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

ZIXUAN KE AND VINCENT NG

HUMAN LANGUAGE TECHNOLOGY RESEARCH INSTITUTE
THE UNIVERSITY OF TEXAS AT DALLAS

08/14/2019

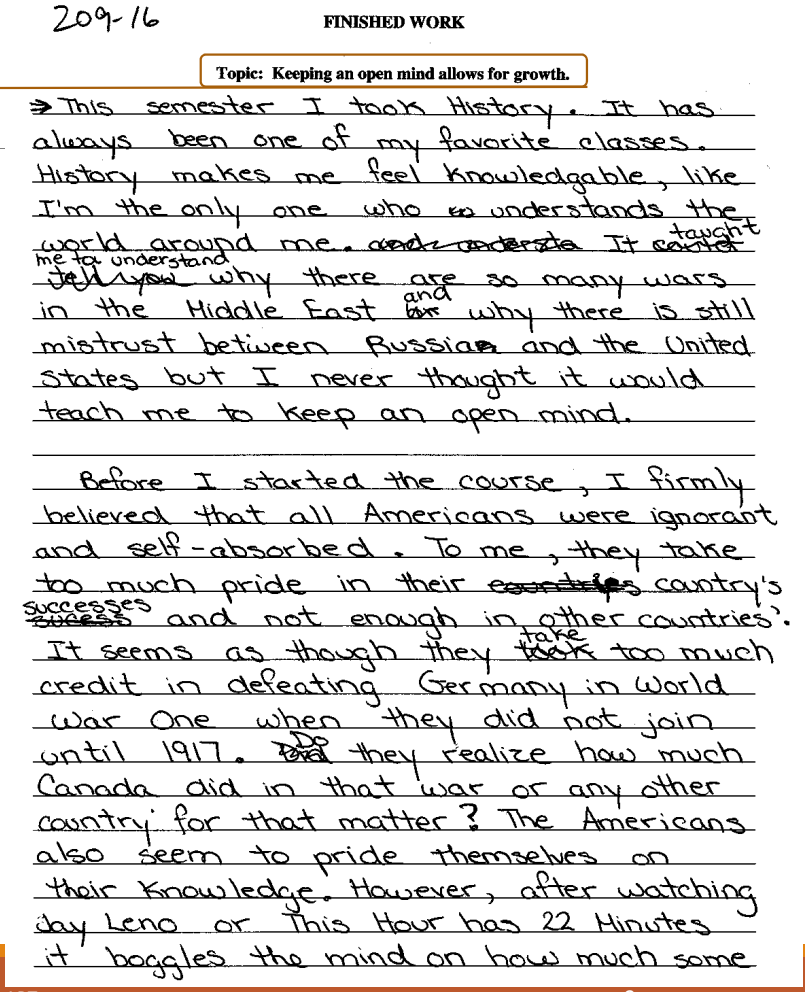
Automated Essay Scoring

Essay

High-stakes Testing, e.g. TOEFL, GRE

Classroom Setting, e.g. homework assignment, in-class writing

Prompt/Topic



209-16

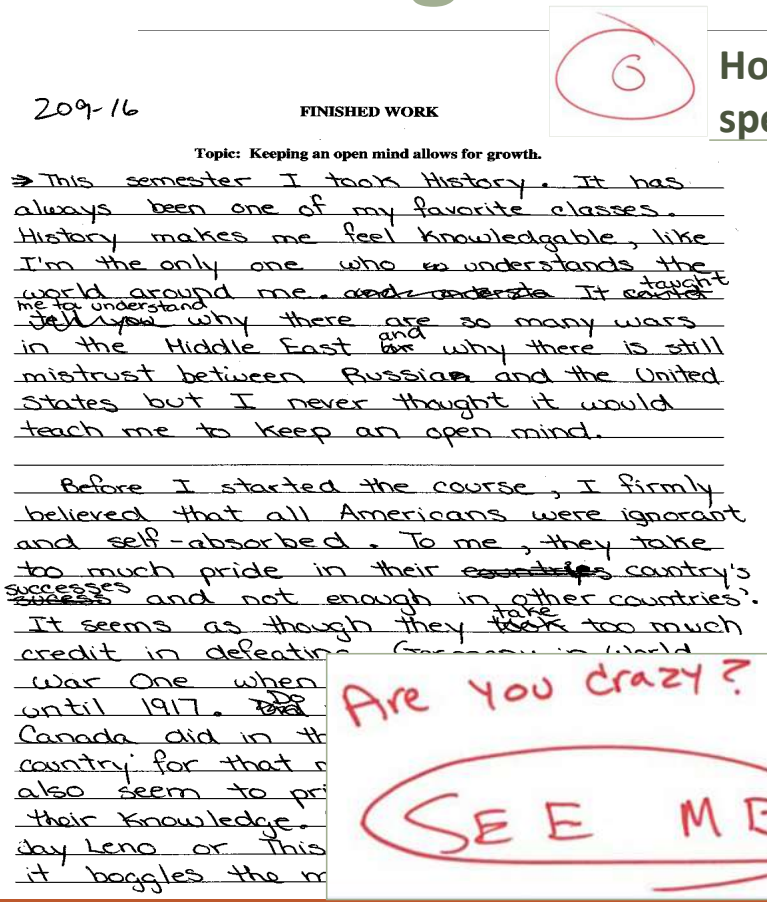
FINISHED WORK

Topic: Keeping an open mind allows for growth.

This semester I took History. It has always been one of my favorite classes. History makes me feel knowledgable, like I'm the only one who understands the world around me. It taught me to understand why there are so many wars in the Middle East and why there is still mistrust between Russia and the United States but I never thought it would teach me to keep an open mind.

Before I started the course, I firmly believed that all Americans were ignorant and self-absorbed. To me, they take too much pride in their country's successes and not enough in other countries'. It seems as though they take too much credit in defeating Germany in World War One when they did not join until 1917. They realize how much Canada did in that war or any other country for that matter? The Americans also seem to pride themselves on their knowledge. However, after watching Jay Leno or This Hour has 22 Minutes it boggles the mind on how much some

Scoring



Holistic Scoring **or Dimension-specific** scoring

How persuasive is the argument? (Persuasiveness)

How does it adherent to the topic? (Adherence)

How is its logical organization? (Coherence)

How clear is its thesis? (Thesis Clarity)

Some useful feedback to student from teacher

Automated

Essay scoring is a time-consuming, laborious task.

For teachers:

Long hours

For standardized testing services (e.g. ETS):

High cost, due to amount of human labor required.

For students:

Inability to judge their own work.

Goals

Provide an overview of the major milestones in AES research since its inception more than 50 years ago

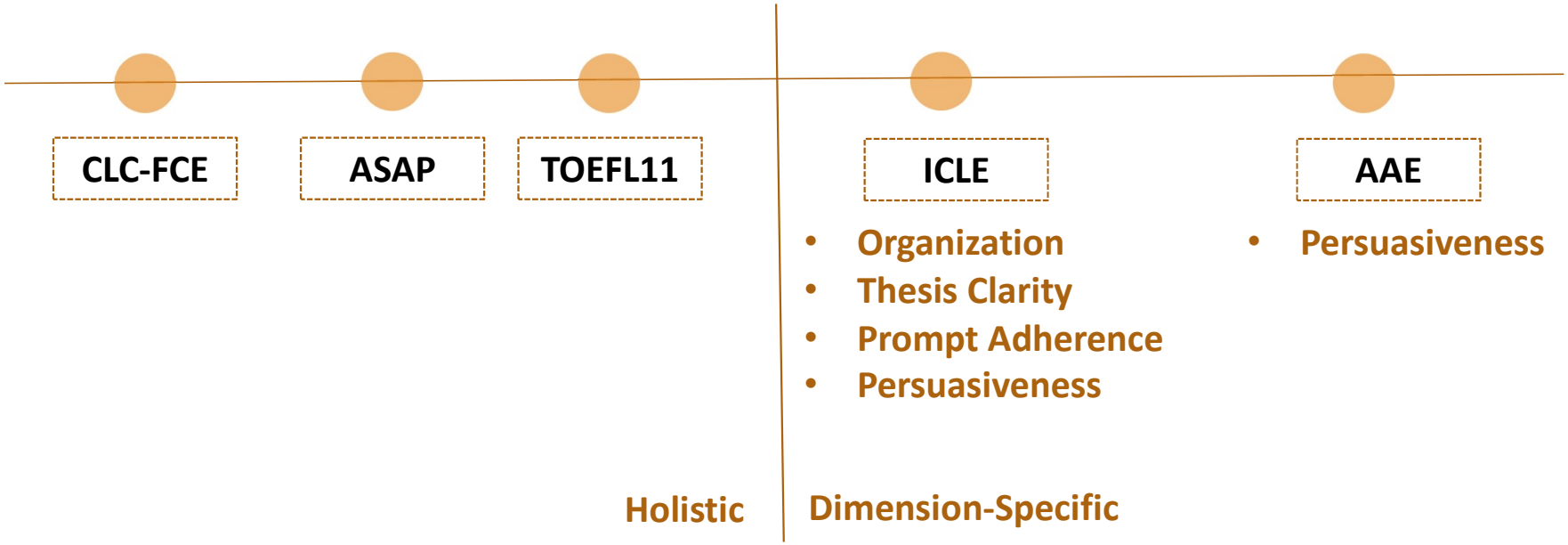
While server books and articles exist, we don't aware of any useful survey on AES that were published in the past three years

Plan for the talk

- Corpora
- Systems
- Evaluation and State of the Art
- Concluding Remarks

Corpora

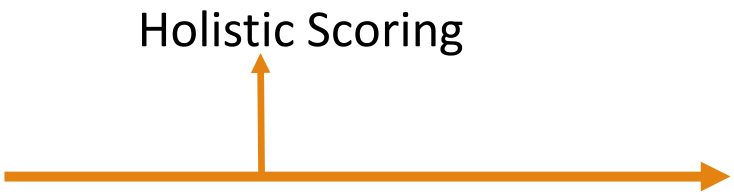
5 publicly available English Corpora



Plan for the talk

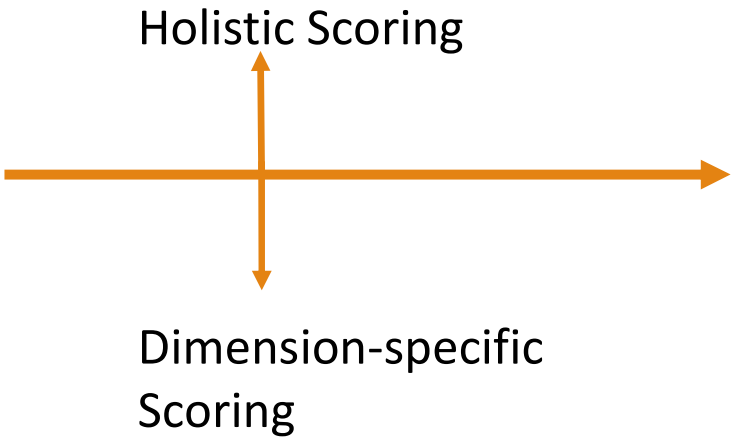
- Corpora
- Systems
- Evaluation and State of the Art
- Concluding Remarks

Tasks



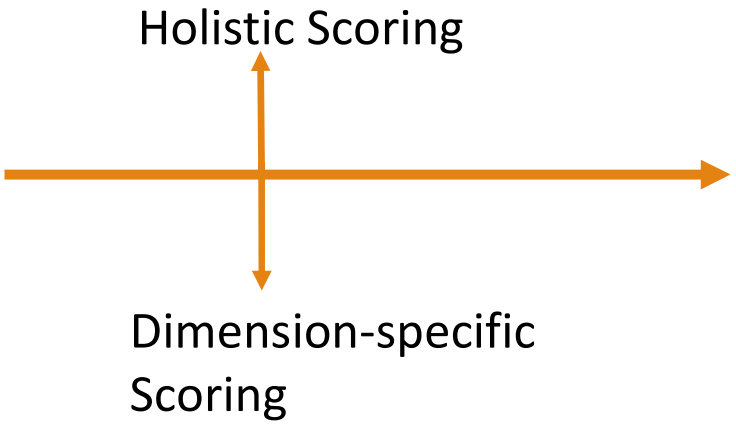
- Vast Majority of existing AES systems were developed for holistic scoring
  - Corpora manually annotated with holistic scoring are publicly available
  - Holistic scoring technologies are commercially valuable

Tasks



- Holistic Scoring is far from adequate for use in classroom settings
  - Merely returning a low holistic score to a student provides essentially **no feedback** to her on **which aspects** of the essay contributed to the low score and **how it can be improved**

Tasks



- They are both challenging
  - Discourse-level problems that involve the computational modeling of different facets of text structure
  - An understanding of essay content is required

Approaches: Off-the-shelf

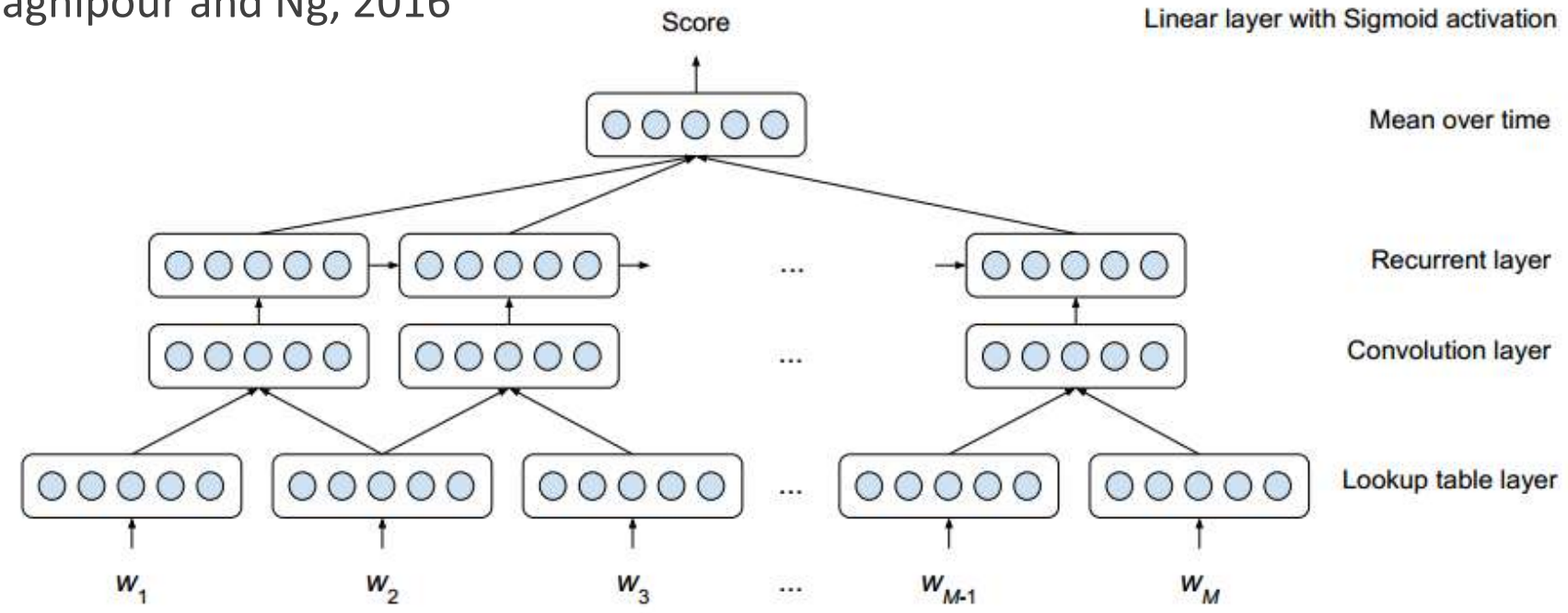
- As an regression task
  - Linear Regression
  - Support Vector Regression
  - Sequential Minimal Optimization
- As a classification task
  - Logistic Regression
  - Sequential Minimal Optimization
  - Bayesian network classification
- As a ranking task
  - SVM ranking
  - LambdaMART

Approaches: Neural Approaches

- Many recent AES systems are neural-based
  - Many traditional work on AES has focused on feature engineering
  - An often-cited advantage of neural approaches is that they obviate the need for feature engineering

Approaches: Neural Approaches

- Taghipour and Ng, 2016

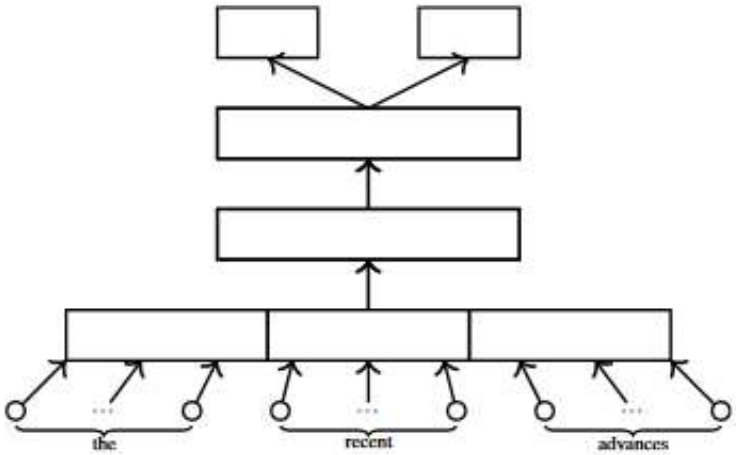


Approaches: Neural Approaches

- Weakness
  - 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

Approaches: Neural Approaches

- Weakness
  - 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
- Solution (Alikaniotis et al., 2016)
  - Train a **task-specific** word embeddingfs by augmenting the CW model
  - These **score-specific** word embeddings (**SSWEs),** are then used as features for training a neural AES model

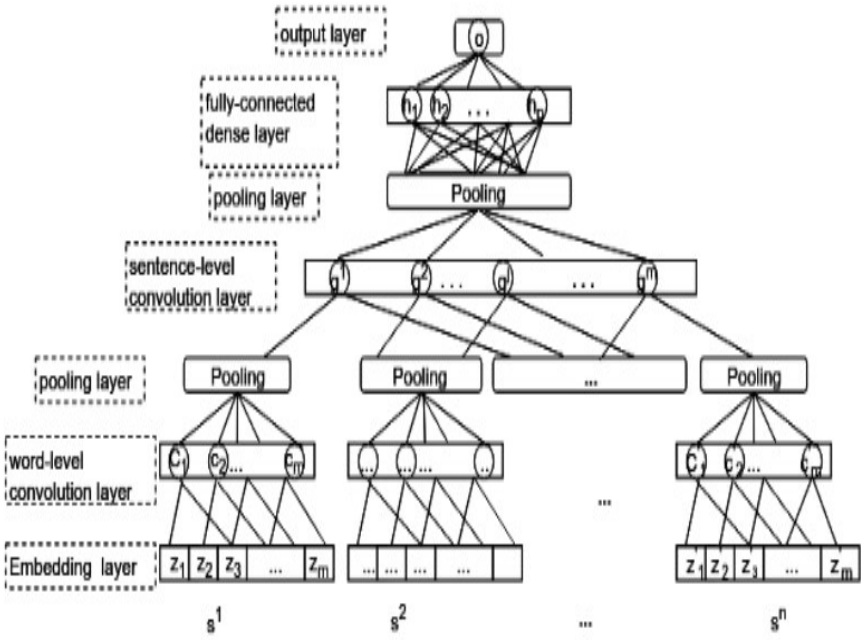


Approaches: Neural Approaches

- Weakness
  - Aforementioned approaches model a document as a linear sequence of words
  - But in reality, document is in **hierarchical** structure

Approaches: Neural Approaches

- Weakness
  - Aforementioned approaches model a document as a linear sequence of words
  - But in reality, document is in **hierarchical** structure
- Solution (Dong and Zhang, 2016)
  - Model the **hierarchical** structure by using two convolution layers that correspond to the two level hierarchical structure (i.e. sentence level and word-level)

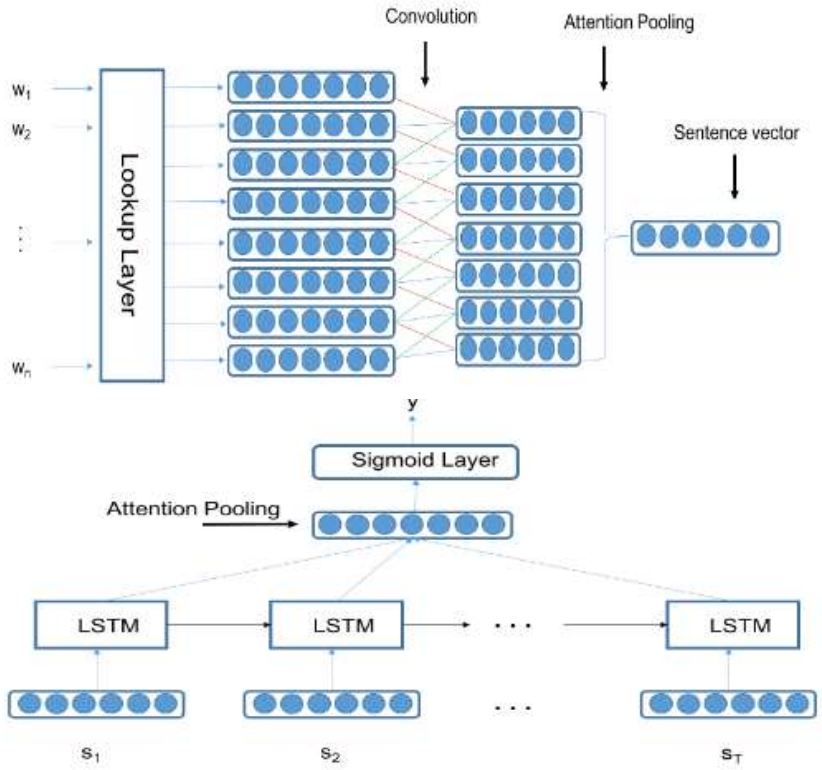


Approaches: Neural Approaches

- Weakness
  - 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

Approaches: Neural Approaches

- Weakness
  - 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
- Solution (Dong et al., 2017)
  - Incorporate an attention mechanism into the neural network by using attention pooling rather then simple pooling

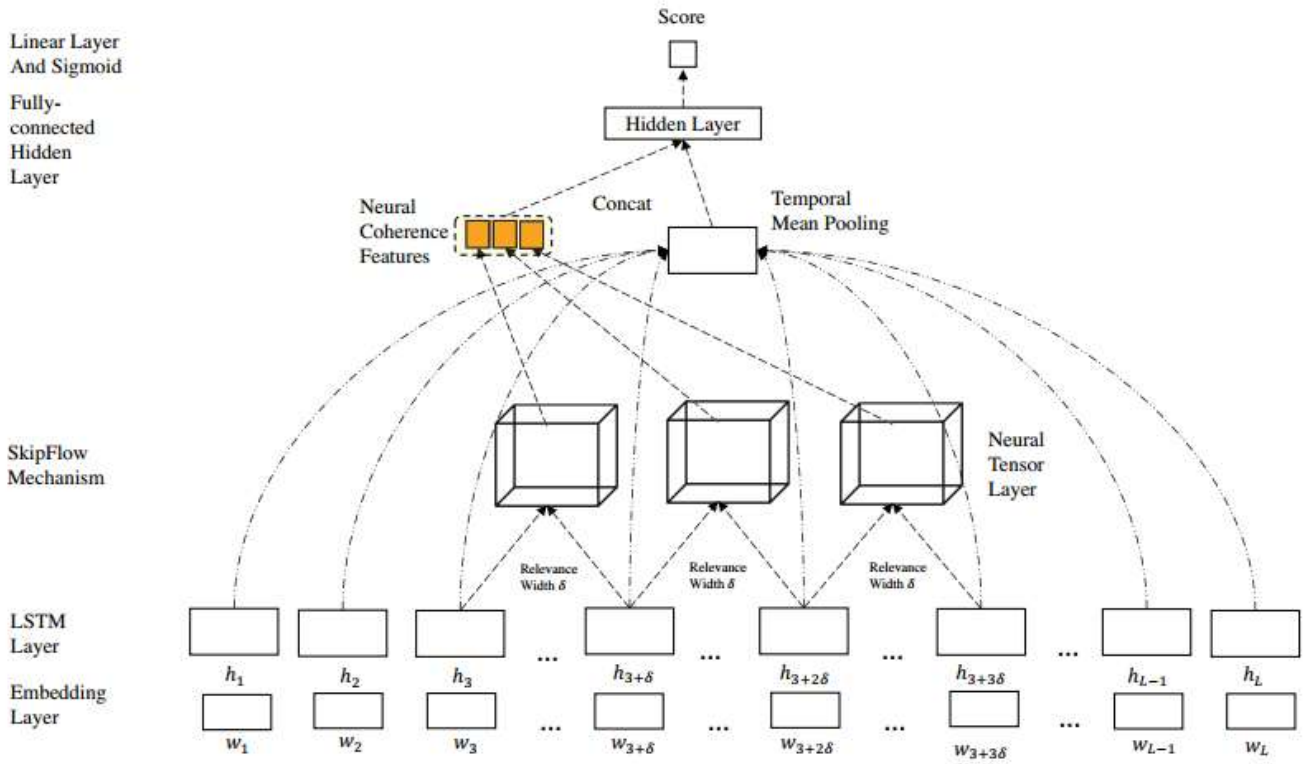


Approaches: Modeling Coherence

- Motivation (Tay et al., 2018)
  - Coherence is an important dimension of essay quality
  - Holistic scoring can also be improved by computing and exploiting the coherence score of an essay

Approaches: Neural Approaches

- Motivation
  - Coherence is an important dimension of essay quality
  - Holistic scoring can also be improved by computing and exploiting the coherence score of an essay



Approaches: Transfer Learning

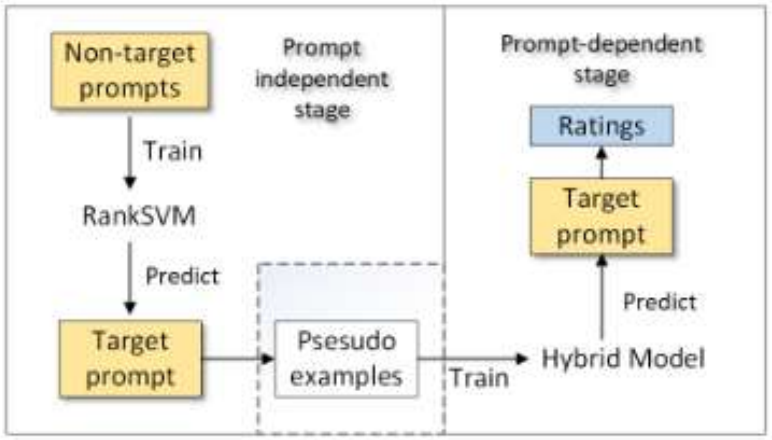
- Motivation (Phai et a;., 2015; Cummins et al., 2016; Jin et al., 2018)
  - Ideally, we can train **prompt-specific** AES systems
  - In practice however, it is **rarely** the case that enough essays for the target prompt are available for training.

Approaches: Transfer Learning

- Motivation (Phai et a;., 2015; Cummins et al., 2016; Jin et al., 2018)
  - Ideally, we can train **prompt-specific** AES systems
  - 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**
  - In that case, the potential **mismatch** in the vocabulary used in the essays written for the source prompt and those for the target prompt may hurt the performance of prompt-independent systems

Approaches: Transfer Learning

- Motivation (Phai et a;., 2015; Cummins et al., 2016; Jin et al., 2018)
  - Ideally, we can train **prompt-specific** AES systems
  - 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**
  - In that case, the potential **mismatch** in the vocabulary used in the essays written for the source prompt and those for the target prompt may hurt the performance of prompt-independent systems



Features

- A large amount of work on AES has involved feature development
  - The amount of training data is limited (which is important for neural models to be effective)
  - Neural model may further improved by incorporating hand-crafted features obtained via feature engineering
  - Feature-based approaches and neural approaches should be viewed as complementary rather than competing approached

Features: Length-based

- Length within certain range is found to be highly positively correlated with the holistic score of an essay
- Commonly used features
  - Number of Sentences
  - Number of Words
  - Number of Characters

Features: Lexical

- Divided into two categories
  - Word unigrams, bigrams and trigrams that appear in an essay.
  - Statics computed based on word n-grams, particularly bi-gram
- These word n-grams are useful because they encode the grammatical, semantic, and discourse information about an essay that could be useful for AES
  - the bigram ”people is” suggests ungrammaticality
  - the use of discourse connectives (e.g., ”moreover", "however”) suggest cohesion

Features: Embeddings

- Embeddings can be seen as a variant of n-ram features, are arguably a better representation of the semantics of a word/phrase than word n-grams
- Three types of Embedding features
  - Features computed based on embeddings **pertained** on a large corpus such as GLoVE
  - **AES-specific** embeddings
  - **Originally** one-hot word vectors, but are being updated as he neural model that uses these feature is trained

Features: Word category

- Computed based on wordlist 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
- 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

Features: Prompt-relevant

- 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
- Common measures of similarity
  - Number of word overlap
  - Word topicality
  - Semantic similarity as measured by random indexing

Features: Readability

- Encode how difficult an essay is to read
  - While good essays should not be overly difficult to read, they should not be too essay to read either
- Common measures of readability in AES
  - Flesch-kindcaid Reading Ease
  - Type-token ration

Features: Syntactic

- Encode the syntactic information about an essay
- Three main types of syntactic features
  - Part-of-speech
  - Parse Tree
  - Grammatical error rates

Features: Argumentation

- Computed based on the argumentative structure
  - Only applicable to persuasive essay
  - Have often been used to predict the persuasiveness of an argument made in an essay
- Argumentative structure
  - Major claim
  - Claim
  - premise
- Computed based on the argument component and relations

Features: Semantic

- Encode the lexical semantic relations between different words in an essay
- Two main types of semantic features
  - Histogram-based features
  - Frame-based features
- Computed based on the argument component and relations

Features: Discourse

- Encode the discourse structure of an essay
- This feature have been derived from
  - Entity grid
  - Rhetorical structure theory (RST) trees
  - Lexical Chain
  - Discourse function labels

Plan for the talk

- Corpora
- Systems
- Evaluation and State of the Art
- Concluding Remarks

Evaluation and State of the Art

Metrics

- Quadratic Weighted Kappa (QWK)
- Mean Absolute Error (MAE)
- Pearson's Correlation Coefficient (PCC)

Evaluation and State of the Art

- Holistic Scoring
  - Both QWK and PCC are quite high (on CLC-FCE, ASAP and TOEFL11)
- Dimension-specific Scoring
  - Worse than their holistic counterparts in terms of PCC (on ICLE and AAE)

Evaluation and State of the Art

- Nevertheless, these results do not necessary suggest that holistic scoring is easier than domain-specific scoring
  - They are not directly comparable as they are obtained on different corpora
  - The number of essay used to train the holistic scoring tend to be larger than those used to train the dimension-specific scores.
  - What these results do suggest, however, is that dimension-specific is far from being solved

Plan for the talk

- Corpora
- Systems
- Evaluation and State of the Art
- Concluding Remarks

Concluding Remarks