AUTOMATED ESSAY SCORING:

ARGUMENT PERSUASIVENESS

by

Zixuan Ke

APPROVED BY SUPERVISORY COMMITTEE:

Vincent Ng, Chair

Nicholas Ruozzi

Sriraam Natarajan

This Master's thesis

is dedicated to my advisor and lab mates,

who help me a lot

in both academia and life.

AUTOMATED ESSAY SCORING:
ARGUMENT PERSUASIVENESS

by

ZIXUAN KE, BS

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER IN

COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

May 2019

ACKNOWLEDGMENTS

The author thanks Dr. Vincent Ng, Jing Lu, Gerardo Ocampo Diaz, Mingqing Ye, Hui Lin and Yize Pang for their time giving me extremely useful advice. He also wants to thank Dr. Nicholas Ruozzi and Dr. Sriraam Natarajan for their time serving on his thesis committee.

April 2019

AUTOMATED ESSAY SCORING:

ARGUMENT PERSUASIVENESS

Zixuan Ke, MS
The University of Texas at Dallas, 2019

Supervising Professor: Vincent Ng, Chair

Despite being investigated for over 50 years, the task of Automated Essay Scoring (AES) is far from being solved. Nevertheless, it 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 research challenges it brings about. While the majority of work on AES has focused on evaluating an essay's overall quality, comparatively less work has focused on scoring an essay along specific dimensions of quality. Among many dimensions, argument persuasiveness is arguably the most important but largely ignored dimension.

In this thesis, we focus on argument persuasiveness scoring. First, we present our publicly available corpus for argument persuasiveness. Our corpus is the first corpus of essays that are simultaneously annotated with argument components, argument persuasiveness scores, and attributes of argument components that impact an argument's persuasiveness. The inter-annotator agreement and the oracle experiments indicate that our annotations are reliable and the attributes can help predict the persuasiveness score. Second, we present the first set of neural models that predict the persuasiveness of an argument and its attributes in a student essay, enabling useful feedback to be provided to students on *why* their arguments are (un)persuasive in addition to *how* persuasive they are. The evaluation of our models shows that automatically computed attributes are useful for persuasiveness scoring and that performance can be improved by improving attribute prediction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Automated Essay Scoring (AES), the task of employing computer technology to evaluate and 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 (PEG) 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 training and evaluation of holistic essay scoring engines. Second, holistic scoring technologies have large commercial values: being able to successfully automate the scoring of the millions of essays written for aptitude tests such as TOEFL, SAT, GRE, and GMAT every year can save a lot of manual grading effort.

Though useful for scoring essays written for aptitude tests, holistic essay 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 an essay provides essentially no feedback to its author 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 (Somasundaran et al., 2014; Higgins et al., 2004), technical errors, and relevance to prompt (Persing and Ng, 2014; Louis and Higgins, 2010). 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 1.1 enumerates the aspects of an essay that could impact its overall quality and hence 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 1.1: Different dimensions of essay quality

Dimension	**Description**
Grammar	Correct use of grammar; correct spelling
Usage	Correct use of preposition & punctuation
Mechanics	Correct use of capitalization
Style	Variety in vocabulary & sentence structure
Relevance	Content addresses the prompt given
Organization	Essay is well-structured
Development	Proper development of ideas w/ examples
Cohesion	Appropriate use of transition phrases
Coherence	Appropriate transitions between ideas
Thesis Clarity	Presents a clear thesis
Persuasiveness	The argument for its thesis is convincing

Table 1.1 are listed roughly in increasing difficulty of the corresponding scoring task. For instance, grammaticality and mechanics are low-level NLP tasks that have 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. 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.

Among these dimensions of essay quality, argument persuasiveness is largely ignored in existing automated essay scoring research despite being one of the most important dimensions of essay quality. Nevertheless, scoring the persuasiveness of arguments in student essays is by no means easy. The difficulty stems in part from the scarcity of persuasiveness-annotated corpora of student essays. While persuasiveness-annotated corpora exist for other domains such as online debates (e.g., Habernal and Gurevych (2016a; 2016b)), to our knowledge only one corpus of persuasiveness-annotated student essays has been made publicly available so far (Persing and Ng, 2015).

Though a valuable resource, Persing and Ng's (2015) (P&N) corpus has several weaknesses that limit its impact on automated essay scoring research. First, P&N assign only *one* persuasive-

ness score to each essay that indicates the persuasiveness of the argument an essay makes for its *thesis*. However, multiple arguments are typically made in a persuasive essay. Specifically, the arguments of an essay are typically structured as an argument tree, where the major claim, which is situated at the root of the tree, is supported by one or more claims (the children of the root node), each of which is in turn supported by one or more premises. Hence, each node and its children constitute an argument. In P&N's dataset, only the persuasiveness of the overall argument (i.e., the argument represented at the root and its children) of each essay is scored. Hence, any system trained on their dataset cannot provide any feedback to students on the persuasiveness of any arguments other than the overall argument. Second, P&N's corpus does not contain annotations that explain *why* the overall argument is not persuasive if its score is low. This is undesirable from a feedback perspective, as a student will not understand why her argument is not persuasive if its score is low.

To address the aforementioned weakness, we annotate and make publicly available a corpus of persuasive student essay. We further develop computational models for scoring argument persuasiveness and providing feedback on the argument persuasiveness score. Our main contributions in this thesis are two-fold. First, we present the first corpus of 102 persuasive student essays that are simultaneously annotated with argument trees, persuasiveness scores, and attributes of argument components that impact these scores. We believe that this corpus will push the frontiers of research in content-based essay scoring by triggering the development of novel computational models concerning argument persuasiveness that could provide useful feedback to students on why their arguments are (un)persuasive in addition to how persuasive they are. Second, we design the first set of neural models for predicting the persuasiveness of an argument and its attributes in a student essay.

CHAPTER 2

RELATED WORK

Across the rich history of AES research, many works have focused on either *holistic scoring* or *dimension-specific scoring*. In this chapter, we detail above two scoring tasks in order to provide a comprehensive overview on AES to readers.

2.1 Holistic Scoring

As mentioned above, many works on AES have focused on *holistic scoring*. Here, we characterize the existing work along three different dimensions: *corpora* it used for training and evaluating, *approaches* it was developed, as well as *features* it employed.

2.1.1 Corpora

Table 2.1: Comparison of several popularly used corpora for holistic AES

Corpora	Essay Types	Writer's Language Level	Statistic	Score Range	Additional Annotation
CLC-FCE	ARG, LET, NAR COM, SEG	Non-native; ESOL Test Taker	1244 essays across 10 prompts	1-40	Linguistic errors (∼80 error types)
ASAP	ARG, RES, NAR	US students; Grade 7 to 10	17450 essays across 8 prompts	can be as small as [0-3] and as large as [0-60]	none
TOEFL11	ARG	Non-native; TOEFL Test Taker	1100 essays across 8 prompts	Low, Medium, High	none

In this section, we present three corpora that have been widely used for training and evaluating AES systems. Table 2.1 compares these corpora along five dimensions: (1) the types of essays present in the corpus: argumentative (ARG), response (RES), narrative (NAR), comment (COM), suggestion (SEG) and letter (LET); (2) the language level of the essay's writer; (3) the number of essays and the number of prompts in the corpus; (4) the score range; and (5) additional annotation on the corpus (if any).

The Cambridge Learner Corpus-First Certificate in English exam (CLC-FCE) is the earliest publicly-available corpus for training and evaluating AES systems (Yannakoudakis et al., 2011).

It provides for each essay both its holistic score and the manually tagged linguistic errors types it contains (e.g., incorrect tense), which makes it possible to build systems not only for scoring but also for grammatical error detection and correction. However, the rather small number of essays per prompt makes it hard to build *prompt-specific* systems (i.e., systems that are trained and tested on the same prompt).

The Automated Student Assessment Prize (Hewlett Foundation, 2012), 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 not only in terms of the total number of essays, but also the number of essays per prompt (up to 3000 essays per prompt). This makes it possible to build prompt-specific systems. However, there are at least three factors that limit its usefulness: (1) the score ranges are different for different prompts, which makes it difficult to train a model on multiple prompts; (2) its essays do not contain any paragraph information; and (3) its aggressive preprocessing expunged both name entities and most other capitalized words, so the essays may not be ”true to the original”.

The TOEFL11 corpus (Blanchard et al., 2013) , which contains essays from a real high-stakes exam (TOEFL), is originally compiled for the Native Language Identification (NLI) task but 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 ”score”/quality of the essays and attempted to train AES systems on them. The corpus is well-balanced across prompts and languages (11 native languages). However, its usefulness is limited by the fact that the label of an essay (i.e., the associated proficiency level) is not necessarily the proficiency of the essay but rather the language proficiency of the writer.

All the corpora shown in Table 2.1 are publicly available. They are all in English. AES corpora in other languages exist, such as Ostling's (2013) Swedish corpus and Horbach *et al.*'s (2017) German corpus.

2.1.2 Approaches

Supervised Learning Approaches Almost all existing AES systems are supervised, recasting the scoring task (1) as a *regression* task, where the goal is to predict the score of an essay; (2) as 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, high as in the aforementioned TOEFL11 corpus); or (3) as a *ranking* task, where the goal is to rank two (i.e., pairwise ranking) or more essays based on their scores.

Off-the-shelf learning algorithms are typically used for model training. For regression, linear regression (Page, 1966, 1994; Landauer et al., 2003; Elliot, 1999; Attali and Burstein, 2006; Miltsakaki and Kukich, 2004; Klebanov et al., 2013; Crossley et al., 2015; Faulkner, 2014; Klebanov et al., 2016), support vector regression (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 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 work on AES has focused on feature engineering (see the next subsection for a detailed discussion on features for AES), a 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 the essay as input, their model first uses a *convolution* layer to extract n-gram level features. These features capture the *local* textual dependencies (among the words in an n-gram). These features are then passed as input to a *recurrent* layer (i.e., a Long-Short Term Memory) (LSTM) network (Hochreiter and Schmidhuber, 1997), which outputs a vector at each time step that captures *long-distance* dependencies of the

words in the essay. The vectors from different timesteps 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 models for AES are extensions of T&N's model, they all address one or more of its weaknesses, as described below:

Learning score-specific word embeddings. SSWEs are motivated by the observation that some words are *under*-informative in that they have little power in discriminating high- and low-scoring essays. To address 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. Though word embeddings can be trained on a large, unannotated corpus using the CW model (Collobert and Weston, 2008), a word embedding learning neural network architecture, 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 informative from under-informative words, are then used as input 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 (1) combining its words to form each of its sentences and then (2) combining the resulting sentences to form the document. Consequently, their model uses two convolution layers that corresponds 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 input sentences. After passing through a pooling layer, these 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. Some characters, words, and sentences in an essay are more important than others as far as scoring is concerned. However, Dong and Zhang's (2016) two convolution layer neural network described above does not distinguish the importance between different words. To identify important characters, words, and sentences, they employ an *attention* mechanism. Rather than using *simple pooling* (such as max or average) after each layer, Dong et al. (2017) use *attention pooling*. 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 the coherence score of an essay, since coherence is an important dimension of essay quality. They model coherence as follows. Like T&N, their neural network employs an LSTM. 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 timesteps and computes the similarity for each such pair of positional outputs. They call these similarities *neural coherence features*. The reason is that intuitively, coherence should correlate positively with high 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 scoring.

Transfer learning. Ideally, we can train prompt-specific AES systems, in which the training (i.e., source) prompt and the test (i.e., target) prompt are the same. 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 larger set of non-target-prompt essays are used for training. However, the potential mismatch in the vocabulary typically 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 (Daume III, 2007), one of the very well-known transfer learning algorithms, assumes as input training data from only two domains (source and target), and the goal is to learn a model that can perform well when classifying the (target domain) test instances. To understand EasyAdapt, recall that a model that does *not* use transfer learning is typically trained by employing a feature space shared by both the source and target domain instances. EasyAdapt augments this feature set by duplicating each feature in the space three times, where the first copy stores information shared by both domains, the second copy stores source-domain information, and the last copy stores 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 target-domain information. Phandi et al. (2015) generalize EasyAdapt to Correlated Bayesian Linear Ridge Regression (CBLRR), enabling the weight given to the target-domain 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 in the same domain (i.e., prompt).

While the above systems assume that a small number of target-domain essays is available for training, Jin et al. (2018) perform transfer learning under the assumption that *no* target-domain

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 *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 features (i.e., prompt-independent features). Stage 2 aims to score the remaining (presumably target-domain essays with non-extreme quality) in the test set. To do so, they first automatically label each low-quality essay and each high-quality essay identified in Stage 1 as 0 and 1, respectively. They then train a regressor on these automatically-labeled essays using *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 essays, whose scores are expected to fall between 0 and 1 given their non-extreme quality.

Reinforcement Learning Approaches Wang et al. (2018) propose the first reinforcement learning (RL) framework for AES. The use of RL for AES is primarily motivated by the metric typically used to evaluate AES systems, quadratic weighted kappa (QWK)[1]. QWK is an agreement metric that ranges from 0 to 1 but can be negative if there is less agreement than expected by chance. Here we note two key properties of QWK that are relevant to the current discussion. First, QWK is not differentiable, so it is hard to optimize it directly (in supervised systems). Second, QWK emphasizes the overall rating schema while the aforementioned supervised systems are trained to score one essay at a time without considering the overall rating schema. The RL framework proposed by Wang et al. seeks to address both of these issues. We refer the reader to their paper for details.

[1]See Hewlett Foundation (2012) for details.

2.1.3 Features

In this section, we give an overview of 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 score of an essay. These features encode the length of an essay in terms of the number of sentences, words, and/or characters in the essay.

Lexical features comprise two categories. One category contains word unigrams, bigrams, and trigrams that appear in an essay. These word n-grams are useful because they encode 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 or important to mention for a particular prompt. The key advantages of n-grams are 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 a particular punctuation in an essay (Page, 1966; Chen and He, 2013; Phandi et al., 2015; Zesch et al., 2015).

Embeddings can be seen as a variant of n-gram features. As mentioned before, embeddings 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 (BOSWE). 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 score-specific word embeddings (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 that uses these features is trained (Taghipour and Ng, 2016; Dong and Zhang, 2016; Jin et al., 2018; Tay et al., 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 that encode whether certain categories of words are present in an essay (McNamara et al., 2015; Cummins et al., 2016; Yannakoudakis and Briscoe, 2012; Amorim et al., 2018; Farra et al., 2015; Chen and He, 2013), as they reflect 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 word levels (see Breland et al. (1994)) that a word belongs to have also been used. Intuitively, the higher a word's level is, the more sophisticated vocabulary usage it indicates. Word category features help generalize n-grams 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).

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: good essays should demonstrate a broad vocabulary and a variety of sentence structures. Readability is typically measured using readability metrics (e.g., 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 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 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 relation distance features encode the distance between a head and its dependent in a grammatical relation. *Grammatical error rates* are used to derive features that encode how frequent grammatical errors appear in an essay, and is computed either using a language model or from hand-annotated grammatical error types (Yannakoudakis et al., 2011; Yannakoudakis and Briscoe, 2012).

Argumentation features are features 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. 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 w.r.t. the essay's topic). The major claim is supported/attacked by one or more claims (controversial statements that should not be readily accepted by the reader without further evidence), each of which is in turn supported/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/premises in a paragraph) and the argument tree (e.g., the tree depth) (Nguyen and Litman, 2018; Ghosh et al., 2016).

Semantic features encode the lexical semantic relation 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 more likely to indicate a more creative use of language. *Frame-based features* are computed based on semantic frames in FrameNet (Baker et al., 1998). Briefly, a frame may describe an event that occurs in a sentence, and the event's frame elements may be the people or objects that participate in the 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. One of the frame elements participating in the frame is the Speaker "they". 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).

Discourse features encode the discourse structure of an essay. Broadly, there are four types of discourse features. *Entity grids* 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). *Discourse parse trees* constructed based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) are used to capture the discourse structure of text (e.g., is one discourse segment a elaboration of the other, or is it in a contrast relation with the other?). These RST trees can be used to capture the local and global coherence of an essay. *Lexical chains*, which are sequences of related words in a document, have been used as an indicator of text cohesion. Intuitively, an essay that contains many lexical chains, especially ones where the beginning and end of the chain cover a large span of an essay, tend to be more cohesive (Somasundaran et al., 2014). *Discourse function labels* have also been used to compute features (Persing et al., 2010). A discourse function label is defined on a sentence/paragraph that indicates its discourse function (e.g., whether the paragraph is an introduction or a conclusion, whether a sentence is the thesis of the essay). Persing et al. (2010) have used n-grams computed based on sentence- and paragraph-based discourse function labels when scoring the organization of an essay.

In order to obtain deeper insight on the usefulness of these features, we consider the feature set used in the state-of-the-art model on above three corpora (i.e., CLC-FEC, ASAP and TOEFL11).

Cozma et al. (2018) propose the state-of-the-art model on ASAP dataset. They use *embeddings* (BOSWE) as the feature to obtain the best result. Note that their result outperforms those methods based on handcrafted features (Phandi et al., 2015) as well as deep features (Dong and Zhang, 2016; Tay et al., 2018; Dong et al., 2017). This feature is powerful based on the fact that word embeddings carry semantic information by projecting semantically related words in the same region of the embedding state. The clustering and reassigning operation in BOSWE make the semantic pattern of a document more obvious and easier to be recognized.

Yannakoudakis and Briscoe (2012) propose the state-of-the-art model on CLC-FEC dataset. The features they used include *length-based* (word length), *lexical* (BOW), *semantic* (semantic similarity), *syntactic* (POS), *word category* (discourse connectives) and *discourse* (entity grids) features. Their result outperforms the best reported by Yannakoudakis et al. (2011). The main difference is that features in Yannakoudakis et al. (2011) only involve *lexical* and *grammar* while features in Yannakoudakis and Briscoe (2012) further assume that adding a coherence metric to the feature set of the holistic scoring AES system would improve the performance, thus many features involved coherence are added (e.g., entity grids).

Vajjala (2018) proposes the state-of-the-art model on TOEFL11 dataset. The features they used include *length-based* (document length) , *syntactic* (POS, syntactic parse tree), *word-category* (spelling error) and *prompt-relevant* (prompt identifier), *readability* (type-token ration) and *discourse* (entity grids) features. This feature set contains a broad range of features and they are developed following two directions: (1) different linguistic aspects related to written language (e.g., lexical, discourse, syntactic and semantic features); and (2) knowledge from second language acquisition about learner writing (e.g., lexical richness, syntactic complexity and readability assessment). While Vajjala (2018) obtains the state-of-the-art performance on *accuracy*, Nguyen and Litman (2018) obtain the state-of-the-art performance on another metric (i.e., QWK) by using

a set of *argumentation* features. It is not surprising since all essays in TOEFL11 dataset are *argumentative* essays and the *holistic* score itself is intuitively argument-related: it is less likely for an essay to achieve high *holistic* score with bad arguments.

2.2 Dimension-specific Scoring

While *holistic* scoring has a long history which can date back to 1960s, *dimension-specific* scoring did not start until 2004. So far, several dimensions of quality, including coherence (Burstein et al., 2010; Somasundaran et al., 2014), organization (Persing et al., 2010; Wachsmuth et al., 2016), relevance to prompt (Louis and Higgins, 2010; Persing and Ng, 2014; Wachsmuth et al., 2016), thesis clarity (Persing and Ng, 2013; Wachsmuth et al., 2016), and argument persuasiveness (Persing and Ng, 2015; Wachsmuth et al., 2016) have been examined.

2.2.1 Corpora

International Corpus of Learner English (ICLE) (Granger et al., 2009) is a popular corpus used in dimension-specific socring. This corpus is originally designed for Native Language Identification task and contains 6085 essays written by undergraduate university students who were non-native English speakers. One issue that hinders the progress on *dimension-specific* essay scoring concerns the scarcity of publicly-available corpora annotated with dimension-specific scores. With the goal of performing dimension-specific scoring, Persing and his colleagues (2010; 2013; 2014; 2015) annotated a subset of essays in the ICLE corpus. They annotated the subset (all essays in the subset are *argumentative* essays) along several dimensions of essay quality, including (1) Organization, which refers to how well-organized an essay is (the subset used for organization annotation contains 1003 essays across 12 prompts); (2) Thesis Clarity, which refers to how clearly an author explains the thesis of her essays (the subset used for Thesis Clarity annotation contains 830 essays across 13 prompts); (3) Prompt Adherence, which refers to how related an essay's content is to the prompt for which it was written (the subset used for Prompt Adherence annotation contains 830

essays across 13 prompts); and (4) Argument Persuasiveness, refers to the persuasiveness of the argument an essay makes for its thesis (the subset used for Argument Persuasiveness annotation contains 1000 essays across 10 prompts).

We note that all these 4 subsets of ICLE are annotated in a range 1-4 (0.5 increments) and only the dimension scores are annotated, which means there is no additional annotation for the use of feedback. Another thing deserved to mention is that while the three corpora introduced in holistic scoring section are all publicly available, the ICLE corpus is proprietary.

2.2.2 Approaches

Supervised learning approaches are extensively used in *dimension-specific* scoring. Specifically, *off-the-shelf* learning algorithms are typically used for the model training, including *support vector regression* (Persing et al., 2010; Persing and Ng, 2013, 2014, 2015, 2016), decision tree (Burstein et al., 2010) and SMO (Wachsmuth et al., 2016).

2.2.3 Features

In this section, we summarize features used in different dimensions of essay quality (i.e., *Coherence*, *Argument Persuasiveness*, *Prompt Adherence*, *Thesis Clarity* and *Organization*)

Coherence refers to the reader's ability to construct meaning from a document. Intuitively, it is greatly influenced by the presence and organization of cohesive elements. There is no publicly available corpora for coherence thus there is no state-of-the-art model. Nevertheless, we can still note that many *discourse* features are used in coherence scoring, such as *RST*, *entity grids* and *lexical-chains* (Somasundaran et al., 2014; Higgins et al., 2004).

Argument persuasiveness is arguably one of the most important dimension of essay quality (note that Argument Persuasiveness only valid in persuasive essay). *Argumentation* features are extensively used for argument persuasiveness scoring (Persing and Ng, 2015; Wachsmuth et al., 2016). By making use of the information of argumentation structure in both *sentence* level

and *paragraph* level, Wachsmuth et al. (2016) obtain the state-of-the-art performance on ICLE dataset. Note that it outperforms Persing and Ng (2015). The main difference is that the features in Wachsmuth et al. (2016) highly emphasize on argumentation structure (e.g. the argument flow among paragraphs) while Persing and Ng (2015) only implement some simple heuristic features for argumentation structure (e.g., number of major claims and number of paragraphs without argument component).

Prompt adherence refers to how related an essay's content is to prompt for which it was written. *Word category* (words in the prompt that the response will likely mention), *Semantic* (semantic similarity as measured by random indexing (Higgins et al., 2004)) are actively used as features for prompt adherence (Persing and Ng, 2014). On ICLE dataset, Persing and Ng (2014) outperform Wachsmuth et al. (2016) (remind that the features in Wachsmuth et al. (2016) focus on extracting the information of *argumentation* structure). The result is not surprising since argumentation structure has low impact on whether an essay is off-topic.

Thesis clarity refers to how clearly an author explains the thesis of her essay. *Syntactic* (POS) and *lexical* (BOW) and *Frame-based* features are deployed. Similar to *Prompt adherence*, Persing and Ng (2013) outperform Wachsmuth et al. (2016) on ICLE dateset. This is again not surprising since the features in Wachsmuth et al. (2016) focus on *argumentation* structure and the impact of *argumentative structure* is low for thesis clarity while the *argumentation* features proposed by Wachsmuth et al. (2016) captured only the structural information of an argument.

Organization refers to the structure of an essay. *Discourse* feature (discourse function labels) is used since organization itself is highly related to the discourse elements (Persing et al., 2010). Intuitively, *organization* also has a strong relation to *argumentation* structure. With the same feature set as in *argument persuasiveness* scoring, Wachsmuth et al. (2016) obtain the state-of-the-art performance on ICLE dataset using *argumentation* features. Note that Wachsmuth et al. (2016) outperform Persing et al. (2010). The main difference is that the feature set in Persing et al. (2010) is some simple heuristic discourse labels (e.g., thesis, conclusion and introduction) while the feature set in Wachsmuth et al. (2016) is a set of rich *argumentation* features.

CHAPTER 3

CORPUS AND ANNOTATION

One of our main contributions is annotating and making publicly available a corpus of persuasive student essays in which we (1) annotate the persuasiveness of each argument; (2) identify a set of attributes that can explain an argument's persuasiveness; and (3) annotate each argument with the values of these attributes. In this chapter, we present our corpus and annotation in detail.

3.1 Corpus

The corpus we chose to annotate is composed of 102 essays randomly chosen from the Argument Annotated Essays corpus (Stab and Gurevych, 2014). This collection of essays was taken from *essayforum* (Essayforum, 2012), a site offering feedback to students wishing to improve their ability to write persuasive essays for tests. Each essay is written in response to a topic such as "should high school make music lessons compulsory?" and has already been annotated by Stab and Gurevych with an argument tree. Hence, rather than annotate everything from scratch, we annotate the persuasiveness score of each argument in the already-annotated argument trees in this essay collection as well as the attributes that potentially impact persuasiveness.

Each argument tree is composed of three types of tree nodes that correspond to argument components. The three annotated argument component types include: **MajorClaim**, which expresses the author's stance with respect to the essay's topic; **Claims**, which are controversial statements that should not be accepted by readers without additional support; and **Premises**, which are reasons authors give to persuade readers about the truth of another argument component statement. The two relation types include: **Support**, which indicates that one argument component supports another, and **Attack**, which indicates that one argument component attacks another.

Each argument tree has three to four levels. The root is a major claim. Each node in the second level is a claim that supports or attacks its parent (i.e., the major claim). Each node is the third

Table 3.1: Corpus statistics

Essays: 102	Sentences: 1462	Tokens: 24518
Major Claims: 185	Claims: 567	Premises: 707
Support Relations: 3615	Attack Relations: 219	

level is a premise that supports or attacks its parent (i.e., a claim). There is an optional fourth level consisting of nodes that correspond to premises. Each of these premises either supports or attacks its (premise) parent. Stab and Gurevych (2014) report high inter-annotator agreement on these annotations: for the annotations of major claims, claims, and premises, the Krippendorff's α values (Krippendorff, 1980) are 0.77, 0.70, and 0.76 respectively, and for the annotations of support and attack relations, the α values are both 0.81.

Note that Stab and Gurevych (2014) determine premises and claims by their position in the argument tree and not by their semantic meaning. Due to the difficulty of treating an opinion as a non-negotiable unit of evidence, we convert all subjective premises into claims to demonstrate that they are subjective and require backing. At the end of this process, several essays contain argument trees that violate the scheme used by Stab and Gurevych, due to some premises supported by opinion premises, now converted to claims. Although the ideal argument should not violate the canonical structure, students attempting to improve their persuasive writing skills may not understand this, and mistakenly support evidence with their own opinions.

Statistics of this corpus are shown in Table 3.1. Its extensive use in argument mining research in recent years together with its reliably annotated argument trees makes it an ideal corpus to use for our annotation task.

3.2 Annotation

Once we set up the corpus, we can annotate the persuasiveness score as well as the attributes which impact the persuasiveness. In this section, we will first introduce our annotation schemes and scoring rubrics, then the annotation procedure and analysis.

3.2.1 Definition

Since persuasiveness is defined on an argument, in order to annotate persuasiveness we need to define precisely what an argument is. Following van Eemeren et al. (2014), we define an argument as consisting of a conclusion that may or may not be supported/attacked by a set of evidences. Given an argument tree, a non-leaf node can be interpreted as a "conclusion" that is supported or attacked by its children, which can therefore be interpreted as "evidences" for the conclusion. In contrast, a leaf node can be interpreted as an unsupported conclusion. Hence, for the purposes of our work, an argument is composed of a node in an argument tree and all of its children, if any.

Table 3.2: Description of the Persuasiveness scores

Score	Description
6	A very strong, clear argument. It would persuade most readers and is devoid of errors that might detract from its strength or make it difficult to understand.
5	A strong, pretty clear argument. It would persuade most readers, but may contain some minor errors that detract from its strength or understandability.
4	A decent, fairly clear argument. It could persuade some readers, but contains errors that detract from its strength or understandability.
3	A poor, understandable argument. It might persuade readers who are already inclined to agree with it, but contains severe errors that detract from its strength or understandability.
2	It is unclear what the author is trying to argue or the argument is poor and just so riddled with errors as to be completely unpersuasive.
1	The author does not appear to make any argument (e.g. he may just describe some incident without explaining why it is important). It could not persuade any readers because there is nothing to be persuaded of. It may or may not contain detectable errors, but errors are moot since there is not an argument for them to interfere with.

3.2.2 Annotation Scheme

Recall that the goal of our annotation is to score each argument w.r.t. its persuasiveness (see Table 3.2 for the rubric for scoring persuasiveness) and annotate each of its components with a set of predefined attributes that could impact the argument's persuasiveness. Table 3.3 presents a summary of the attributes we annotate. The rest of this subsection describes these attributes.

Table 3.3: Summary of the attributes together with their possible values, the argument component type(s) each attribute is applicable to (**MC**: MajorClaim, **C**: Claim, **P**: Premise), and a brief description

Attribute	**Possible Values**	**Applicability**	**Description**
Specificity	1–5	MC,C,P	How detailed and specific the statement is
Eloquence	1–5	MC,C,P	How well the idea is presented
Evidence	1–6	MC,C,P	How well the supporting statements support their parent
Logos/Pathos/Ethos	yes,no	MC,C	Whether the argument uses the respective persuasive strategy
Relevance	1–6	C,P	The relevance of the statement to the parent statement
ClaimType	value,fact,policy	C	The category of what is being claimed
PremiseType	see Section 3.2.2	P	The type of Premise, e.g. statistics, definition, real example, etc.
Strength	1–6	P	How well a single statement contributes to persuasiveness

Table 3.4: Description of the Eloquence scores

Score	**Description**
5	Demonstrates mastery of English. There are no grammatical errors that distract from the meaning of the sentence. Exhibits a well thought out, flowing sentence structure that is easy to read and conveys the idea exceptionally well.
4	Demonstrates fluency in English. If there are any grammatical or syntactical errors, their affect on the meaning is negligible. Word choice suggests a broad vocabulary.
3	Demonstrates competence in English. There might be a few errors that are noticeable but forgivable, such as an incorrect verb tense or unnecessary pluralization. Demonstrates a typical vocabulary and a simple sentence structure.
2	Demonstrates poor understanding of sentence composition and/or poor vocabulary. The choice of words or grammatical errors force the reader to reread the sentence before moving on.
1	Demonstrates minimal eloquence. The sentence contains errors so severe that the sentence must be carefully analyzed to deduce its meaning.

Each component type (MajorClaim, Claim, Premise) has a distinct set of attributes. All component types have three attributes in common: Eloquence, Specificity, and Evidence. *Eloquence* is how well the author uses language to convey ideas, similar to clarity and fluency. *Specificity* refers to the narrowness of a statement's scope. Statements that are specific are more believable because they indicate an author's confidence and depth of knowledge about a subject matter. Argument assertions (major claims and claims) need not be believable on their own since that is the

Table 3.5: Description of the Evidence scores

Score	Description
6	A very strong, very persuasive argument body. There are many supporting components that have high Relevance scores. There may be a few attacking child components, but these components must be used for either concession or refuting counterarguments as opposed to making the argument indecisive or contradictory.
5	A strong, persuasive argument body. There are sufficient supporting components with respectable scores.
4	A decent, fairly persuasive argument body.
3	A poor, possibly persuasive argument body.
2	A totally unpersuasive argument body.
1	There is no argument body for the given component.

Table 3.6: Description of the Claim and MajorClaim Specificity scores

Score	Description
5	The claim summarizes the argument well and has a qualifier that indicates the extent to which the claim holds true. Claims that summarize the argument well must reference most or all of the supporting components.
4	The claim summarizes the argument very well by mentioning most or all of the supporting components, but does not have a qualifier indicating the conditions under which the claim holds true. Alternatively, the claim may moderately summarize the argument by referencing a minority of supporting components and contain qualifier.
3	The claim has a qualifier clause or references a minority of the supporting components, but not both.
2	The claim does not make an attempt to summarize the argument nor does it contain a qualifier clause.
1	Simply rephrases the major claim or is outside scope of the major claim (argument components were annotated incorrectly: major claim could be used to support claim).

job of the supporting evidence. The *Evidence* score describes how well the supporting components support the parent component. The rubrics for scoring Eloquence, Evidence, Claim/MajorClaim Specificity, and Premise Specificity are shown in Tables 3.4, 3.5, 3.6, and 3.7 respectively.

MajorClaim Since the major claim represents the overall argument of the essay, it is in this component that we annotate the persuasive strategies employed (i.e., *Ethos, Pathos* and *Logos*). These three attributes are not inherent to the text identifying the major claim but instead summarize the child components in the argument tree.

Table 3.7: Description of the Premise Specificity scores

Score	Description
5	An elaborate, very specific statement. The statement contains numerical data, or a historical example from the real world. There is (1) both a sufficient qualifier indicating the extent to which the statement holds true and an explanation of why the statement is true, or (2) at least one real world example, or (3) a sufficient description of a hypothetical situation that would evoke a mental image of the situation in the minds of most readers.
4	A more specific statement. It is characterized by either an explanation of why the statement is true, or a qualifier indicating when/to what extent the statement is true. Alternatively, it may list examples of items that do not qualify as historical events.
3	A sufficiently specific statement. It simply states a relationship or a fact with little ambiguity.
2	A broad statement. A statement with hedge words and without other redeeming factors such as explicit examples, or elaborate reasoning. Additionally, there are few adjectives or adverbs.
1	An extremely broad statement. There is no underlying explanation, qualifiers, or real-world examples.

Table 3.8: Description of the Relevance scores

Score	Description
6	Anyone can see how the support relates to the parent claim. The relationship between the two components is either explicit or extremely easy to infer. The relationship is thoroughly explained in the text because the two components contain the same words or exhibit coreference.
5	There is an implied relationship that is obvious, but it could be improved upon to remove all doubt. If the relationship is obvious, both relating components must have high Eloquence and Specificity scores.
4	The relationship is fairly clear. The relationship can be inferred from the context of the two statements. One component must have a high Eloquence and Specificity scores and the other must have lower but sufficient Eloquence and Specificity scores for the relationship to be fairly clear.
3	Somewhat related. It takes some thinking to imagine how the components relate. The parent component or the child component have low clarity scores. The two statements are about the same topic but unrelated ideas within the domain of said topic.
2	Mostly unrelated. It takes some major assumptions to relate the two components. A component may also receive this score if both components have low clarity scores.
1	Totally unrelated. Very few people could see how the two components relate to each other. The statement was annotated to show that it relates to the claim, but this was clearly in error.

Claim The claim argument component possesses all of the attributes of a major claim in addition to a *Relevance* score and a *ClaimType*. In order for an argument to be persuasive, all supporting components must be relevant to the component that they support/attack. The scoring rubric for

Table 3.9: Description of the Strength scores

Score	Description
6	A very strong premise. Not much can be improved in order to contribute better to the argument.
5	A strong premise. It contributes to the persuasiveness of the argument very well on its own.
4	A decent premise. It is a fairly strong point but lacking in one or more areas possibly affecting its perception by the audience.
3	A fairly weak premise. It is not a strong point and might only resonate with a minority of readers.
2	A totally weak statement. May only help to persuade a small number of readers.
1	The statement does not contribute at all.

Relevance is shown in Table 3.8. The ClaimType can be *value* (e.g., something is good or bad, important or not important, etc.), *fact* (e.g. something is true or false), or *policy* (claiming that some action should or should not be taken).

Premise The attributes exclusive to premises are *PremiseType* and *Strength*. To understand Strength, recall that only premises can persuade readers, but also that an argument can be composed of a premise and a set of supporting/attacking premises. In an argument of this kind, Strength refers to how well the parent premise contributes to the persuasiveness independently of the contributions from its children. The scoring rubric for Strength is shown in Table 3.9. PremiseType takes on a discrete value from one of the following: real_example, invented_instance, analogy, testimony, statistics, definition, common_knowledge, and warrant. Analogy, testimony, statistics, and definition are self-explanatory. A premise is labeled *invented_instance* when it describes a hypothetical situation, and *definition* when it provides a definition to be used elsewhere in the argument. A premise has type *warrant* when it does not fit any other type, but serves a functional purpose to explain the relationship between two entities or clarify/quantify another statement. The *real_example* premise type indicates that the statement is a historical event that actually occurred, or something that is verfiably true about the real world.

Table 3.10: Class/Score distributions by component type

Attribute	Value	MC	C	P
Specificity	1	0	80	64
	2	73	259	134
	3	72	155	238
	4	32	59	173
	5	8	14	98
Logos	Yes	181	304	
	No	4	263	
Pathos	Yes	67	59	
	No	118	508	
Ethos	Yes	16	9	
	No	169	558	
Relevance	1		1	5
	2		33	45
	3		58	59
	4		132	145
	5		97	147
	6		246	306
Evidence	1	3	246	614
	2	62	115	28
	3	57	85	12
	4	33	80	26
	5	16	35	15
	6	14	6	12
Eloquence	1	3	23	24
	2	19	106	97
	3	116	320	383
	4	42	102	154
	5	5	16	49
ClaimType	fact		368	
	value		145	
	policy		54	
PremiseType	real_example			93
	invented_instance			53
	analogy			2
	testimony			4
	statistics			15
	definition			3
	common_know.			493
	warrant			44
Persuasiveness	1	3	82	8
	2	62	278	112
	3	60	84	145
	4	28	74	249
	5	17	39	123
	6	15	10	70

Table 3.11: Krippendorff's α agreement on each attribute by component type

Attribute	MC	C	P
Persuasiveness	.739	.701	.552
Eloquence	.590	.580	.557
Specificity	.560	.530	.690
Evidence	.755	.878	.928
Relevance		.678	.555
Strength			.549
Logos	1	.842	
Pathos	.654	.637	
Ethos	1	1	
ClaimType		.589	
PremiseType			.553

3.2.3 Annotation Procedure

Our 102 essays were annotated by two native speakers of English. We first familiarized them with the rubrics and definitions and then trained them on five essays (not included in our corpus). After that, they were both asked to annotate a randomly selected set of 30 essays and discuss the resulting annotations to resolve any discrepancies. Finally, the remaining essays were partitioned into two sets, and each annotator received one set to annotate. The resulting distributions of scores/classes for persuasiveness and the attributes are shown in Table 3.10.

3.2.4 Inter-Annotator Agreement

We use Krippendorff's α to measure inter-annotator agreement. Results are shown in Table 3.11. As we can see, all attributes exhibit an agreement above 0.5, showing a correlation much more significant than random chance. Persuasiveness has an agreement of 0.688, which suggests that it can be agreed upon in a reasonably general sense. The MajorClaim components have the highest Persuasiveness agreement, and it declines as the type changes to Claim and then to Premise. This would indicate that persuasiveness is easier to articulate in a wholistic sense, but difficult to explain as the number of details involved in the explanation increases.

The agreement scores that immediately stand out are the perfect 1.0's for Logos and Ethos. The perfect Logos score is explained by the fact that every major claim was marked to use logos. Although ethos is far less common, both annotators easily recognized it. This is largely due to the indisputability of recognizing a reference to an accepted authority on a given subject. Very few authors utilize this approach, so when they do it is extremely apparent. Contrary to Persuasiveness, Evidence agreement exhibits an upward trend as the component scope narrows. Even with this pattern, the Evidence agreement is always higher than Persuasiveness agreement, which suggests that it is not the only determiner of persuasiveness.

In spite of a rubric defining how to score Eloquence, it remains one of the attributes with the lowest agreement. This indicates that it is difficult to agree on exact eloquence levels beyond basic English fluency. Additionally, Specificity produced unexpectedly low agreement in claims and major claims. Precisely quantifying how well a claim summarizes its argument turned out to be a complicated and subjective task. Relevance agreement for premises is one of the lowest, partly because there are multiple scores for high relevance, and no examples were given in the rubric.

All attributes but those with the highest agreement are plagued by inherent subjectivity, regardless of how specific the rubric is written. There are often multiple interpretations of a given sentence, sometimes due to the complexity of natural language, and sometimes due to the poor writing of the author. Naturally, this makes it difficult to identify certain attributes such as Pathos, ClaimType, and PremiseType.

Although great care was taken to make each attribute as independent of the others as possible, they are all related to each other to a minuscule degree (e.g., Eloquence and Specificity). While annotators generally agree on what makes a persuasive argument, the act of assigning blame to the persuasiveness (or lack thereof) is tainted by this overlapping of attributes.

3.2.5 Analysis of Annotations

To understand whether the attributes we annotated are indeed useful for predicting persuasiveness, we compute the Pearson's Correlation Coefficient (PC) between persuasiveness and each of the

Table 3.12: Correlation of each attribute with Persuasiveness and the corresponding p-value

Attribute	PC	p**-value**
Specificity	.5680	0
Relevance	−.0435	.163
Eloquence	.4723	0
Evidence	.2658	0
Strength	.9456	0
Logos	−.1618	0
Ethos	−.0616	.1666
Pathos	−.0835	.0605
ClaimType:fact	.0901	.1072
ClaimType:value	−.0858	.1251
ClaimType:policy	−.0212	.7046
PremiseType:real_example	.2414	0
PremiseType:invented_instance	.0829	.0276
PremiseType:analogy	.0300	.4261
PremiseType:testimony	.0269	.4746
PremiseType:statistics	.1515	0
PremiseType:definition	.0278	.4608
PremiseType:common_knowledge	−.2948	0
PremiseType:warrant	.0198	.6009

attributes along with the corresponding p-values. Results are shown in Table 3.12. Among the correlations that are statistically significant at the $p < .05$ level, we see, as expected, that Persuasiveness is positively correlated with Specificity, Evidence, Eloquence, and Strength. Neither is it surprising that support provided by a premise in the form of statistics and examples is positively correlated with Persuasiveness. While Logos and invented_instance also have significant correlations with Persuasiveness, the correlation is very weak.

Next, we conduct an oracle experiment in an attempt to understand how well these attributes, when used together, can explain the persuasiveness of an argument. Specifically, we train three linear SVM regressors (using the SVM^{light} software (Joachims, 1999) with default learning parameters except for C (the regularization parameter), which is tuned on development data using grid search) to score an argument's persuasiveness using the *gold* attributes as features. The three regressors are trained on arguments having MajorClaims, Claims, and Premises as parents. For

Table 3.13: Persuasiveness scoring using gold attributes

	MC	C	P	Avg
PC	.969	.945	.942	.952
ME	.150	.250	.251	.217

instance, to train the regressor involving MajorClaims, each instance corresponds to an argument represented by all and only those attributes involved in the major claim and all of its children.[1]

Five-fold cross-validation results, which are shown in Table 3.13, are expressed in terms of two evaluation metrics, PC and ME (the mean absolute distance between a system's prediction and the gold score). Since PC is a *correlation* metric, higher correlation implies better performance. In contrast, ME is an *error* metric, so lower scores imply better performance. As we can see, the large PC values and the relatively low ME values provide suggestive evidence that these attributes, when used in combination, can largely explain the persuasiveness of an argument.

What these results imply in practice is that models that are trained on these attributes for persuasiveness scoring could provide useful feedback to students on *why* their arguments are (un)persuasive. For instance, one can build a pipeline system for persuasiveness scoring as follows. Given an argument, this system first predicts its attributes and then scores its persuasiveness using the predicted attribute values computed in the first step. Since the persuasiveness score of an argument is computed using its predicted attributes, these attributes can explain the persuasiveness score. Hence, a student can figure out which aspect of persuasiveness needs improvements by examining the values of the predicted attributes.

Table 3.14: An example essay. Owing to space limitations, only its first two paragraphs are shown

Prompt: Government budget focus, young children or university?

Education plays a significant role in a country's long-lasting prosperity. It is no wonder that governments throughout the world lay special emphasis on education development. As for the two integral components within the system, elementary and advanced education, there's no doubt that a government is supposed to offer sufficient financial support for both.

Concerning that elementary education is the fundamental requirement to be a qualified citizen in today's society, government should guarantee that all people have equal and convenient access to it. So a lack of well-established primary education goes hand in hand with a high rate of illiteracy, and this interplay compromises a country's future development. In other words, if countries, especially developing ones, are determined to take off, one of the key points governments should set on agenda is to educate more qualified future citizens through elementary education.

...

Table 3.15: The argument components in the example in Table 3.14 and the scores of their associated attributes: **P**ersuasiveness, **E**loquence, **S**pecificity, **Ev**idence, **R**elevance, **St**rength, **Lo**gos, **Pa**thos, **Et**hos, **c**laim**Type**, and **p**remise**Type**

		P	E	S	Ev	R	St	Lo	Pa	Et	cType	pType
M1	government is supposed to offer sufficient financial support for both	3	4	2	3			T	F	F		
C1	if countries, especially developing ones, are determined to take off, one of the key points governments should set on agenda is to educate more qualified future citizens through elementary education	4	5	4	4	6		T	F	F	policy	
P1	elementary education is the fundamental requirement to be a qualified citizen in today's society	4	5	3	1	6	4					A
C2	government should guarantee that all people have equal and convenient access to it	2	3	1	1	6		F	F	F	policy	
P2	a lack of well-established primary education goes hand in hand with a high rate of illiteracy, and this interplay compromises a country's future development	4	5	3	1	6	4					C

3.2.6 Example

To better understand our annotation scheme, we use the essay in Table 3.14 to illustrate how we obtain the attribute values in Table 3.15. In this essay, Claim **C1**, which supports MajorClaim **M1**, is supported by three children, Premises **P1** and **P2** as well as Claim **C2**.

After reading the essay in its entirety and acquiring a holistic impression of the argument's strengths and weaknesses, we begin annotating the atomic argument components bottom up, starting with the leaf nodes of the argument tree. First, we consider **P2**. Its Evidence score is 1 because it is a leaf node with no supporting evidence. Its Eloquence score is 5 because the sentence has no serious grammatical or syntactic errors, has a flowing, well thought out sentence structure, and uses articulate vocabulary. Its Specificity score is 3 because it is essentially saying that poor primary education causes illiteracy and consequently inhibits a country's development. It does not state why or to what extent, so we cannot assign a score of 4. However, it does explain a simple relationship with little ambiguity due to the lack of hedge words, so we can assign a score of 3. Its PremiseType is *common_knowledge* because it is reasonable to assume most people would agree that poor primary education causes illiteracy, and also that illiteracy inhibits a country's development. Its Relevance score is 6: its relationship with its parent is clear because the two components exhibit coreference. Specifically, **P2** contains a reference to primary/elementary education and shows how this affects a country's inability to transition from developing to developed. Its Strength is 4: though eloquent and relevant, **P2** is lacking substance in order to be considered for a score of 5 or 6. The PremiseType is *common_knowledge*, which is mediocre compared to statistics and real_example. In order for a premise that is not grounded in the real world to be strong, it must be very specific. **P2** only scored a 3 in Specificity, so we assign a Strength score of 4. Finally, the argument headed by **P2**, which does not have any children, has a Persuasiveness

[1]There is a caveat. If we define features for each of the children, the number of features will be proportional to the number of children. However, SVMs cannot handle a variable number of features. Hence, all of the children will be represented by one set of features. For instance, the Specificity feature value of the children will be the Specificity values averaged over all of the children.

score of 4, which is obtained by summarizing the inherent strength of the premise and the supporting evidence. Although there is no supporting evidence for this premise, this does not adversely affect persuasiveness due to the standalone nature of premises. In this case the persuasiveness is derived totally from the strength.

Next, the annotator would score **C2** and **P1**, but for demonstration purposes we will examine the scoring of **C1**. **C1**'s Eloquence score is 5 because it shows fluency, broad vocabulary, and attention to how well the sentence structure reads. Its ClaimType is *policy* because it specifically says that the government should put something on their agenda. Its Specificity score is 4: while it contains information relevant to all the child premises (i.e., creating qualified citizens, whose role it is to provide the education, and the effect of education on a country's development), it does not contain a qualifier stating the extent to which the assertion holds true. Its Evidence score is 4: **C1** has two premises with decent persuasiveness scores and one claim with a poor persuasiveness score, and there are no attacking premises, so intuitively, we may say that this is a midpoint between many low quality premises and few high quality premises. We mark Logos as true, Pathos as false, and Ethos as false: rather than use an emotional appeal or an appeal to authority of any sort, the author attempts to use logical reasoning in order to prove their point. Its Persuasiveness score is 4: this score is mainly determined by the strength of the supporting evidence, given that the assertion is precise and clear as determined by the specificity and eloquence. Its Relevance score is 6, as anyone can see how endorsement of elementary education in **C1** relates to the endorsement of elementary and university education in its parent (i.e., **M1**).

After all of the claims have been annotated in the bottom-up method, the annotator moves on to the major claim, **M1**. **M1**'s Eloquence score is 4: while it shows fluency and a large vocabulary, it is terse and does not convey the idea exceptionally well. Its persuasion strategies are obtained by simply taking the logical disjunction of those used in its child claims. Since every claim in this essay relied on logos and did not employ pathos nor ethos, **M1** is marked with Logos as true, Pathos as false, and Ethos as false. Its Evidence score is 3: in this essay there are two other

supporting claims not in the excerpt, with persuasiveness scores of only 3 and 2, so **M1**'s evidence has one decently persuasive claim, one claim that is poor but understandable, and one claim that is so poor as to be completely unpersuasive (in this case it has no supporting premises). Its Specificity score is 2 because it does not have a quantifier nor does it attempt to summarize the main points of the evidence. Finally, its Persuasiveness score is 3: all supporting claims rely on logos, so there is no added persuasiveness from a variety of persuasion strategies, and since the eloquence and specificity are adequate, they do not detract from the Evidence score.

CHAPTER 4

MODELS

The analysis in the last chapter has implied that models that are trained on these attributes for persuasiveness scoring could provide useful feedback to students on *why* their arguments are (un)persuasive. Motivated by this implication, in this chapter, we propose the first set of neural models for predicting persuasiveness and attributes impacting persuasiveness.

4.1 Baseline Model

Since one of the goals of our evaluation is to determine the usefulness of the automatically predicted attributes for persuasiveness scoring, we design a Baseline model that scores persuasiveness *without* using any attributes.

Figure 4.1 shows the Baseline model. Baseline takes as input an argument, which corresponds to a node in an argument tree and its n children, if any, and scores the argument's persuasiveness. Baseline relies on bidirectional long short term memory networks (biLSTMs) (Schuster and Paliwal, 1997). Recall that biLSTMs use both the previous and future context by processing the input sequence in two directions. The final representation is the concatenation of the (last timesteps of each of the) forward and backward steps.

Specifically, Baseline first uses n+1 biLSTMs: one for creating a representation of the parent's word sequence and the remaining n for creating representations of its n children. It then concatenates these n+1 representations. The resulting vector first goes through a dense layer, which reduces the vector's dimension to 150 (with Leaky ReLU as the activation function), then goes through another dense layer for scoring (again with Leaky ReLU as the activation function). To represent the words, we use the 300-dimensional Facebook FastText pre-trained word embeddings (Bojanowski et al., 2017). To handle out-of-word vocabulary words, we create random word vectors and map each of them to the same random vector. The network is trained to minimize mean

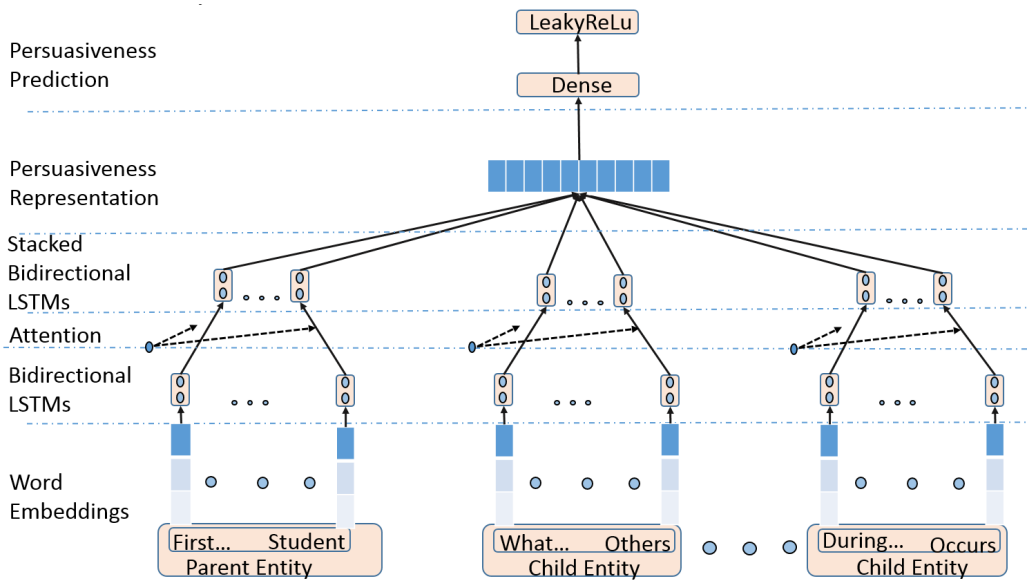

Figure 4.1: Baseline neural network architecture for persuasiveness scoring

absolute error. Early stopping is used to choose the best epoch. Specifically, training stops when the loss on development data stops improving after 20 epochs.

In addition, we evaluate two extensions to Baseline.

Attention mechanism. In order for the model to focus on the relevant parts of the n+1 representations created by the biLSTMs, we apply an attention mechanism to each of these representations separately. In our attention mechanism, we determine the importance weighting α_t for each hidden state h_t of a biLSTM as follows:

$$e_t = \tanh(Wh_t + b); \quad \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)};$$

where W (the kernel weight matrix) and b (the bias) are tunable parameters of the mechanism, and e_t is the hidden representation of h_t. Using α_t, we determine the context vector c_t at timestep t as follows:

$$c_t = \alpha_t e_t$$

This yields a sequence of context vectors $c_1 c_2 \ldots c_T$ for each of the n+1 biLSTMs. These n+1 sequences then serve as the inputs for a second set of n+1 biLSTMs, whose output vectors are concatenated and passed to the first dense layer.

Additional features. We determine whether incorporating additional features can improve Baseline's performance. Specifically, we employ 17 of the linguistic features originally used by Tan et al. (2016) for a task related to argument persuasiveness. The 17 features, which are defined on the argument under consideration, include: #words, #definite/indefinite articles, #positive/negative words, #1st/2nd person pronouns, #1st person plural pronouns, #hedges, #examples, #quotations, #sentences, #quantifiers, #children, type-token ratio, fraction of definite articles, and fraction of positive words. When applied, these features will be concatenated with the representations created by the n+1 biLSTMs.

Note that the attention mechanism and the additional features can be applied in isolation and in combination. When used in combination, the additional features will be concatenated with the vectors created by the second (rather than the first) set of n+1 biLSTMs described above.

4.2 Pipeline Model

Figure 4.2, 4.3 and 4.4 show the Pipeline model. Our Pipeline model operates in two steps. First, it predicts each attribute in Table 3.3 independently of other attributes. Then, it uses the predicted attributes to score persuasiveness. Below we describe these two steps in detail.

We note that a Type 1 attribute, which includes Eloquence, Specificity, Strength, ClaimType, and PremiseType, can be computed using a single argument component (i.e., either the parent or a child), as exemplified by Attribute 1 in the Attribute Representation layer of Figure 4.2. A Type 2 attribute, which includes Logos, Ethos, and Pathos, is computed using both the parent and all of its children, as exemplified by Attribute 2 in Attribute Representation layer of Figure 4.3. A Type 3 attribute, which includes Evidence, is computed using all of the children, as exemplified by Attribute 3 in the Attribute Representation layer of Figure 4.4.

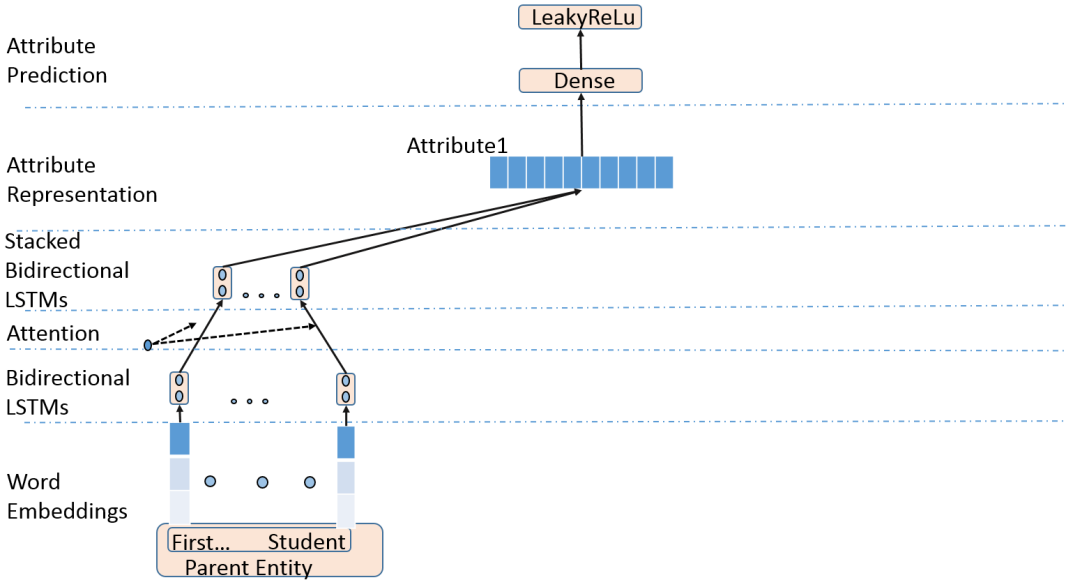

Figure 4.2: Pipeline step1 neural network architecture for attribute type 1 scoring

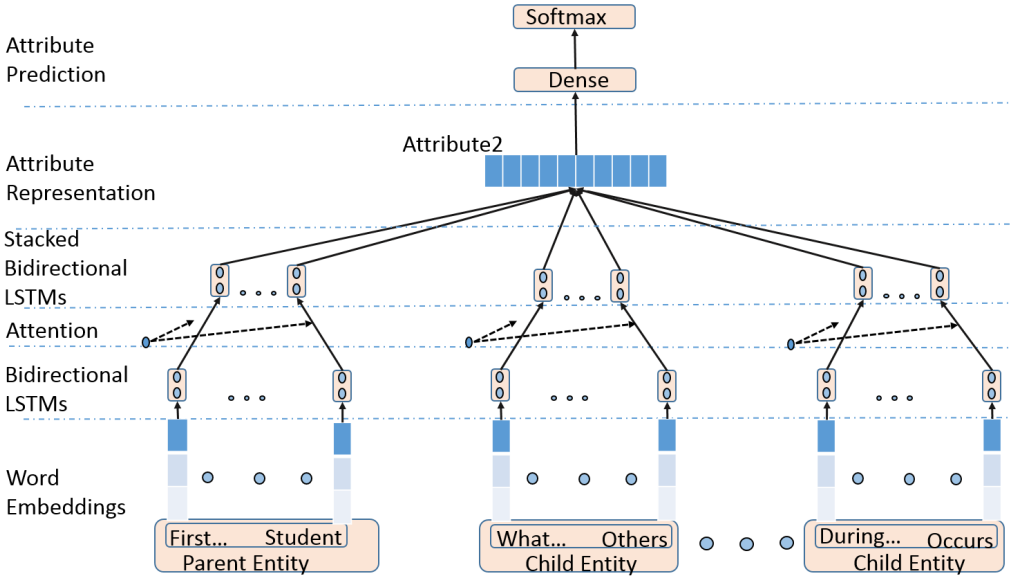

Figure 4.3: Pipeline step1 neural network architecture for attribute type 2 scoring

Step 1: Type 1 attribute attributes, including Specificity, Eloquence, Strength, ClaimType, and PremiseType, are defined on an argument *component* (as opposed to an argument, which involves more than one argument component). To predict each of these attributes, we employ a network whose architecture is the same as that of Baseline except that it uses one biLSTM (rather than n+1 biLSTMs) because the input is composed solely of the word sequence appearing in the argument

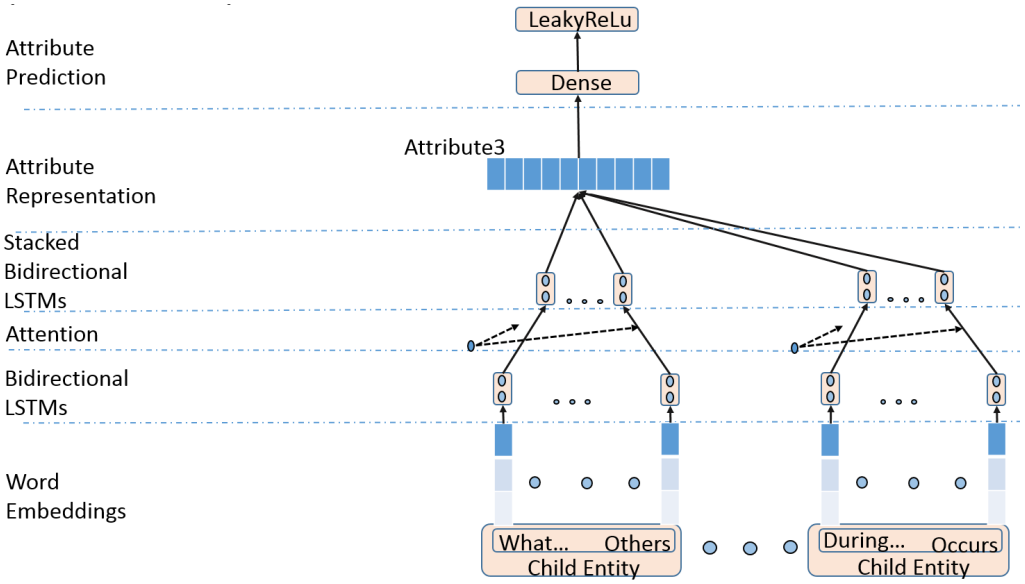

Figure 4.4: Pipeline step1 neural network architecture for attribute type 3 scoring

component whose attributes are to be predicted, as shown in Figure 4.2. If a discrete- (rather than a real-)valued attribute is to be predicted (e.g., ClaimType), we need to replace Leaky ReLU with softmax as the activation function and mean absolute error (ME) with cross entropy as the objective function in the dense layers.

Type 3 attribute, Evidence, in contrast, is defined on the n children of an argument. So, to predict evidence, we use the Baseline network architecture but with n biLSTMs, each of which creates a representation of one of the children, as shown in Figure 4.4.

Type 2 attribute (Logos, Ethos, Pathos) is defined on an entire argument and will be predicted using a network that has the same architecture as that of Baseline, as shown in Figure 4.3.

Finally, note that the attention mechanism can be applied to the networks described in this step in the same way as in the Baseline.

Step 2: As shown in Figure 4.5, to score the persuasiveness of an argument, we feed the vector of attributes predicted in Step 1 that involve all of its components into a dense layer that has Leaky ReLU as its activation function. This network is trained on vectors of gold attribute values. Note that the additional features described in Baseline can be applied to train this network simply by concatenating them with the vector of attributes.

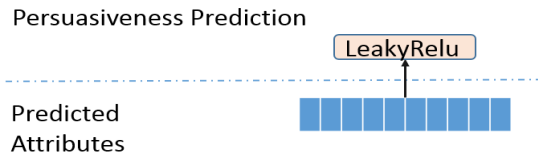

Figure 4.5: Pipeline step2 neural network architecture for persuasiveness scoring

4.3 Joint Model

Figure 4.6 shows the Joint model, which is a neural network that simultaneously scores the persuasiveness of an argument and predicts its attributes. Note that the Baseline network is a special case of this network where only one attribute is predicted, namely persuasiveness. Similarly for the networks used in Step 1 of the Pipeline model: each of them is a special case of this network where only one attribute is predicted.

Similar to Pipeline model, We also note that Figure 4.6 is a simplified view of the Joint network. Specifically, while the output layer seems to suggest that the network predicts only three things, in reality there is one output node for each of the attributes of the argument under consideration and its persuasiveness. One representation will be created for each such attribute as well as persuasiveness in the Attribute Representation layer. An attribute belongs to one of three *types* mentioned in the pipeline models sections (note that Persuasiveness belongs to Type 2).

Each of the representations in the Attribute Representation layer is created using an attribute-specific biLSTM in the Attribute-specific Bidirectional LSTMs layer of Figure 4.6. For instance, to predict the parent's Eloquence, one biLSTM will be created specifically for it in this layer.

Like other networks for multi-task learning, this network has n+1 "shared" biLSTMs (see the Shared Bidirectional LSTMs layer) that create representations for the parent and its n children that are shared by multiple prediction tasks.

Like the Baseline and Pipeline models, the attention mechanism and the additional features can be optionally applied in the Joint model. If the attention mechanism is not applied, the outputs of the biLSTMs in the lower layer will directly become the inputs for the biLSTMs in the upper layer.

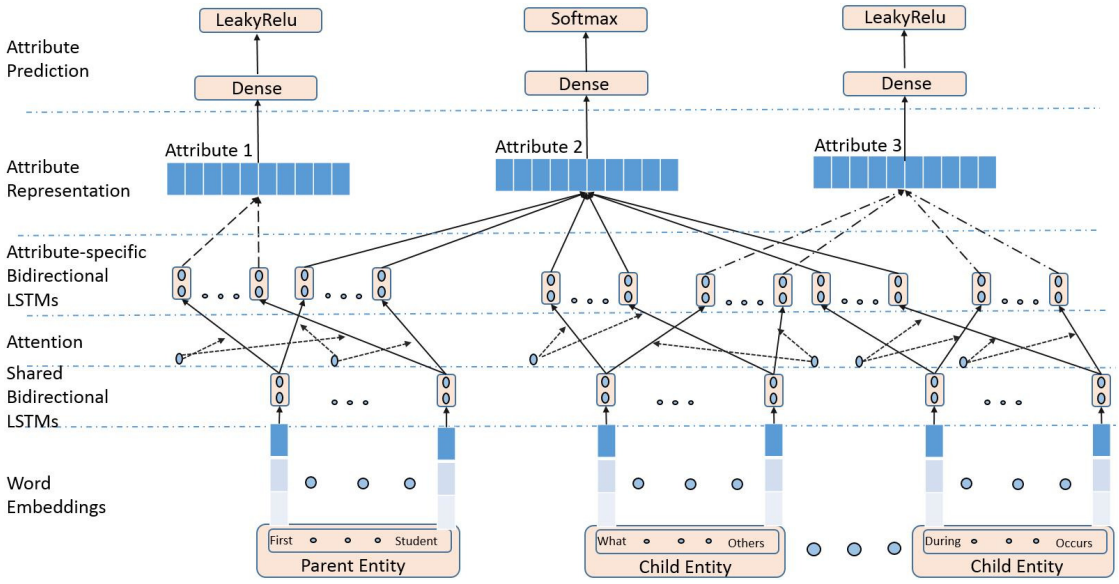

Figure 4.6: Neural network architecture for joint persuasiveness scoring and attribute prediction

If additional features are used, they will be concatenated with each of the vectors in the Attribute Representation layer.

CHAPTER 5

EVALUATION

In this chapter, we evaluate our three neural models and discuss the results in detail.

5.1 Experimental Setup

We first randomly partition our 102 essays into five folds, each of which contains 20–21 essays, and then conduct five-fold cross-validation experiments. In each fold experiment, we employ three folds for training, one fold for development, and one fold for testing. Given a training/development/test set, we first divide the available arguments into three subsets depending on whether the argument's parent node is a MajorClaim, Claim, or Premise. We then train one model on each of the three subsets of training arguments and apply each model to classify the corresponding development/test arguments.

For persuasiveness scoring, we employ two evaluation metrics, PC and ME. PC computes the Pearson's Correlation Coefficient between a model's predicted scores and the annotator assigned scores. In contrast, ME measures the mean absolute distance between a system's prediction and the gold score. Note that PC is a *correlation* metric, so higher correlation implies better performance. In contrast, ME is an *error* metric, so lower scores imply better performance.

5.2 Results and Discussion

Persuasiveness scoring results of the three models on the *development* set obtained via five-fold cross validation are shown in Table 5.1. Four variants of each model are evaluated. The U variants are trained *without* the 17 additional features and attention; the UF variants are trained with the 17 features but without attention; the UA variants are trained with attention but without the 17 features; and the UFA variants are trained with both attention and the 17 features. Results, expressed in terms of the PC and ME scoring metrics, are first computed separately for arguments

whose parent nodes correspond to a MajorClaim (MC), Claim (C), and Premise (P) before being micro-averaged (Avg). The strongest result in each column w.r.t. each metric is boldfaced.

First, to determine whether automatically computed attributes are useful for persuasiveness scoring, we compare Baseline, which does not use attributes, with Pipeline and Joint, both of which predict attributes. W.r.t. ME, we see that Joint's Avg scores are consistently better than those of Baseline, and Pipeline's Avg scores are better than those of Baseline on all but one variant (UF). The best Avg score is achieved using Pipeline-UA. If we examine the MC, C, and P results, we see some consistency across the three models. Specifically, the best MC results come from the U variant, and the best C and P results both come from the UA variant. These results seem to suggest that scoring a MC's persuasiveness does not benefit from the addition of features and the use of an attention mechanism, whereas scoring a C or P's persuasiveness benefits from applying attention in the absence of additional features. W.r.t. PC, the Avg results are somewhat mixed. Specifically, Joint outperforms Baseline on two of the four variants (U and UF), whereas Pipeline outperforms Baseline on all but the UFA variant. Examining the MC, C, and P results, we no longer see any consistency across the three models. Specifically, for Baseline, the best MC, C, and P results all come from the UFA variant; for Pipeline, the best MC come from UF, and the best C and P results come from UA; and for Joint, the best MC, C, and P results are from different variants. These results suggest that while Baseline consistently benefits from employing attention and the additional features, the same is not true for Pipeline and Joint. For instance, Pipeline benefits from attention when scoring a C or P's persuasiveness and from using additional features when scoring a MC's persuasiveness.

Given these development results, we hypothesize that our three persuasiveness scoring models could be improved by scoring a MC, C, and P's persuasiveness using different variants. To test this hypothesis, we conduct the following experiment. For each of the three models, we score a MC/C/P's persuasiveness on the *test* set using the variant that achieved the best performance on the development set w.r.t. a particular scoring metric. Doing so gives each model the flexibility to use

Table 5.1: Persuasiveness scoring results of the four variants (U, UF, UA, UFA) of the three models (Baseline, Pipeline, and Joint) on the development set as measured by the two scoring metrics (PC and ME)

		Baseline				**Pipeline**				**Joint**			
	System	MC	C	P	Avg	MC	C	P	Avg	MC	C	P	Avg
PC	U	.115	.093	.207	.158	.035	.136	.263	.196	**.165**	.074	.199	.163
	UF	$-.041$	.204	.221	.177	**.155**	.025	.303	.210	.060	**.207**	.276	**.226**
	UA	.022	.181	.214	.179	.088	**.169**	**.343**	**.259**	.054	.182	.081	.104
	UFA	**.165**	**.245**	**.293**	**.261**	.047	.082	.204	.149	.052	.087	**.304**	.209
ME	U	**1.028**	1.099	1.058	1.065	**.991**	1.035	1.015	1.017	**1.021**	1.060	1.063	1.056
	UF	1.150	1.039	1.047	1.062	1.272	1.418	.977	1.135	1.097	1.008	.989	1.011
	UA	1.107	**.968**	**1.016**	**1.018**	1.081	**.970**	**.948**	**.975**	1.141	**.969**	**.965**	**.993**
	UFA	1.178	.996	1.032	1.046	1.229	1.034	.957	1.019	1.217	1.016	.999	1.036

different variants when scoring MC's, C's, and P's persuasiveness. For instance, it is possible for Pipeline to use UF when scoring a MC's persuasiveness and UA when scoring a C's persuasiveness, and such choices can change depending on the scoring metric.

Five-fold cross-validation results of the aforementioned experiment are shown in Table 5.2. As mentioned before, these are results on the *test* set. A few points deserve mention. First, Joint consistently beats Baseline, outperforming it on MC, C, P, and Avg w.r.t. both scoring metrics. Second, while Pipeline outperforms Baseline w.r.t. Avg PC (primarily because of its superior performance on P), it underperforms Baseline w.r.t. Avg ME (primarily because of its inferior performance on MC and C). Finally, Joint consistently outperforms Pipeline w.r.t. ME, but underperforms it w.r.t. Avg PC only because of its inferior performance on P. Hence, we may be able to obtain further gains by creating an "ensemble" model where we apply Pipeline when scoring a P's persuasiveness w.r.t. PC and use Joint otherwise. Overall, given that Joint has consistently superior performance to Baseline, we conclude that automatically computed attributes are useful for persuasiveness scoring. Nevertheless, the usefulness of these automatically computed attributes depends in part on how they are used, as shown by the difference in Pipeline's and Joint's results.

To gain additional insights into the usefulness of these attributes, we conduct an oracle experiment where we use *gold* attribute values for persuasiveness scoring by training the same neural network that was used in the second step of the Pipeline model. Cross-validation results, which are

Table 5.2: Persuasiveness scoring results on the test set obtained by employing the variant that performs the best on the development set w.r.t. the scoring of MC/C/P's persuasiveness

		Baseline				**Pipeline**				**Joint**			
	System	MC	C	P	Avg	MC	C	P	Avg	MC	C	P	Avg
PC	Best	.034	.145	.269	.205	.038	.138	.353	.248	.148	.163	.290	.236
ME	Best	1.280	1.036	1.056	1.086	1.363	1.237	1.041	1.147	1.220	1.032	.983	1.035

Table 5.3: Attribute prediction results of different variants of Pipeline and Joint on the test set

	Pipeline									**Joint**								
	Evid.	**Eloq.**	**Spec.**	**Stre.**	**Log.**	**Path.**	**Eth.**	**CT**	**PT**	**Evid.**	**Eloq.**	**Spec.**	**Stre.**	**Log.**	**Path.**	**Eth.**	**CT**	**PT**
U	.353	.126	.284	.132	.281	.045	.045	.692	.200	.335	.106	.243	.697	.281	.863	.967	.695	.207
UF	.309	.206	.417	.132	.281	.045	.045	.692	.210	.340	.178	.409	.697	.281	.829	.961	.676	.189
UA	.347	.148	.306	.132	.281	.045	.045	.692	.308	.366	.168	.332	.697	.281	.967	.967	.692	.268
UFA	.368	.202	.414	.132	.281	.045	.045	.692	.239	.341	.173	.429	.697	.281	.868	.967	.692	.312

shown in Table 3.13, provide strong evidence that these attributes are very useful for persuasiveness scoring.

Finally, we report attribute prediction performance on the test set in Table 5.3 where each attribute's results are micro-averaged over its respective argument component types. Real- and discrete-valued attributes are evaluated using PC and $F1$ respectively. As we can see, Joint outperforms Pipeline on predicting Strength, Pathos and Ethos, but the two yield similar results otherwise. Overall, attribute prediction performance is rather mediocre. Comparing the results in Tables 5.1, 5.2, and 5.3 we see that while the attributes are useful for persuasiveness scoring, their usefulness in our models is limited by the accuracy with which they are computed.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The development of argument persuasiveness, which is arguably the most important dimension of essay quality, has been heavily hindered by the scarcity of annotated corpora needed for model training. In this thesis, we present the first corpus of 102 persuasive student essays that are simultaneously annotated with argument trees, persuasiveness scores, and attributes of argument components that impact these scores. The inter-annotator agreement has shown that this corpus is reliable and the oracle experiment has implied that models that are trained on these attributes for persuasiveness scoring could provide useful feedback to students on why their arguments are (un)persuasive. Motivated by this implication, we further propose the first set of neural models for predicting the persuasiveness of an argument and its attributes in a student essay. The evaluation of our proposed models have shown that automatically computed attributes are useful for persuasiveness scoring and performance could be improved by improving attribute prediction. Our approach to argument persuasiveness scoring is *different* from the existing work at least in two dimensions: (1) we annotate our own, reliable corpus of argument persuasiveness as well as attributes impacting persuasiveness; and (2) we proposes a first set of neural models for predicting the persuasiveness of an argument and its attributes in a student essay.

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

One direction concerns attribute scoring. As we have mentioned in the last section, the attribute prediction is rather mediocre and their usefulness in our models is limited by the accuracy with which they are computed. Future work should be done in order to make a better performance on attribute scoring. For instance, the attribute *specificity* in our model may be further improved by using some specificity scoring technologies, as shown in Li and Nenkova (2015) and Ko et al. (2018)

In addition to exploring the interaction between different dimensions, we believe it is worthwhile to examine how *argument persuasiveness* interacts with other dimensions, such as *thesis clarity* and *prompt adherence*. For instance, intuitively, an argument is less likely to be persuasive if it is off-topic or unclear, hence *thesis clarity* and *prompt adherence* may help predict the argument persuasiveness. Based on such intuitive observation, some *multi-task* learning or *transfer* learning techniques may be in use for improving *argument persuasiveness* scoring.

Finally, to enable argument persuasiveness scoring 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 argument persuasiveness scoring can help improve their writing skills.

REFERENCES

Alikaniotis, D., H. Yannakoudakis, and M. Rei (2016). Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Amorim, E., M. Cançado, and A. Veloso (2018). Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 229–237.

Attali, Y. and J. Burstein (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment 4*(3).

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, Stroudsburg, PA, USA, pp. 86–90. Association for Computational Linguistics.

Barzilay, R. and M. Lapata (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics 34*(1), 1–34.

Blanchard, D., J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow (2013). Toefl11: A corpus of non-native english. *ETS Research Report Series 2013*(2), i–15.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics 5*, 135–146.

Breland, H. M., R. J. Jones, L. Jenkins, M. Paynter, J. Pollack, and Y. F. Fong (1994). The college board vocabulary study. *ETS Research Report Series 1994*(1), i–51.

Burstein, J., M. Chodorow, and C. Leacock (2004). Automated essay evaluation: The criterion online writing service. *Ai Magazine 25*(3), 27.

Burstein, J., J. Tetreault, and S. Andreyev (2010). Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pp. 681–684. Association for Computational Linguistics.

Chen, H. and B. He (2013). Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1741–1752.

Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM.

Cozma, M., A. M. Butnaru, and R. T. Ionescu (2018). Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.

Crossley, S., L. K. Allen, E. L. Snow, and D. S. McNamara (2015). Pssst... textual features... there is more to automatic essay scoring than just you! In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pp. 203–207. ACM.

Cummins, R., M. Zhang, and E. J. Briscoe (2016). Constrained multi-task learning for automated essay scoring. Association for Computational Linguistics.

Daume III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263.

Dong, F. and Y. Zhang (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.

Dong, F., Y. Zhang, and J. Yang (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 153–162. Association for Computational Linguistics.

Elliot, S. (1999). Construct validity of intellimetric with international assessment. *Yardley, PA: Vantage Technologies (RB-323)*.

Essayforum (2012). https://essayforum.com/.

Farra, N., S. Somasundaran, and J. Burstein (2015). Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 64–74.

Faulkner, A. (2014). Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. *Science 376*(12), 86.

Ghosh, D., A. Khanam, Y. Han, and S. Muresan (2016). Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Volume 2, pp. 549–554.

Granger, S., E. Dagneaux, F. Meunier, and M. Paquot (2009). *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.

Grosz, B. J., S. Weinstein, and A. K. Joshi (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics 21*(2), 203–225.

Habernal, I. and I. Gurevych (2016a, November). What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1214–1223.

Habernal, I. and I. Gurevych (2016b, August). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1589–1599.

Hewlett Foundation (2012). https://www.kaggle.com/c/asap-aes.

Higgins, D., J. Burstein, D. Marcu, and C. Gentile (2004). Evaluating multiple aspects of coherence in student essays. In *Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 185–192.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Horbach, A., D. Scholten-Akoun, Y. Ding, and T. Zesch (2017). Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 357–366. Association for Computational Linguistics.

Jin, C., B. He, K. Hui, and L. Sun (2018). Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1088–1097.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*, Chapter 11, pp. 169–184. Cambridge, MA: MIT Press.

Klebanov, B. B. and M. Flor (2013). Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Volume 1, pp. 1148–1158.

Klebanov, B. B., M. Flor, and B. Gyawali (2016). Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 63–72.

Klebanov, B. B., N. Madnani, and J. Burstein (2013). Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *Transactions of the Association for Computational Linguistics 1*, 99–110.

Ko, W.-J., G. Durrett, and J. J. Li (2018). Domain agnostic real-valued specificity prediction. *arXiv preprint arXiv:1811.05085*.

Krippendorff, K. (1980). Content analysis; an introduction to its methodology. Technical report.

Landauer, T. K., D. Laham, and P. W. Foltz (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pp. 87–112. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Li, J. J. and A. Nenkova (2015). Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Louis, A. and D. Higgins (2010). Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 92–95.

Mann, W. C. and S. A. Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse 8*(3), 243–281.

McNamara, D. S., S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing 23*, 35–59.

Miltsakaki, E. and K. Kukich (2004). Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering 10*(1), 25–55.

Nguyen, H. V. and D. J. Litman (2018). Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Östling, R., A. Smolentzov, B. Tyrefors Hinnerich, and E. Höglin (2013, June). Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, pp. 42–47. Association for Computational Linguistics.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan 47*(5), 238–243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of experimental education 62*(2), 127–142.

Pennington, J., R. Socher, and C. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Persing, I., A. Davis, and V. Ng (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 229–239.

Persing, I. and V. Ng (2013). Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 260–269.

Persing, I. and V. Ng (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1534–1543.

Persing, I. and V. Ng (2015, July). Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 543–552.

Persing, I. and V. Ng (2016). Modeling stance in student essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2174–2184.

Phandi, P., K. M. A. Chai, and H. T. Ng (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 431–439.

Schuster, M. and K. K. Paliwal (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing 45*(11), 2673–2681.

Somasundaran, S., J. Burstein, and M. Chodorow (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *COLING*, pp. 950–961.

Stab, C. and I. Gurevych (2014, August). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, pp. 1501–1510. Dublin City University and Association for Computational Linguistics.

Taghipour, K. and H. T. Ng (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891. Association for Computational Linguistics.

Tan, C., V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International World Wide Web Conference*.

Tay, Y., M. C. Phan, L. A. Tuan, and S. C. Hui (2018). Skipflow: incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education 28*(1), 79–105.

van Eemeren, F. H., B. Garssen, E. C. W. Krabbe, F. A. S. Henkemans, B. Verheij, and J. H. M. Wagemans (2014). In *Handbook of Argumentation Theory*. Dordrecht: Springer.

Wachsmuth, H., K. Al Khatib, and B. Stein (2016). Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1680–1691.

Wang, Y., Z. Wei, Y. Zhou, and X. Huang (2018). Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 791–797. Association for Computational Linguistics.

Yannakoudakis, H. and T. Briscoe (2012). Modeling coherence in esol learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 33–43. Association for Computational Linguistics.

Yannakoudakis, H., T. Briscoe, and B. Medlock (2011, June). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 180–189. Association for Computational Linguistics.

Zesch, T., M. Wojatzki, and D. Scholten-Akoun (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 224–232.

BIOGRAPHICAL SKETCH

Zixuan Ke has worked on multi-dimensional essay scoring for the last two years under Professor Vincent Ng at UTD. By the completion of his degree, he has published two papers in ACL and IJCAI respectively. Before joining to UTD, Zixuan was an undergraduate student at South China Agricultural University, where he worked on Dialogue Act Classification, which is an essential part of dialogue system. During the 4 years he spent there, he published one paper and two journal articles on dialogue area.

CURRICULUM VITAE

Zixuan Ke

Curriculum Vitae

Research Interests

Human language technologies, *Artificial intelligence*, Machine learning.

Education

2017–2019 **M.Sc., Computer Science**, *The University of Texas at Dallas*, Machine Learning Doctoral Qualifying Exam Passed.

2013–2017 **B.Sc., Network Engineering**, *South China Agricultural University*, *GPA – 3.7*. Rank: 3/100

2007–2013 **High School Diploma**, *Guangdong Experimental High School, China*.

Publications

[1] **Zixuan Ke** ,Winston Carlile, Nishant Gurrapadi and Vincent Ng, Learning to Give Feedback: Modeling Attributes Affecting Argument Persuasiveness in Student Essays, Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI), 2018.

[2] Winston Carlile, Nishant Gurrapadi **Zixuan Ke** and Vincent Ng, Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL, Volume 1: Long Papers), 2018

[3] **KE Zixuan** ,HUANG Peijie, ZENG Zhen, Domain Classification Based on Undefined Utterances Detection Optimization, Journal of Chinese Information Processing 2018, Vol. 32, No. 4, Apr. 2018

[4] HUANG Peijie, WANG Jundong, **KE Zixuan** , Lin Piyuan, Dialogue Act Recognition for Out-of-Domain Utterances in Restricted Domain Spoken Dialogue System, Journal of Chinese Information Processing 2016, 30(6): 200-208

[5] Jundong Wang, Peijie Huang, Qiangjia Huang, **Zixuan Ke**, Piyuan Lin. Dialogue Act Recognition for Chinese Out-of-Domain Utterances using Hybrid CNN-RF, In Proceedings of the 20th International Conference on Asian Language Processing (IALP 2016)

Awards

2017 **First place**, *User Intention and Domain Classification*, SMP2017-ECDT.

2016 **First-class Prize**, *Stance Detection in Chinese Micro blogs*, NLPCC2016-SDCM.

2016 **Top 20 Cognition Application (APP) within IBM General China Group (IBM GCG)** , *International Business Machines Corporation (IBM).*

2015,2016,2017 **Annual Scholarship for Top 10% students**, *South China Agricultural University.*

Experience

Research

2017–present **Research Group Member**, Machine Learning Lab, The University of Texas at Dallas.
Supervisor: Dr. Vincent Ng

- Built Machine Learning Systems (both deep learning and traditional methods) to automatically grade student essays (in different dimensions, such as persuasiveness, adherence and organization etc.)

2014–2017 **Research Assistant**, Machine Learning Lab, South China Agricultural University.
Supervisor: Dr. Peijie Huang

- Developed and designed the Spoken Language Understanding (SLU) module for the school's internal chat bot.

Internships

2016 **Summer Intern**, IBM Procurement China Limited (IBM), Shenzhen.
Projects: Audit Tool and Electronic Functional Control

- Developed a real-time business intelligence server module for statistical analysis of sales data using Node-Red (IBM internal framework) and Django.