Automated Essay Scoring: A Survey of the State of the Art

Zixuan Ke and Vincent Ng
Human Language Technology Research Institute, University of Texas at Dallas, USA
{zixuan, vince}@hlt.utdallas.edu

Abstract
Despite being investigated for over 50 years, the task of automated essay scoring continues to draw a lot of attention in the natural language processing community in part because of its commercial and educational values as well as the associated research challenges. This paper presents an overview of the major milestones made in automated essay scoring research since its inception.

1 Introduction
Automated essay scoring (AES), the task of employing computer technology to score written text, is one of the most important educational applications of natural language processing (NLP). This area of research began with Page’s [1966] pioneering work on the Project Essay Grader system and has remained active since then. The vast majority of work on AES has focused on holistic scoring, which summarizes the quality of an essay with a single score. There are at least two reasons for this focus. First, corpora manually annotated with holistic scores are publicly available, facilitating the development of learning-based holistic scoring engines. Second, holistic scoring technologies are commercially valuable: being able to automate the scoring of the millions of essays written for standardized aptitude tests such as the SAT and the GRE every year can save a lot of manual grading effort.

Though useful for scoring essays written for aptitude tests, holistic scoring technologies are far from adequate for use in classroom settings, where providing students with feedback on how to improve their essays is of utmost importance. Specifically, merely returning a low holistic score to a student provides essentially no feedback to her on which aspect(s) of the essay contributed to the low score and how it can be improved. In light of this weakness, researchers have recently begun work on scoring a particular dimension of essay quality such as coherence [Higgins et al., 2004; Somasundaran et al., 2014], technical errors, and relevance to prompt [Louis and Higgins, 2010; Persing and Ng, 2014]. Automated systems that provide instructional feedback along multiple dimensions of essay quality such as Criterion [Burstein et al., 2004] have also begun to emerge. Table 1 enumerates the aspects of an essay that could impact its holistic score. Providing scores along different dimensions of essay quality could help an author identify which aspects of her essay need improvements.

From a research perspective, one of the most interesting aspects of the AES task is that it encompasses a set of NLP problems that vary in the level of difficulty. The dimensions of quality in Table 1 are listed roughly in increasing difficulty of the corresponding scoring tasks. For instance, the detection of grammatical and mechanical errors has been extensively investigated with great successes. Towards the end of the list, we have a number of relatively less-studied but arguably rather challenging discourse-level problems that involve the computational modeling of different facets of text structure, such as coherence, thesis clarity, and persuasiveness. Modeling some of these challenging dimensions may even require an understanding of essay content, which is largely beyond the reach of state-of-the-art essay scoring engines.

Our goal in this paper is to provide the AI audience with an overview of the major milestones in AES research since its inception more than 50 years ago. While several books [Shermis and Burstein, 2003; Shermis and Burstein, 2013] and articles [Zupanc and Bosnic, 2016] exist that provide an overview of the state of the art in this area, we are not aware of any useful survey on AES that were published in the past three years. We therefore believe that this timely survey can provide up-to-date knowledge of the field to AI researchers. It is worth noting that another major subarea of automated essay grading concerns correcting the errors in an essay. Error correction is beyond the scope of this survey, but we refer the interested reader to Leacock et al. [2014] for an overview.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammaticality</td>
<td>Grammar</td>
</tr>
<tr>
<td>Usage</td>
<td>Use of prepositions, word usage</td>
</tr>
<tr>
<td>Mechanics</td>
<td>Spelling, punctuation, capitalization</td>
</tr>
<tr>
<td>Style</td>
<td>Word choice, sentence structure variety</td>
</tr>
<tr>
<td>Relevance</td>
<td>Relevance of the content to the prompt</td>
</tr>
<tr>
<td>Organization</td>
<td>How well the essay is structured</td>
</tr>
<tr>
<td>Development</td>
<td>Development of ideas with examples</td>
</tr>
<tr>
<td>Cohesion</td>
<td>Appropriate use of transition phrases</td>
</tr>
<tr>
<td>Coherence</td>
<td>Appropriate transitions between ideas</td>
</tr>
<tr>
<td>Thesis Clarity</td>
<td>Clarity of the thesis</td>
</tr>
<tr>
<td>Persuasiveness</td>
<td>Convincingness of the major argument</td>
</tr>
</tbody>
</table>

Table 1: Different dimensions of essay quality.
### Corpora Table

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Essay Types</th>
<th>Writer’s Language Level</th>
<th>No. of Essays</th>
<th>No. of Prompts</th>
<th>Scoring Task</th>
<th>Score Range</th>
<th>Additional Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLC-FCE</td>
<td>A,N,C,S,L</td>
<td>Non-native; ESOL test takers</td>
<td>1244</td>
<td>10</td>
<td>Holistic</td>
<td>1-40</td>
<td>Linguistic errors (~80 error types)</td>
</tr>
<tr>
<td>ASAP</td>
<td>A,R,N</td>
<td>US students; Grades 7 to 10</td>
<td>17450</td>
<td>8</td>
<td>Holistic</td>
<td>as small as [0-3]; as large as [0-60]</td>
<td>none</td>
</tr>
<tr>
<td>TOEFL11</td>
<td>A</td>
<td>Non-native; TOEFL test takers</td>
<td>1100</td>
<td>8</td>
<td>Holistic</td>
<td>Low, Medium, High</td>
<td>none</td>
</tr>
<tr>
<td>ICLE</td>
<td>A</td>
<td>Non-native; undergraduate students</td>
<td>1003</td>
<td>12</td>
<td>Organization</td>
<td>1-4 (at half-point increments)</td>
<td>none</td>
</tr>
<tr>
<td>AAE</td>
<td>A</td>
<td>Online community</td>
<td>102</td>
<td>101</td>
<td>Persuasiveness</td>
<td>1-6</td>
<td>Attributes impacting persuasiveness</td>
</tr>
</tbody>
</table>

Table 2: Comparison of several popularly used corpora for holistic and dimension-specific AES.

### 2 Corpora

In this section, we present five corpora that have been widely used for training and evaluating AES systems. Table 2 compares these corpora along eight dimensions: (1) the types of essays present in the corpus (argumentative (A), response (R), narrative (N), comment (C), suggestion (S) and letter (L)); (2) the language level of the essay writers; (3) the number of essays; (4) the number of prompts; (5) whether the scoring task is holistic or dimension-specific; (6) the score range of the essays; and (7) additional annotations on the corpus (if any).

The Cambridge Learner Corpus-First Certificate in English exam (CLC-FCE) [Yannakoudakis et al., 2011] provides for each essay both its holistic score and the manually tagged linguistic error types it contains (e.g., incorrect tense), which make it possible to build systems not only for holistic scoring but also for grammatical error detection and correction. However, the rather small number of essays per prompt makes it difficult to train high-performance prompt-specific systems (i.e., systems that are trained and tested on the same prompt).

The Automated Student Assessment Prize (ASAP) corpus was released as part of a Kaggle competition in 2012. Since then, it has become a widely used corpus for holistic scoring. The corpus is large in terms of not only the total number of essays, but also the number of essays per prompt (with up to 3000 essays per prompt). This makes it possible to build high-performance prompt-specific systems. However, it has at least two weaknesses that could limit its usefulness. First, the score ranges are different for different prompts, so it is difficult to train a model on multiple prompts. Second, the essays may not be "true to the original", as they do not contain any paragraph information and have gone through an aggressive preprocessing process that expunged both name entities and most other capitalized words.

The TOEFL11 corpus [Blanchard et al., 2013] contains essays from a real high-stakes exam, TOEFL. These essays are evenly distributed over eight prompts and 11 native languages spoken by the essay writers. The corpus is originally compiled for the Native Language Identification task, but it comes with a coarse level of proficiency consisting of only three levels, Low, Medium, and High. Some researchers have taken these proficiency labels as the holistic scores of the essays and attempted to train AES systems on them, but the underlying assumption that an essay’s quality can be represented by the language proficiency of its author is questionable.

One issue that hinders progress in dimension-specific essay scoring research concerns the scarcity of corpora manually annotated with dimension-specific scores. With the goal of performing dimension-specific scoring, members of our research group have annotated a subset of the essays in the International Corpus of Learner English (ICLE) [Granger et al., 2009] along several dimensions of essay quality, including (1) Organization, which refers to how well-organized an essay is [Persing et al., 2010]; (2) Thesis Clarity, which refers to how clearly an author explains the thesis of her essay [Persing and Ng, 2013]; (3) Prompt Adherence, which refers to how related an essay’s content is to the prompt for which it was written [Persing and Ng, 2014]; and (4) Argument Persuasiveness, which refers to the persuasiveness of the argument an essay makes for its thesis [Persing and Ng, 2015].

Another corpus annotated with dimension-specific scores is Argument Annotated Essays (AAE) [Stab and Gurevych, 2014]. The corpus contains 402 essays taken from essayforum2, a site offering feedback to students wishing to improve their ability to write persuasive essays for tests. Each essay was annotated with its argumentative structure (i.e., argument components such as claims and premises as well as the relationships between them (e.g., support, attack)). Recently, Carlile et al. [2018] scored each argument in 100 essays randomly selected from the corpus w.r.t. its persuasiveness.

All the corpora shown in Table 2 are in English. AES corpora in other languages exist, such as Ostling’s [2013] Swedish corpus and Horbach et al.’s [2017] German corpus.

### 3 Systems

Next, we characterize existing AES systems along three dimensions: the scoring task a system was developed for as well as the approach and the features it employed.

#### 3.1 Tasks

The vast majority of existing AES systems were developed for holistic scoring. Dimension-specific scoring did not start until 2004. So far, several dimensions of quality have been examined, including organization [Persing et al., 2010], thesis clarity [Persing and Ng, 2013], argument persuasiveness...
3.2 Approaches

Virtually all existing AES systems are learning-based and can be classified based on whether they employ supervised, weakly supervised, or reinforcement learning. Since state-of-the-art AES systems are all supervised, we will focus our discussion on supervised approaches to AES in this subsection, and refer the reader to Chen et al. [2010] and Wang et al. [2018] for the application of weakly supervised learning and reinforcement learning to AES, respectively.

Researchers adopting supervised approaches to AES have recast the task as (1) a regression task, where the goal is to predict the score of an essay; (2) a classification task, where the goal is to classify an essay as belonging to one of a small number of classes (e.g., low, medium, or high, as in the aforementioned TOEFL11 corpus); or (3) a ranking task, where the goal is to rank two or more essays based on their quality.

Off-the-shelf learning algorithms are typically used for model training. For regression, linear regression [Page, 1966; Landauer et al., 2003; Miltsakaki and Kukich, 2004; Attali and Burstein, 2006; Klebanov et al., 2013; Faulkner, 2014; Crossley et al., 2015; Klebanov et al., 2016], support vector regression [Persing et al., 2010; Persing and Ng, 2013; Persing and Ng, 2014; Persing and Ng, 2015; Cozma et al., 2018], and sequential minimal optimization (SMO, a variant of support vector machines) [Vajjala, 2018] are typically used. For classification, SMO [Vajjala, 2018], logistic regression [Farra et al., 2015; Nguyen and Litman, 2018] and Bayesian network classification [Rudner and Liang, 2002] have been used. Finally, for ranking, SVM ranking [Yannakoudakis et al., 2011; Yannakoudakis and Briscoe, 2012] and LambdaMART [Chen and He, 2013] have been used.

Neural Approaches

Many recent AES systems are neural-based. While a lot of traditional work on AES has focused on feature engineering (see Section 3.3 for a detailed discussion on features for AES), an often-cited advantage of neural approaches is that they obviate the need for feature engineering.

The first neural approach to holistic essay scoring was proposed by Taghipour and Ng [2016] (T&N). Taking the sequence of (one-hot vectors of the) words in an essay as input, their model first uses a convolution layer to extract n-gram level features. These features, which capture the local textual dependencies among the words in an n-gram, are then passed to a recurrent layer composed of a Long-Short Term Memory (LSTM) network [Hochreiter and Schmidhuber, 1997], which outputs a vector at each time step that captures the long-distance dependencies of the words in the essay. The vectors from different time steps are then concatenated to form a vector that serves as the input to a dense layer to predict the essay’s score. As the model is trained, the one-hot input vectors mentioned above are being updated.

Though not all subsequent neural AES models are extensions of T&N’s model, they all attempt to address one or more of its weaknesses, as described in the following subsections.

Learning Score-Specific Word Embeddings

Some words have little power in discriminating between good and bad essays. Failure to distinguish these under-informative words from their informative counterparts may hurt AES performance. In light of this problem, Alikaniotis et al. [2016] train word embeddings. Informally, a word embedding is a low-dimensional real-valued vector representation of a word that can be trained so that two words that are semantically similar are close to each other in the embedding space. For instance, “king” and “queen” should have similar embeddings, whereas “king” and “table” should not. Hence, word embeddings are generally considered a better representation of word semantics than the one-hot word vectors used by T&N. Although word embeddings can be trained on a large, unannotated corpus using a word embedding learning neural network architecture known as the CW model [Collobert and Weston, 2008], Alikaniotis et al. propose to train task-specific word embeddings by augmenting the CW model with an additional output that corresponds to the score of the essay in which the input word appears. These score-specific word embeddings (SSWEs), which they believe can better discriminate between informative and under-informative words, are then used as features for training a neural AES model.

Modeling Document Structure

Both T&N and Alikaniotis et al. [2016] model a document as a linear sequence of words. Dong and Zhang [2016] hypothesize that a neural AES model can be improved by modeling the hierarchical structure of a document, wherein a document is assumed to be created by first (1) combining its words to form its sentences and then (2) combining the resulting sentences to form the document. Consequently, their model uses two convolution layers that correspond to this two-level hierarchical structure, a word-level convolution layer and a sentence-level convolution layer. Like T&N, the word-level convolution layer takes the one-hot word vectors as input and extracts the n-gram level features from each sentence independently of other sentences. After passing through a pooling layer, the n-gram level features extracted from each sentence are then condensed into a “sentence” vector. The sentence-level convolution layer then takes the sentence vectors generated from different sentences of the essay as input and extracts n-gram level features over different sentences.

Using Attention

As mentioned above, some characters, words and sentences in an essay are more important than the others as far as scoring is concerned and therefore should be given more attention. However, Dong and Zhang’s two convolution layer neural network fails to do so. To automatically identify important characters, words and sentences, Dong et al. [2017] incorporate an attention mechanism [Sutskever et al., 2014] into the network by using attention pooling rather than simple pooling such as max or average pooling after each layer. Specifically, each attention pooling layer takes the output of the corresponding convolution layer as input, leveraging a trainable weight matrix to output vectors that are a weighted combination of the input vectors.
Modeling Coherence

Tay et al. [2018] hypothesize that holistic scoring can be improved by computing and exploiting the coherence score of an essay, since coherence is an important dimension of essay quality. They model coherence as follows. Like T&N, they employ a LSTM as their neural network. Unlike T&N, however, they employ an additional layer in their neural model that takes as inputs two positional outputs of the LSTM collected from different time steps and compute the similarity for each such pair of positional outputs. They call these similarity values \textit{neural coherence features}. The reason is that intuitively, coherence should correlate positively with similarity. These neural coherence features are then used to augment the vector that the LSTM outputs (i.e., the vector that encodes local and long-distance dependencies, as in T&N). Finally, they predict the holistic score using the augmented vector, effectively exploiting coherence in the scoring process.

Transfer Learning

Ideally, we can train prompt-specific AES systems, in which the training prompt and the test (i.e., target) prompt are the same, because this would allow AES systems to exploit the prompt-specific knowledge they learned from the training essays to more accurately score the test essays. In practice, however, it is rarely the case that enough essays for the target prompt are available for training. As a result, many AES systems are trained in a prompt-independent manner, meaning that a small number of target-prompt essays and a comparatively larger set of non-target-prompt (i.e., source-prompt) essays are typically used for training. However, the potential mismatch in the vocabulary used in the essays written for the source prompt(s) and those for the target prompt may hurt the performance of prompt-independent systems. To address this issue, researchers have investigated the use of transfer learning (i.e., domain adaptation) techniques to adapt the source prompt(s)/domain(s) to the target prompt/domain.

EasyAdapt [Daumé III, 2007], one of the simple but effective transfer learning algorithms, assumes as input training data from only two domains (the source domain and the target domain), and the goal is to learn a model that can perform well when classifying the test instances from the target domain. To understand EasyAdapt, recall that a model that does \textit{not} use transfer learning is typically trained by employing a feature space that is shared by the instances from both the source domain and the target domain. EasyAdapt augments this feature set by duplicating each feature in the space three times, where the first copy stores the information shared by both domains, the second copy stores the source-domain information, and the last copy stores the target-domain information. It can be proven that in this augmented feature space, the target-domain information will be given twice as much importance as the source-domain information, thus allowing the model to better adapt to the target-domain information.

When applying transfer learning to AES, we can view prompts as domains. In a realistic scenario, there are one target prompt and multiple source prompts available for training. However, since EasyAdapt can only handle one target domain and one source domain, researchers who have applied EasyAdapt to AES treat all source prompts as belonging to the same source domain. In their transfer learning work, Phandi et al. [2015] generalize EasyAdapt to Correlated Bayesian Linear Ridge Regression, enabling the weight given to the target-prompt information to be learned (rather than fixed to 2 as in EasyAdapt). Cummins et al. [2016] also perform transfer learning, employing EasyAdapt to augment the feature space and training a pairwise ranker to rank two essays that are constrained to be from the same prompt.

While the above systems assume that a small number of essays from the target prompt is available for training, Jin et al. [2018] perform transfer learning under the assumption that no target-prompt essays are available for training via a two-stage framework. Stage 1 aims to identify the (target-prompt) essays in the test set with extreme quality (i.e., those that should receive very high or very low scores). To do so, they train a model on the (source-prompt) essays using \textit{prompt-independent} features (e.g., those based on grammatical and spelling errors) and use it to score the (target-domain) test essays. The underlying assumption is that those test essays with extreme quality can be identified with general (i.e., prompt-independent) features. Stage 2 aims to score the remaining essays in the test set (i.e., those with non-extreme quality). To do so, they first automatically label each low-quality essay and each high-quality essay identified in the first stage as 0 and 1, respectively. They then train a regressor on these automatically-labeled essays using \textit{prompt-specific} features, under the assumption that these specific features are needed to properly capture the meaning of the essays with non-extreme quality. Finally, they use the regressor to score the remaining test essays, whose scores are expected to fall between 0 and 1 given their non-extreme quality.

3.3 Features

A large amount of work on AES has involved feature development. While the recently developed neural models for AES obviate the need for feature engineering, we believe that feature development will continue to play a crucial role in AES research, for the following reasons. First, for neural models to be effective, they need to be trained on a large amount of annotated data. Even if we believe we have enough data for training accurate AES models for English, the same is not true for the vast majority of natural languages. To build AES systems for these languages, the most practical way is to employ a feature-based approach. Second, even for English, the amount of data available for training dimension-specific AES systems is fairly limited. Until we have bigger annotated corpora, feature engineering will remain an important step when building dimension-specific AES systems. Third, while many neural holistic scoring models have achieved state-of-the-art results, it is possible that these models can be further improved by incorporating hand-crafted features obtained via feature engineering. Overall, we believe that feature-based approaches and neural approaches should be viewed as complementary rather than competing approaches. In this subsection, we describe the features that have been used for AES.

Length-based features are one of the most important feature types for AES, as length is found to be highly positively correlated with the holistic score of an essay. These features encode the length of an essay in terms of the number of sen-
tences, words, and/or characters in the essay.

Lexical features can be divided into two categories. One category contains the word unigrams, bigrams, and trigrams that appear in an essay. These word n-grams are useful because they encode the grammatical, semantic, and discourse information about an essay that could be useful for AES. For instance, the bigram "people is" suggests ungrammaticality; the use of discourse connectives (e.g., "moreover", "however") suggest cohesion; and certain n-grams indicate the presence of topics that may be relevant to a particular prompt. The key advantage of using n-grams as features is that they are language-independent. The downside, however, is that lots of training data are typically needed to learn which word n-grams are useful. Another category contains statistics computed based on word n-grams, particularly unigrams. For instance, there are features that encode the number of occurrences of a particular punctuation in an essay [Page, 1966; Chen and He, 2013; Phandi et al., 2015; Zesch et al., 2015].

Embeddings, which can be seen as a variant of n-gram features, are arguably a better representation of the semantics of a word/phrase than word n-grams. Three types of embedding-based features have been used for AES. The first type contains features computed based on embeddings pretrained on a large corpus such as GLoVe [Pennington et al., 2014]. For instance, Cozma et al. [2018] use bag-of-super-word-embeddings. Specifically, they cluster the pretrained word embeddings using k-means and represent each word using the centroid of the cluster it belongs to. The second type contains features computed based on AES-specific embeddings, such as the SSWEs [Alikaniotis et al., 2016] mentioned earlier. The third type contains features that are originally one-hot word vectors, but are being updated as the neural model learns them (e.g., the number of claims and premises in a paragraph) as well as the structure of the argument tree where the nodes correspond to the argument components (e.g., claims, premises) and the edges correspond to the relationship between two components (e.g., whether one component support or attack the other). For instance, an essay typically has a major claim, which encodes the stance of the author w.r.t. the essay’s topic. The major claim is supported or attacked by one or more claims (controversial statements that should not be readily accepted by the reader without further evidences), each of which is in turn supported or attacked by one or more premises (evidences for the corresponding claim). Argumentation features are computed based on the argument components and the relationships between them (e.g., the number of claims and premises in a paragraph) as well as the structure of the argument tree (e.g., the tree depth) [Ghosh et al., 2016; Wachsmuth et al., 2016; Nguyen and Litman, 2018].

Word category features are computed based on wordlists or dictionaries, each of which contains words that belong to a particular lexical, syntactic, or semantic category. For instance, features are computed based on lists containing discourse connectives, correctly spelled words, sentiment words, and modals [Yannakoudakis and Briscoe, 2012; Farra et al., 2015; McNamara et al., 2015; Cummins et al., 2016; Amorim et al., 2018], as the presence of certain categories of words in an essay could reveal a writer’s ability to organize her ideas, compose a cohesive and coherent response to the prompt, and master standard English. Wordlists that encode which of the eight levels of word complexity that a word belongs to have also been used [Breland et al., 1994]. Intuitively, a higher word level indicates a more sophisticated vocabulary usage. Word category features help generalize word n-gram features and are particularly useful when only a small amount of training data is available.

Prompt-relevant features encode the relevance of the essay to the prompt it was written for. Intuitively, an essay that is not adherent to the prompt cannot receive a high score. Different measures of similarity are used to compute the relevance of an essay to the prompt, such as the number of word overlap and its variants [Louis and Higgins, 2010], word topicality [Klebanov et al., 2016], and semantic similarity as measured by random indexing [Higgins et al., 2004].

Readability features encode how difficult an essay is to read. Readability is largely dependent on word choice. While good essays should not be overly difficult to read, they should not be too easy to read either: in a good essay, the writer should demonstrate a broad vocabulary and a variety of sentence structures. Readability is typically measured using readability metrics such as Flesch-Kindcaid Reading Ease [Zesch et al., 2015] and simple measures such as the type-token ratio (the number of unique words to the total number of words in an essay).

Syntactic features encode the syntactic information about an essay. There are three main types of syntactic features. Part-of-speech (POS) tag sequences provide syntactic generalizations of word n-grams and are used to encode ungrammaticality (e.g., plural nouns followed by singular verbs) and style (using the ratio of POS tags) [Zesch et al., 2015]. Parse trees have also been used. For instance, the depth of a parse tree is used to encode how complex the syntactic structure of a sentence is [Chen and He, 2013]; phrase structure rules are used to encode the presence of different grammatical constructions; and grammatical/dependency relations are used to compute the syntactic distance between a head and its dependent. Grammatical error rates are used to derive features that encode how frequently grammatical errors appear in an essay, and are computed either using a language model or from hand-annotated grammatical error types [Yannakoudakis et al., 2011; Yannakoudakis and Briscoe, 2012].

Argumentation features are computed based on the argumentative structure of an essay. As a result, these features are only applicable to a persuasive essay, where an argumentative structure is present, and have often been used to predict the persuasiveness of an argument made in an essay [Persing and Ng, 2015]. The argumentative structure of an essay is a tree structure where the nodes correspond to the argument components (e.g., claims, premises) and the edges correspond to the relationship between two components (e.g., whether one component support or attack the other). For instance, an essay typically has a major claim, which encodes the stance of the author with respect to the essay’s topic. The major claim is supported or attacked by one or more claims (controversial statements that should not be readily accepted by the reader without further evidences), each of which is in turn supported or attacked by one or more premises (evidences for the corresponding claim). Argumentation features are computed based on the argument components and the relationships between them (e.g., the number of claims and premises in a paragraph) as well as the structure of the argument tree (e.g., the tree depth) [Ghosh et al., 2016; Wachsmuth et al., 2016; Nguyen and Litman, 2018].

Semantic features encode the lexical semantic relations between different words in an essay. There are two main types of semantic features. Histogram-based features [Klebanov and Flor, 2013] are computed as follows. First, the pointwise mutual information (PMI), which measures the degree of association between two words based on co-occurrence, is computed between each pair of words in an essay. Second, a histogram is constructed by binning the PMI values, where the value of a bin is the percentage of word pairs having a PMI value that falls within the bin. Finally,
features are computed based on the histogram. Intuitively, a higher proportion of highly associated pairs is likely to indicate a better development of topics, and a higher proportion of lowly associated pairs is likely to indicate a more creative use of language. Frame-based features are computed based on the semantic frames in FrameNet [Baker et al., 1998]. Briefly, a frame may describe an event that occurs in a sentence, and the frame’s event elements may be the people or the objects that participate in the corresponding event. For a more concrete example, consider the sentence “they said they do not believe that the prison system is outdated”. This sentence contains a Statement frame because a statement is made in it, describing an event in which “they” participate as a Speaker. Knowing that this opinion was expressed by someone other than the author can be helpful for scoring the clarity of the thesis of an essay [Persing and Ng, 2013], as thesis clarity should be measured based on the author’s opinion.

Discourse features, which encode the discourse structure of an essay, have been derived from (1) entity grids, (2) Rhetorical Structure Theory (RST) trees, (3) lexical chains, and (4) discourse function labels. Entity grids, which are a discourse representation designed by Barzilay and Lapata [2008] to capture the local coherence of text based on Centering Theory [Grosz et al., 1995], have been used to derive local coherence features [Yannakoudakis and Briscoe, 2012]. Discourse parse trees constructed based on RST [Mann and Thompson, 1988] encode the hierarchical discourse structure of text (e.g., is one discourse segment an elaboration of the other, or is it in contrast relation with the other?) and have been used to derive features that capture the local and global coherence of an essay [Somasundaran et al., 2014]. Lexical chains, which are sequences of related words in a document, have been used as an indicator of text cohesion [Morris and Hirst, 1991]. Intuitively, an essay that contains many lexical chains, especially ones where the beginning and end of the chain cover a large span of the essay, tend to be more cohesive [Somasundaran et al., 2014]. A discourse function label is defined on a sentence or paragraph that indicates its discourse function in a given essay (e.g., whether the paragraph is an introduction or a conclusion, whether a sentence is the thesis of the essay). These labels have been used to derive features for scoring organization [Persing et al., 2010].

4 Evaluation

In this section, we discuss the metrics and schemas used to evaluate AES systems.

The most widely adopted evaluation metric is Quadratic weighted Kappa (QWK), an agreement metric that ranges from 0 to 1 but can be negative if there is less agreement than what is expected by chance. Other widely used metrics include error metrics such as Mean Absolute Error (MAE) and Mean Square Error (MSE) and correlation metrics such as Pearson’s Correlation Coefficient (PCC) and Spearman’s Correlation Coefficient (SCC). A detailed discussion of the appropriateness of these and other metrics for AES can be found in Yannakoudakis and Cummins [2015].

There are two evaluation schemas in AES. In an in-domain evaluation, a system is trained and evaluated on the same prompt and its overall performance is measured by averaging its performance across all prompts. In a cross-domain evaluation, a system is trained and evaluated on different prompts. This evaluation schema is typically used to evaluate AES systems that perform transfer learning.

5 The State of the Art

In this section, we provide an overview of the systems that have achieved state-of-the-art results on the five evaluation corpora described in Section 2. Results, which are expressed in terms of QWK, PCC and MAE, are shown in Table 3.

Several points deserve mention. First, for holistic scoring (CLC-FCE, ASAP and TOEFL11), both QWK and PCC are quite high: e.g., both in-domain and cross-domain scores are above 0.6. Second, the dimension-specific scoring results (on ICLE and AAE) in terms of PCC are worse than their holistic counterparts. Nevertheless, these results do not necessarily suggest that holistic scoring is easier than domain-specific scoring, for at least two reasons. First, these results are not directly comparable as they are obtained on different corpora. Second, the number of essays used to train the holistic scorers tend to be larger than those used to train the dimensionally.

See https://www.kaggle.com/c/asap-aes#evaluation for details.

In-domain and cross-domain results are available for ASAP, so we report both. For the cross-domain results, we use the notation “X→Y” to denote “training on prompt X and testing on prompt Y”.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>System</th>
<th>Scoring Task</th>
<th>Approach</th>
<th>Features</th>
<th>Evaluation Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>E</td>
</tr>
<tr>
<td>CLC-FCE</td>
<td>Yannakoudakis and Briscoe [2012]</td>
<td>Holistic</td>
<td>Ranking</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ASAP</td>
<td>Cozma et al. [2018] (In-domain)</td>
<td>Holistic</td>
<td>Regression</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TOEFL11</td>
<td>Vajjala [2018]</td>
<td>Holistic</td>
<td>Regression</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ICLE</td>
<td>Persing and Ng [2013]</td>
<td>Thesis Clarity</td>
<td>Regression</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Wachsmuth et al. [2016]</td>
<td>Persuasiveness</td>
<td>Regression</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
| AAE | Ke et al. [2018] | Persuasiveness | (Neural) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |         |         |         | Table 3: Performance of state-of-the-art AES systems on commonly-used evaluation corpora. The features are divided into ten categories: length-based (L), lexical (X), word embeddings (E), category-based (C), prompt-relevant (P), readability (R), syntactic (S), argumentation (A), semantic (M), and discourse (D).
specific scorers. What these results do suggest, however, is that dimension-specific scoring is far from being solved.

6 Concluding Remarks

While researchers are making continued progress on AES despite its difficulty, a natural question is: what are the promising directions for future work?

One concerns feedback to students. As mentioned before, there have been recent attempts to improve the feedback provided to students by scoring an essay along specific dimensions of quality, so that if a student receives a low holistic score, she will have an idea of which dimensions of quality need improvement. However, one can argue that this feedback is still not adequate, as a student who receives a low score for a particular dimension may not know why the score is low. Recent work by Ke et al. [2018] has begun examining this problem by identifying the attributes of an argument that could impact its persuasiveness score. Two of the attributes they identified are Specificity (how specific the statements in the argument are) and Evidence (how strong the evidences are in support of the claim being made in the argument). Intuitively, a persuasive argument should be specific and have strong evidences in support of the claim. Hence, scoring these attributes of an argument in addition to its persuasiveness will enable additional feedback to be provided to students: if a student’s argument receives a low persuasiveness score, she will have an idea of which aspect(s) of the argument should be improved by examining the attribute scores. Overall, we believe feedback is an area that deserves more attention.

Another direction concerns data annotation. As mentioned before, progress in dimension-specific scoring research is hindered in part by the scarcity of annotated corpora needed for model training. An issue that is often under-emphasized is which corpora one should choose for data annotation. We envision that in the long run, substantial progress in AES research can only be made if different researchers create their annotations on the same corpus. For instance, having a corpus of essays that are scored along multiple dimensions of quality can facilitate the study of how these dimensions interact with each other to produce a holistic score, allowing us to train joint models that enable these challenging dimension-specific scoring tasks to help each other via multi-task learning.

Large-scale data annotation takes time, but it by no means implies that progress in AES research cannot be made before data annotation is complete. One can explore methods for learning robust models in the absence of large amounts of annotated training data. For instance, one can leverage BERT [Devlin et al., 2019], a new language representation model that has recently been used to achieve state-of-the-art results on a variety of NLP tasks. The idea is to first use BERT to pretrain deep bidirectional representations from a large amount of unlabeled data, and then fine-tune the resulting model with one additional output layer, which in our case is the layer for scoring. Another possibility is to augment an input essay with hand-crafted features when training neural models for AES.

In addition to exploring the interaction between different dimensions, we believe it is worthwhile to examine how AES interacts with other areas of essay grading research, such as automated essay revision [Zhang et al., 2017], where the goal is to revise, for instance, a thesis or an argument in an essay to make it stronger. Automated essay revision could benefit from argument persuasiveness scores. Specifically, the first step in deciding how to revise a argument to make it stronger is to understand why it is weak, and the aforementioned attributes Ke et al. [2018] identified can provide insights into what makes an argument weak and therefore how to revise it.

Finally, to enable AES technologies to be deployed in a classroom setting, it is important to conduct user studies that allow students to report whether the feedback they obtain from AES systems can help improve their writing skills.

Acknowledgments

We thank the anonymous reviewer for his/her helpful comments on an earlier draft of the paper. This work was supported in part by NSF Grants IIS-1528037 and CCF-1848608.

References


[Crossley et al., 2015] Scott Crossley, Laura K. Allen, Erica L. Snow, and Danielle S. McNamara. Pssst... textual features... there is more to automatic essay scoring than just you! In Proc. of ICLACL, pages 203–207, 2015.


[McNamara et al., 2015] Danielle S. McNamara, Scott A. Crossley, Rod D. Roscoe, Laura K. Allen, and Jianmin


