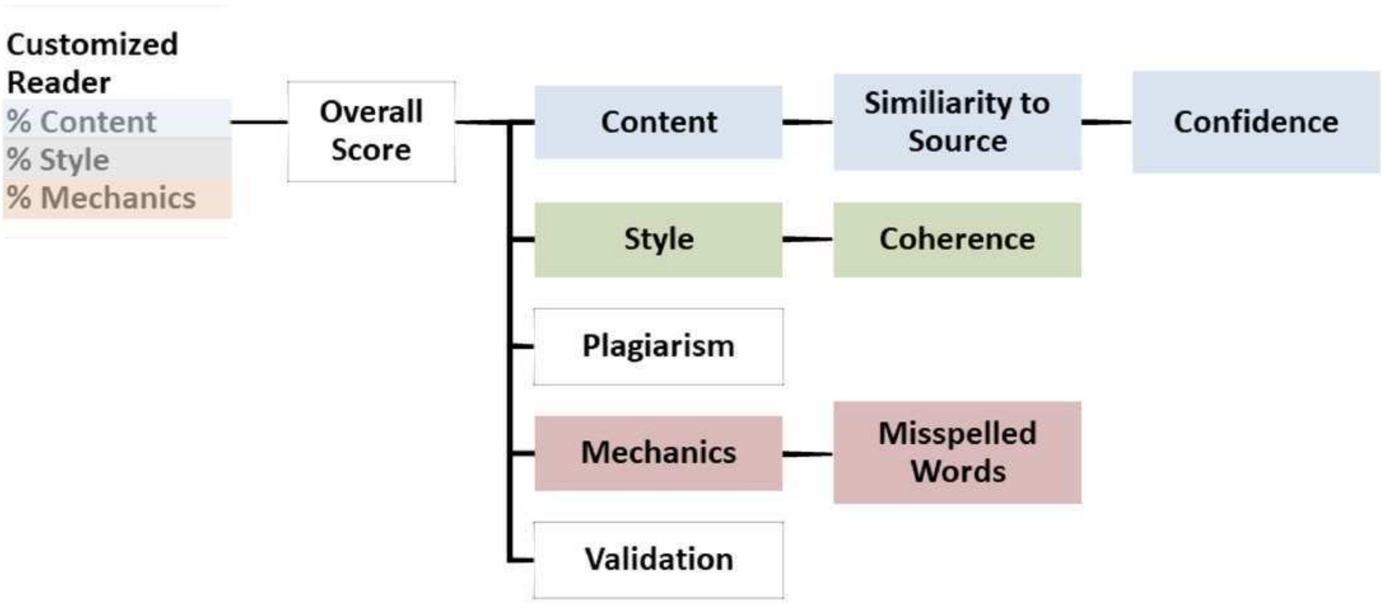
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27715v1 | CC BY 4.0 Open Access | rec: 9 May 2019, publ: 9 May 2019

AES/Parameter	Vendor	Release date	Primary focus	Technique(s) used	Training data	Feedback Application	Correlation with human scorers
PEG™	Ellis Page	1966	Style	Statistical	Yes (100 – 400)	No	0.87
IEA™	Landauer, Foltz, & Laham	1997	Content	LSA (KAT engine by PEARSON)	Yes (~100)	Yes	0.90
E-rater®	ETS development team	1998	Style & Content	NLP	Yes (~400)	Yes (Criterion)	~ 0.91
IntelliMetric™	Vantage Learning	1998	Style & Content	NLP	Yes (~300)	Yes (MY Access!)	~ 0.83
BETSY™	Rudner	1998	Style & Content	Bayesian text classification	Yes (1000)	No	~ 0.80
D. Alikaniotis, H. Yannakoudakis, and M. Rei (Alikaniotis, Yannakoudakis, & Rei, 2016)	Alikaniotis, Yannakoudakis, and Rei	2016	Style & Content	SSWE + Two-layer Bi-LSTM	Yes (~ 8000)	No	~0.91 (Spearman) ~0.96 (Pearson)
Taghipour and Ng (Taghipour & Ng, 2016)	Taghipour and Ng	2016	Style & Content	Adopted LSTM	Yes (~7786)	NO	QWK for LSTM ~0.761
Dong and Zhang (Dong & Zhang, 2016)	Dong and Zhang	2016	Syntactic and semantic features	Word embedding and a two-layer Convolution Neural Network	Yes (~1500 to ~1800)	NO	average kappa ~ 0.734 versus 0.754 for human
T. Dasgupta, A. Naskar, L. Dey and R. Saha (Dasgupta, Naskar, Saha, & Dey, 2018)	Dasgupta, T., Naskar, A., Dey, L., & Saha, R.	2018	Style, Content, linguistic and psychological	Deep Convolution Recurrent Neural Network	Yes (~8000 to 10000)	NO	Pearson's and Spearman's correlation of 0.94 and 0.97 respectively

1

Figure 1

The IEA architecture



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27715v1 | CC BY 4.0 Open Access | rec: 9 May 2019, publ: 9 May 2019

Figure 2

The IntelliMetric features model

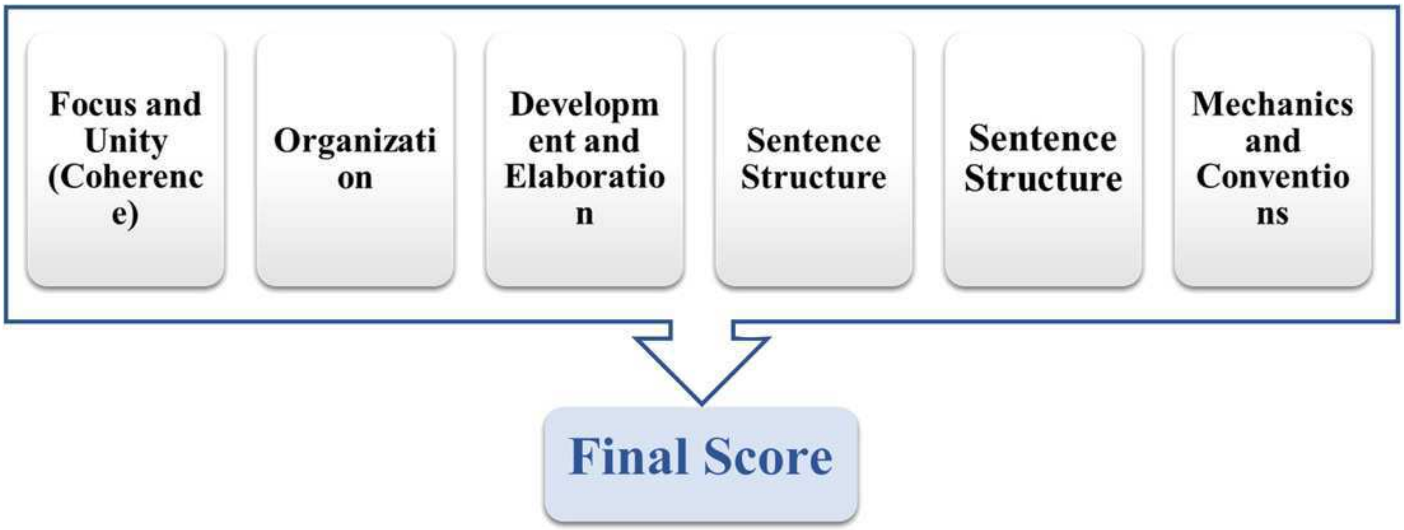
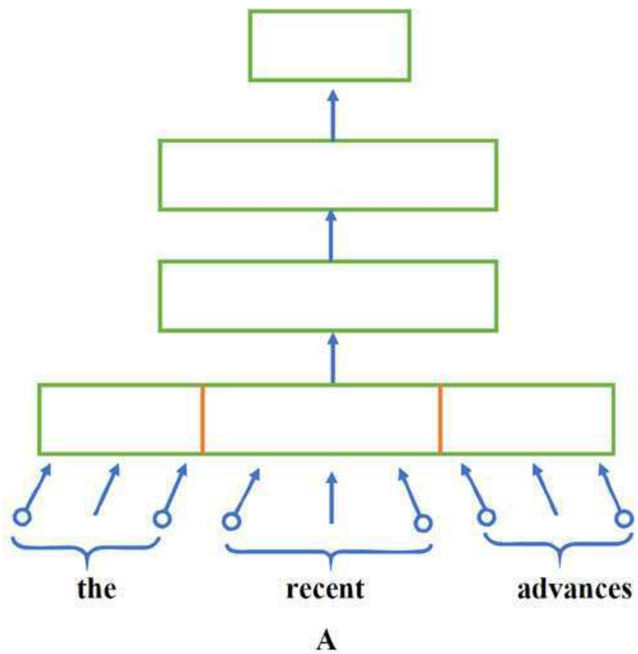


Figure 3

The architectures of two models

(A) Original C&W model. (B) Augmented C&W model



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27715v1 | CC BY 4.0 Open Access | rec: 9 May 2019, publ: 9 May 2019

Figure 4

The example of embeddings

(A) standard neural embeddings. (B) *SSWE* word embeddings

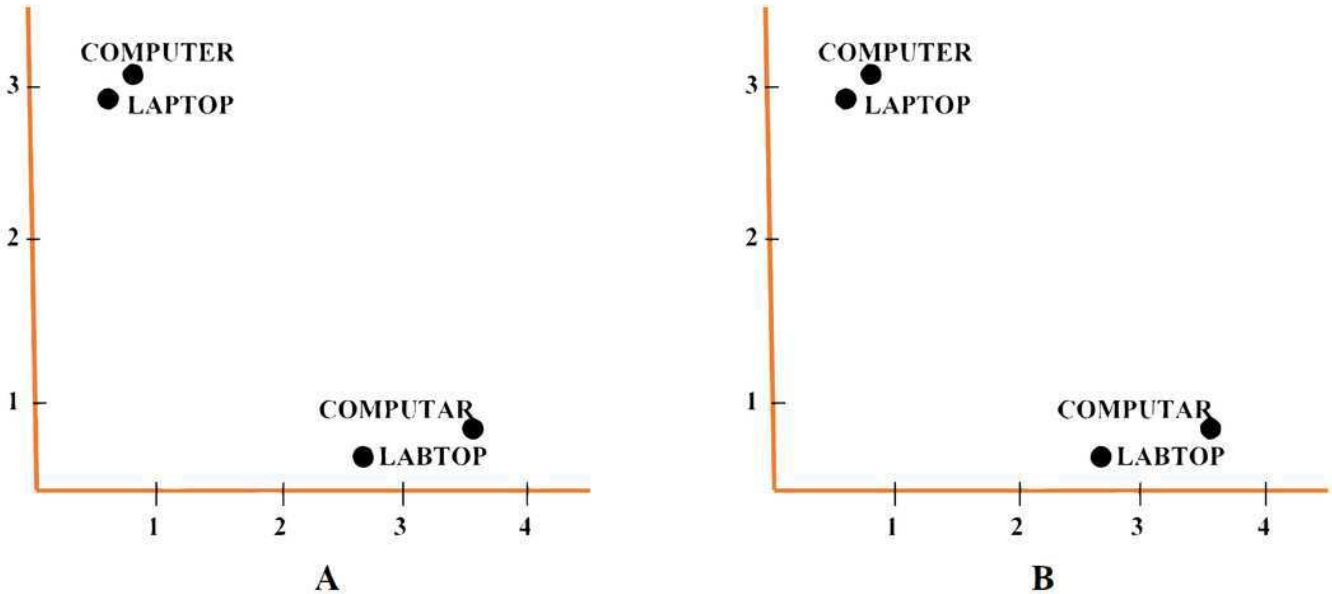
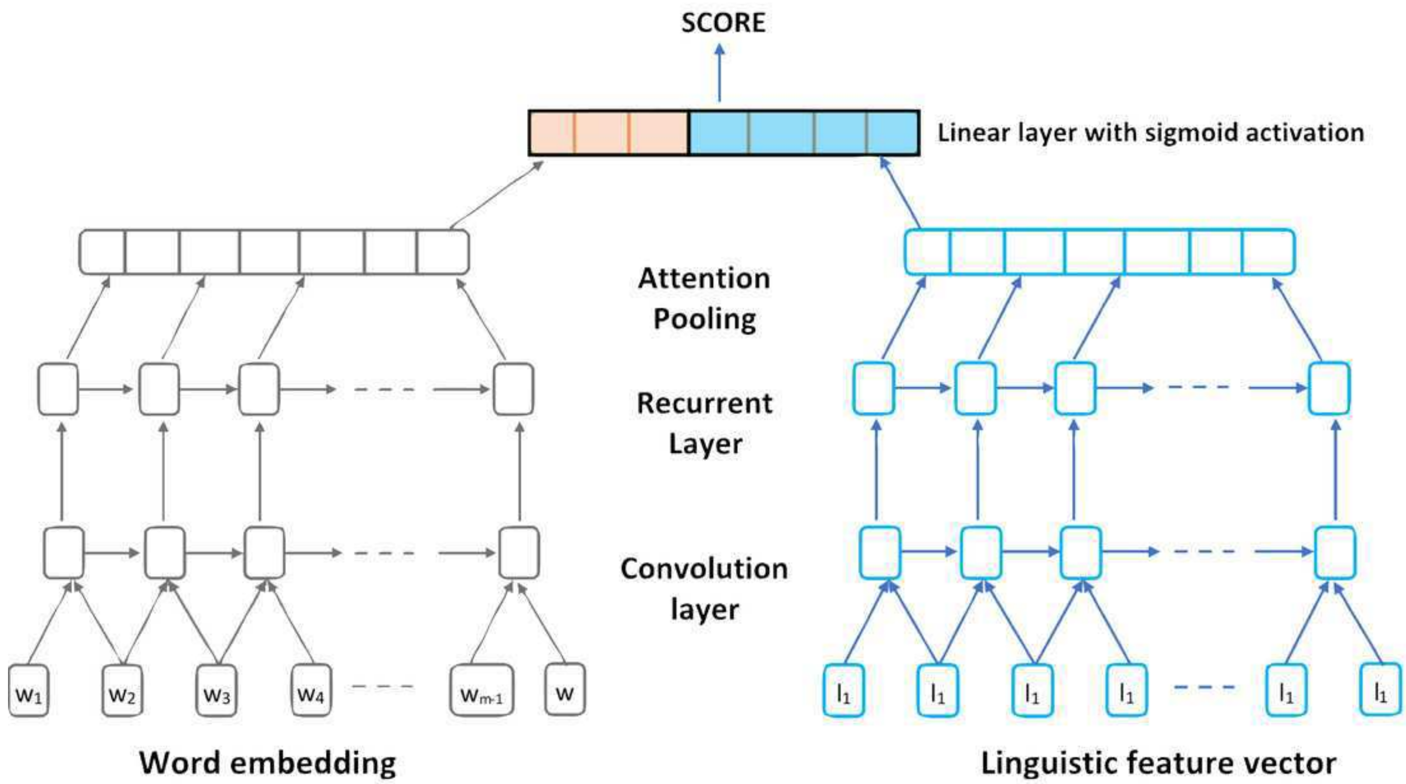


Figure 5

The proposed linguistically informed Convolution Recurrent Neural Network architecture



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27715v1 | CC BY 4.0 Open Access | rec: 9 May 2019, publ: 9 May 2019