The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Argument Mining for Improving the Automated Scoring of Persuasive Essays

Huy V. Nguyen
Computer Science Department
University of Pittsburgh, PA 15260
hvn3@pitt.edu

Diane J. Litman
Computer Science Department
University of Pittsburgh, PA 15260
litman@cs.pitt.edu

Abstract

End-to-end argument mining has enabled the development of new automated essay scoring (AES) systems that use argumentative features (e.g., number of claims, number of support relations) in addition to traditional legacy features (e.g., grammar, discourse structure) when scoring persuasive essays. While prior research has proposed different argumentative features as well as empirically demonstrated their utility for AES, these studies have all had important limitations. In this paper we identify a set of desiderata for evaluating the use of argument mining for AES, introduce an end-to-end argument mining system and associated argumentative feature sets, and present the results of several studies that both satisfy the desiderata and demonstrate the value-added of argument mining for scoring persuasive essays.

Introduction

Argument mining is an emerging field which aims to automatically identify argumentative text portions and the relevant components of the presented argument (Peldszus and Stede 2013). Recent years have seen the development of argument mining architectures and systems (Stab and Gurevych 2014; Peldszus and Stede 2015; Stab and Gurevych 2017) as well as the use of argument mining output to improve applications such as summarization, opinion mining, and automated essay scoring (Boltužić and Šnajder 2014; Egan, Siddharthan, and Wyner 2016; Barker and Gaizauskas 2016; Ghosh et al. 2016; Klebanov et al. 2016; Wachsmuth, Al Khatib, and Stein 2016). As an example, Figure 1 depicts our argument mining system based on our prior studies (Nguyen and Litman 2016b; 2016a).

In automated essay scoring (AES) – "the process of scoring written prose via computer program" (Shermis and Burstein 2013), argumentative features have been extracted from the output of a range of argument mining systems to help predict both argument-related (e.g., argument strength) and holistic essay scores. It was hypothesized that the use of good argumentive structures would correlate with essay quality, and that argument mining - which extracts argumentative structures from essays - should thus be able to help improve AES (Klebanov et al. 2016). We identify a set of 7 key attributes that we believe to be important when evaluating argument mining for AES.

An **end-to-end** (e2e) argument mining system takes a text as input and outputs argumentative structure(s) in a fully automated manner, and thus is more applicable to AES in practice. (Persing and Ng 2015) as well as (Wachsmuth, Al Khatib, and Stein 2016) each used the output of an end-to-end argument mining system to improve scoring the essay dimension of argument strength. While it may be intuitive that argument mining can help predict argument-related scores[1], predicting **holistic scores** (hs) of essays could be much more challenging as many other criteria are also taken into account, e.g., grammar, fluency, coherence. (Ghosh et al. 2016) and (Klebanov et al. 2016) indeed examined this more difficult scenario by proposing a wide range of argumentative features for holistic score prediction. However, their experiments used simple baselines, i.e., word and sentence counts, but not **advanced** (av) AES systems. Moreover, findings in (Ghosh et al. 2016) were limited in that their argument mining system is not end-to-end but depended on human-annotated boundaries (i.e., character indices) of argument components.

Despite their limitations, findings of prior research were impressive in that AES performance gain could be demonstrated using at least some noisy outputs of an argument mining system. For example, all of the above studies have extracted features from argument components and their labels, and achieved AES benefits. However, features from **argumentative relations** (re) between components were not addressed in (Persing and Ng 2015; Wachsmuth, Al Khatib, and Stein 2016). In contrast, (Ghosh et al. 2016) performed **feature ablation** (fa) studies spanning multiple tasks from Figure 1 to compare argument component and argumentative relation features. However as noted above, not all of their argumentative features were fully automatically extracted. An examination of the most useful outputs for AES from end-to-end argument mining tasks is thus needed.

Finally, from a practical perspective, since student populations can vary and even the same population can have a wide range of writing assignments, AES approaches that work well across **multiple data sets** (md) (e.g., produced

[1] See (Wachsmuth et al. 2017) for a survey of how argumentation quality has been computationally modeled and assessed.

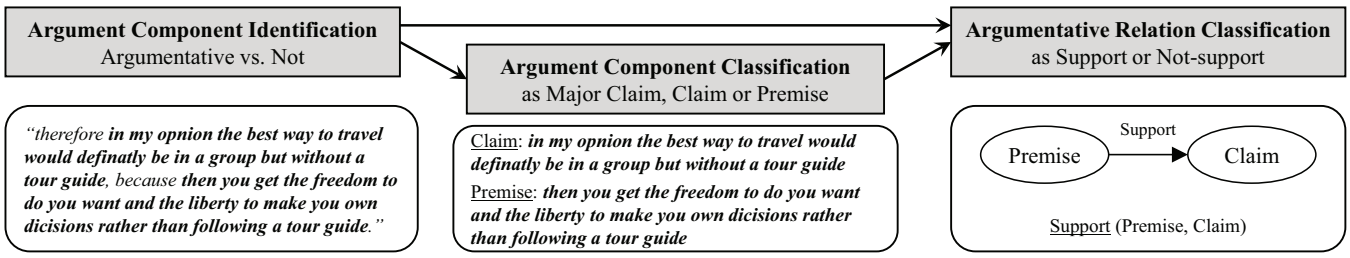

Figure 1: The three argument mining tasks performed by our system (grey boxes). The rounded boxes show each task's output when processing an excerpt of a persuasive TOEFL essay. In the left rounded box, the automatically identified argument components are shown in boldface. These components are then classified in terms of a set of argumentative purpose labels (center rounded box) and a set of relations between labeled components (right rounded box).

Study	hs	av	e2e	re	md	fa	xp
Persing 2015	–	y	y	–	–	–	–
Wachsmuth 2016	–	y	y	–	–	–	–
Ghosh 2016	y	–	–	y	–	y	–
Klebanov 2016	y	–	y	y	–	–	–
This study	y	y	y	y	y	y	y

Table 1: A concise review of recent studies on argument mining for AES. Column headers are explained in the text.

by native vs. non-native speakers), or that are robust when trained and tested on different prompts within a data set, are highly desirable. Because argument mining aims for extracting argumentative structures of text that abstract over specific essay content, features based on argument mining have the potential of being more independent from writing prompts or data set characteristics than many legacy AES features (particularly those that are lexically based). Therefore, it is expected that argumentative features can improve AES in **cross-prompt** (xp) validation. However, the studies discussed above have only evaluated argumentative features for AES using cross-fold validation in which training and test data share the same set of writing prompts, and where all prompts were from a single AES data set.

Table 1 compares recent work on argument mining for AES and summarizes our current study using the 7 key attributes. For each attribute, we indicate whether a study offers that in its AES experiments (y) or not. As can be seen, none of the prior studies addressed all 7 key attributes. To more strongly support the hypothesis that the output of argument mining can improve holistic AES for persuasive essays, our study thus aims to address all 7 key attributes. Our ultimate goal is to give insights on how argument mining can improve already competitive AES methods from different perspectives: feature extraction from noisy end-to-end argument mining output, result generalization across multiple AES data sets, utility of different types of argumentative features, and cross-prompt AES robustness.

To achieve this goal, we first implement an end-to-end argument mining system that parses argumentative structures of free-text essays and creates argumentative features from these structures. Building our argument mining system does not require any training data from the AES datasets. Next, we use machine learning to develop argument mining-enabled AES models that incorporate different combinations of argumentative as well as baseline AES features, training and evaluating all models using two different persuasive essay data sets from prior AES studies. Our experimental results, using both within and cross-prompt evaluation methods, strongly support the hypothesis that adding argumentative features can improve AES performance even under our most challenging experimental settings. Moreover, our ablation studies on features from end-to-end argument mining obtain conflicting findings against a prior study which used human-identified argument components.

Related Work

An end-to-end argument mining system typically consists of three components (Mochales and Moens 2011; Peldszus and Stede 2015; Stab and Gurevych 2017) as depicted in Figure 1: (1) *Argument component identification*, which identifies the boundaries of argument components, e.g., sentences/clauses with specific roles in forming arguments in the text (Peldszus and Stede 2013). (2) *Argument component classification*, which labels each component for its argumentative role (e.g., claim or premise). (3) *Argumentative relation classification*, which determines if an ordered pair of argument components are related (e.g., supports or not).[2]

Different approaches have been proposed to solve these tasks, e.g., sequence labeling to determine boundaries between essay tokens (Stab and Gurevych 2017), text classifiers to determine argumentative roles of clauses (Levy et al. 2014; Persing and Ng 2016). In addition, argument component and argumentative relation classification have been solved using both pipelined (Nguyen and Litman 2016b; 2016a) and joint approaches (Peldszus and Stede 2015; Stab and Gurevych 2017; Persing and Ng 2016). In our study, we build on this prior literature to implement an end-to-end argument mining system that can achieve competitive argument mining performance on community benchmarks. Our goal in this paper is to show how such a system can add value to persuasive essay scoring, rather than to improve further the state-of-the-art in argument mining.

[2]The specific classification labels often vary by system, e.g., including an Attack relation.

Prompt	Essays	Avg. len.	Score range	Median
1	1783	350	2–12	8
2	1800	350	1–6	3

Table 2: Essay score distribution of ASAP data.

Many argumentative features for AES have been proposed in the prior literature, ranging from counts (e.g., number of claims, number of supporting premises) to structure typology (e.g., tree versus chains (Ghosh et al. 2016)) to argument flow patterns (e.g., conclusion–premise versus conclusion–premise–conclusion (Wachsmuth, Al Khatib, and Stein 2016)). However, different levels of argument mining automation have been employed to extract such features. (Klebanov et al. 2016) built a full end-to-end argument mining system, (Ghosh et al. 2016) developed a partially-automated system that started with human-identified argument components, while (Persing and Ng 2015) and (Wachsmuth, Al Khatib, and Stein 2016) did not implement a relation classification component. Our current study investigates all argumentative features from prior studies as well as new features to provide a comprehensive evaluation. In addition, all of our features are based on fully automatic argument mining. Argumentative features are grouped according to the argument mining component enabling the features, and are evaluated together and ablated with respect to their utility for improving AES.

AES Data

To demonstrate the value of argumentative features for AES, we use two corpora of holistically scored persuasive essays that have been studied in prior AES research.

ASAP data

Our first corpus consists of the essays written for prompts 1 and 2 in Kaggle's Automated Student Assessment Prize (ASAP)[3]. The full ASAP corpus has been widely used for AES research (Phandi, Chai, and Ng 2015; Dong and Zhang 2016; Taghipour and Ng 2016) and consists of 8 essay sets, each containing essays for a single prompt. Essays are written by students in the Grade 7 to Grade 10 range and have an average length of 150 to 550 words. From the 8 essay sets, we only use the essays for prompts 1 and 2, which are argumentative. The two essay sets have topics in computer usage and library censorship, respectively. Data statistics of the two essay sets are shown in Table 2. Essays of both sets were double-graded by experts but while the prompt 2 essays have resolved scores, the final prompt 1 scores are the summation of the two expert scores. A computer usage essay excerpt with the argument component boundaries and labels predicted by our argument mining system is given below. The claim in the first sentence was predicted to be supported by the three premises.

> ... The second reason is [*you can learn about far away places and people*]$_{Claim}$. [*Like how we are in the @LOCATION1 and we want to learn about @LOCATION3*]$_{Premise}$ or say that [*we are in france and we want to know about the german people are look it up we dont go all the way down there jest to see the german people we look it up*]$_{Premise}$. [*Now think about it you jest going to fly there and come back the same day*]$_{Premise}$.

[3]https://www.kaggle.com/c/asap-aes/data. All essays have named entities replaced by corresponding NER labels.

	Training	Test
#essays	6074	2023
#prompts	8	
Low score	655	222
Medium score	3318	1101
High score	2101	700

Table 3: Essay score distribution of TOEFL11 data.

TOEFL11 data

Our second corpus includes over 8000 essays from the TOEFL11 corpus (Blanchard et al. 2013). Essays in the corpus were written by non-native test takers to argue for opinions towards issues stated in 8 writing prompts. An excerpt is given in Figure 1. Although the corpus was first introduced for a Native Language Identification shared task, the coarse-grained holistic scores (i.e., Low, Medium, and High) of essays were provided. The corpus we use was compiled by (Klebanov et al. 2016), then split into training and test sets to study relationships between automatically parsed argumentation structures and essay quality. The essay score distributions are reported in Table 3.

Baseline AES Systems

ASAP data

To create our baseline AES model, we use a publicly available open-source AES system called "Enhanced AI Scoring Engine" (EASE: github.com/edx/ease). EASE was ranked in the top three of the Kaggle ASAP competition despite the fact that it used simple features as described in (Phandi, Chai, and Ng 2015):

Length: Numbers of characters, words, commas, apostrophes. Number of sentence ending punctuations (".", "?", "!"). Average word length (in characters). *Prompt*: Number and fraction of essay words that appear in the prompt divided by the total number of essay words. Number and fraction of words in the essay that are a word or a synonym of a word that appears in the prompt. *Bag-of-words*: Count of useful unigrams and bigrams (unstemmed). Count of stemmed and spell corrected useful unigrams and bigrams. *Part-of-speech* (POS): Number and fraction of good POS sequences over the total number of words.

While bag of words and POS sequences are commonly used in AES, EASE proposed using refined n-grams and POS features for better performance. Useful n-grams were defined as n-grams that separate high-score essays and low-score essays. Good POS sequences are collected from a set of novels and have length 2 to 4.

TOEFL11 data

Due to a limited amount of AES research on TOEFL11 data, we implement a simple yet strong baseline for essay score prediction by employing a variety of features found to be effective in the AES literature (Shermis and Burstein 2013; Dikli 2006; Phandi, Chai, and Ng 2015).

Our first group of features (LENGTH) includes 5 numerical features that model fluency and readability of the writing: *word count, sentence count, character count, average sentence length, average word length*. While we do not have a direct model for writing fluency, we use essay length features as an estimate because it is believed that a more fluent writer will be able to write more (Klebanov et al. 2016). Readability features are adapted from an Automated Readability Index formula which involves average sentence length and average word length.[4]

Our second group of features (CONTENT) aim for modeling different aspects of writing mechanics including spelling errors, content-richness and sentence complexity: *number and percentage of spelling errors; number and percentage of stop-words; number and percentage of words found in the writing prompt; number and percentage of words found in the SAT 5000-words;*[5] *numbers of commas, semi-colons, and colons; numbers of question marks, exclamation marks and double quote symbols*.

AES models are evaluated using quadratic-weighted kappa (qwk) which is a standard measure in the AES literature (Shermis and Burstein 2013). We observe that our baseline is more competitive than EASE in this corpus. EASE achieves 10-fold $qwk = 0.447$, which is lower than our baseline's $qwk = 0.599$ (see Table 5). Indeed, EASE depends heavily on n-gram features and was designed to score essays of the same prompt which is not the case in this data. That is why we create our own AES baseline for TOEFLL11 data. While the utilized features are simple, they yield competitive performance as shown in our next experiments.

Our End-to-End Argument Mining Pipeline

With the motivation of building an argument mining system that works for a wide variety of persuasive essays, we employ the argument mining corpus in (Stab and Gurevych 2017) for training our system. Following the *Macro-structure of Argument* theory (Freeman 1991), the authors proposed an argument annotation scheme which assumes argumentation structures as trees, where each argument consists of components (nodes) linked through argumentative relations (directed edges). The corpus consists of 402 persuasive essays which are practice writings in response to sample test questions of standardized English tests for ESL learners. In the essays, writers state their opinions (labeled as *MajorClaim*) towards the writing topics and validate those opinions with convincing arguments consisting of controversial statements (*Claim*) that support or attack the Major Claims, and evidences (*Premise*) that underpin the validity of the Claims. Expert annotators were asked to identify such argument components in the essays, and direct argumentative relations, i.e., support vs. attack, between Premises, Premises to Claims, and Claims to Major Claims.

Our argument mining system implements the pipeline paradigm as depicted in Figure 1. We improve the argument component identification (ACI) model in (Stab and Gurevych 2017) with features derived from an argument and domain word lexicon pre-compiled in (Nguyen and Litman 2015). In particular, our features identify whether the current token, the one preceding and the one following are argument, domain or stop words. For argument component classification (ACC) and support relation identification, we implement our models in (Nguyen and Litman 2016b; 2016a). We follow (Stab and Gurevych 2017) to split the corpus into training and test sets. Our argument mining system achieves F-measure score (F1) on the test set (with true-label input): ACI's F1 = 0.872, ACC's F1 = 0.825, and F1 = 0.730 for support vs. not-support classification. (Stab and Gurevych 2017) achieved the state of the art with a joint model and reported ACI's F1 = 0.867, ACC's F1 = 0.826, and support vs. attack classification F1 = 0.680.

Note that our end-to-end argument mining system is neither trained nor tested on ASAP or TOEFLL11, as these corpora have only been annotated with holistic scores and not argumentative structures. The two AES datasets expose significant differences in writing topic, style and quality compared to our argument mining corpus. Thus we can neither conclude how well our argument mining system performs on AES data nor reason about how AES performance of argumentative features relates to the output quality of argument mining.[6] Instead, we examine whether an off-the-shell argument mining system with decent experimental accuracy on an argument mining benchmark can yield features that improve automated essay scoring tasks in practice.

Argumentative Features for AES

From our argument mining output, we extract 33 features as described in Table 4. Because the relation model that we implemented only identifies in-paragraph support relations, we do not include argumentative features that involve attack relations or cross-paragraph argument component pairs.

For argument component (AC) features, we use raw counts as well as the ratios of argument components and argumentative sentences (i.e., sentences that contain at least one argument component) over the total number of sentences in the essay. Numbers of argument components and argumentative sentences were widely used in prior studies (Ghosh et al. 2016; Klebanov et al. 2016). Our preliminary analysis found moderate correlations ($r > 0.7$) between number of argument components (also argumentative sentences) and essay length (i.e., word and sentence counts). Therefore, argument count features are expected to simulate the effect of essay length features.

(Wachsmuth, Al Khatib, and Stein 2016) hypothesized that essays largely argue sequentially, so they restricted to

[4]https://en.wikipedia.org/wiki/Automated_readability_index

[5]http://www.freevocabulary.com

[6]For example, in Figure 1, our ACI model failed to recognize some organizational elements, e.g., "in my opinion" and "then".

	Argument component features (AC)
1, 2	Number and fraction of argument components over total number of sentences in essay (Ghosh et al. 2016)
3, 4	Number and fraction of argumentative sentences (Ghosh et al. 2016)
5	Total number of words in argument components (*this study*)
6	Number of paragraphs containing argument components (Persing and Ng 2015)
7	Whether the essay has paragraph without any argument component (Persing and Ng 2015)
	Component label features (CL)
8	Number of Major Claims (*this study*)
9, 10	Number and fraction of Claims over total number of sentences (Persing and Ng 2015; Ghosh et al. 2016)
11, 12	Number and fraction of Premises (Persing and Ng 2015; Ghosh et al. 2016)
13	Average number of Premises per Claim (Klebanov et al. 2016)
	Argument flow features (AF)
14	Number of paragraphs that contain Major Claims and Claims (Persing and Ng 2015)
15	Number of paragraphs that contain Major Claims and Premises (*this study*)
16	Number of paragraphs that contain Claims and Premises (*this study*)
17–24	Frequency of 8 typed bigrams of argument components (*this study*)
	Argumentative relation features (RL)
25	Number of supported Claims (Ghosh et al. 2016)
26	Number of dangling Claims (Ghosh et al. 2016)
27	Number of supporting Premises (*this study*)
28	Number of paragraphs that have support relations (*this study*)
	Argumentation structure typology features (TS)
29	Number of *Chain*-structures (Ghosh et al. 2016)
30	Number of *Tree*-structures (*this study*)
31	Number of *Tree*-structures with height = 1 (Ghosh et al. 2016)
32	Number of paragraphs that contain *Chain*-structures (*this study*)
33	Number of paragraphs that contain *Tree*-structures (*this study*)

Table 4: Argumentative features for essay score prediction.

sequences of types (i.e., Thesis, Conclusion, Premise) of argumentative discourse units (i.e., argument flow) in paragraphs to mine reliable patterns of argumentation structure of persuasive essays. For example, argument flows (Conclusion, Premise) and (Conclusion, Premise, Premise) are found to be the most frequent in the International Corpus of Learner English (ICLE) (Granger et al. 2009). We adapt their idea to extract typed-bigrams of argument components from paragraphs of essays to use as features. With three possible argumentative labels: MajorClaim, Claim and Premise, we have 9 possible typed bigrams. We do not consider the MajorClaim–MajorClaim bigrams which do not hold an argumentative relation, and retain 8 remaining typed bigrams. Also, we count number of paragraphs that have simultaneously MajorClaim and Claim, Claim and Premise, or MajorClaim and Premise.

For argumentative relation features, we count Claims that are supported by Premises, dangling Claims which are not supported by any Premises, and Premises that support Claims. As noted above, we do not have features representing if major claims are supported or attacked. Argumentation structure typology features (TS) were proposed in (Ghosh et al. 2016). The authors constructed a directed acyclic graph of support relations for each paragraph, and defined three argumentation structure typologies: *Chain*-structure, *Tree*-structure of height > 1 ($Tree_{h>1}$), and *Tree*-structure of height = 1 ($Tree_{h=1}$). Typology features are essentially different from argument flow features. While the former requires the existence of support relations, the other merely considers the appearance order of argument components. Due to the rare occurrence of $Tree$-structures in essays (Wachsmuth, Al Khatib, and Stein 2016), we group $Tree_{h=1}$ and $Tree_{h>1}$-structures together.

Because essays of ASAP data do not have paragraphs, there are two impacts on our AES models for this data. First, our argument mining models cannot utilize paragraph-position features, e.g., first vs. last paragraph indicators, as depicted in (Nguyen and Litman 2016b; 2016a). Thus, the argument mining performance may reduce for ASAP essays. Second, our AES models cannot have paragraph-related features in Table 4, which reduces number of argumentative features for ASAP essays to 25.

Persuasive Essay Score Prediction Results

Given the baseline models for AES (BASE), our experiments evaluate whether performance can be improved by adding argumentative features (ARG). We conduct both cross-fold and cross-prompt validations. Our experiments successively increase the difficulty of the essay scoring test by increasing the difference between the training and test data. When possible, performance of our AES models are compared with the best results in the literature (LIT). We also report hold-out results on the test set of TOEFL11 data to directly compare with (Klebanov et al. 2016).

Cross-fold validation

For TOEFL11 data, we use the training set and conduct 10-fold cross validation. Regarding ASAP data, we perform 5×5-fold cross validation on each set so we can compare

	ASAP 1	ASAP 2	TOEFL11
LIT	0.821	0.688	–
BASE	0.831	0.680	0.599
ARG	0.790 †	0.620 †	0.494 †
All	0.830	**0.689** *	**0.611** *
All – AC	0.830	0.676	0.610 *
All – CL	**0.832**	**0.689** *	0.608 *
All – AF	0.830	0.688 *	0.604
All – RL	0.831	0.687	0.606 *
All – TS	0.831	0.688	**0.611** *

Table 5: AES cross-fold qwk of different data sets. Symbols * and † indicate significantly higher and lower than BASE values ($p < 0.05$). Best models are highlighted in boldface.

	ASAP	TOEFL11
LIT	0.569	–
BASE	0.585	0.591
ARG	0.567	0.492 †
All	0.622	0.600 *
All – AC	0.610	0.600 *
All – CL	0.596	0.600 *
All – AF	0.611	**0.601** *
All – RL	**0.626**	0.595 *
All – TS	0.622	**0.601** *

Table 6: AES cross-prompt qwk of different data sets

with 5-fold cross validation in prior studies without knowing the data folds. AES models are trained using the Logistic Regression algorithm in Weka (Hall et al. 2009). By not setting ridge regularization, we wanted to have all features of a set included in the training process to obtain a fair performance representing collaborative effectiveness of features in the set. Cross-fold validation qwk's are reported in Table 5. First of all, while using only argumentative features (ARG) performed significantly worse than BASE features, combining ARG with BASE (All) significantly improved BASE performance in TOEFL11 and ASAP prompt 2 essays. Using argumentative features in ASAP prompt 1 data did not gain improvement in AES performance.

Regarding the ablation test, we see that argumentative features performed differently in different data sets. In TOEFL11 data, removing AF features decreased performance the most, while in ASAP prompt 2 data, it was AC feature set. Removing TS features did not yield performance loss in any of the data sets. TS features were shown helpful in (Ghosh et al. 2016) but it might reflect the high performance of their argumentative relation classification. In fact, the authors ran AES experiments on argument-annotated data and solved argumentative relation classification using true component labels. Our experiments are more challenging when AES data is unseen by the argument mining system, and the argumentative relation model has to rely on (noisy) prediction output of two argument component models. We hypothesize that our TS features might be less useful due to a less accurate argumentative relation model.

(Taghipour and Ng 2016) achieved the state-of-the-art for ASAP data by combining different neural network architectures. 5-fold cross validation results of their best system are reported in row LIT of the table. In all different 5-fold runs, our BASE performed better than LIT in ASAP prompt 1, but worse in ASAP prompt 2. Combining ARG with BASE yields a comparable AES model to LIT in ASAP prompt 2 essays. In conjunction with TOEFL11, our results demonstrate the benefit of argumentative features in AES tasks.

Cross-prompt validation

In this experiment, we conduct cross-prompt validation within each of the AES corpora: TOEFL11 and ASAP. This experiment offers a more difficult evaluation than k-fold cross validation because now training and test essays are from disjoint writing prompts; further, the ASAP data even has different essay score ranges for each prompt. We expect that argumentative features which abstract over the argument content and argumentative structure of the writing will work effectively even in cross-prompt AES. Again, our AES models are trained using Logistic Regression.

Using TOEFL11 training data, we iteratively use each combination of 7 out of 8 prompts to train a model and test with the remaining prompt. A number of our prior cross-fold validation findings are also confirmed in this cross-prompt validation. As shown in the right column of Table 6, the AES improvements by adding argumentative features (All) are again significant. Now even all ablated sets yield significant improvements, demonstrating the topic-independent advantage of argumentative features when given more difficult test data. Also in the ablation test, the best performance is obtained when removing TS or AF. However, in the cross-fold setting, removing AF features had negative impact. These findings expose a need for argumentative feature selection to optimize for different experimental settings.

Our interest with ASAP data is to compare with recent domain-adaptation AES studies (Phandi, Chai, and Ng 2015; Dong and Zhang 2016). While we are not developing a domain-adaptation AES algorithm, given the results in TOEFL11 data, we think that our investigation on using argumentative features to improve cross-prompt score prediction may contribute to the advancement of the problem.

Following (Phandi, Chai, and Ng 2015), we use essays of prompt 1 for training and prompt 2 for testing. Because essays of the two prompts have different score ranges, essay scores were scaled to an intermediate range [-1, 1] for training and testing essay score regression models. Then, predicted values are re-scaled back to the score range of test essays so that qwk can be computed. (Dong and Zhang 2016) developed a neural network AES model and further improved the domain-adaptation AES. Our results and the best in (Dong and Zhang 2016) (LIT) are shown in Table 6.

First of all, our use of EASE (BASE) obtained higher qwk than the prior studies when the training and test sets are prompts 1 and 2. This is probably because the models in prior studies were optimized for the best average performance across different training/test pairs. Therefore, results in Table 6 are not evidence to conclude that EASE is gen-

erally better than domain-adaptation algorithms proposed in prior studies. However, because the main focus of our current study is the impact of argumentative features in cross-prompt AES, using a learning algorithm that is particularly good for the data of interest gives us an ideal context.

In our cross-prompt experiment with ASAP data, we again observe that adding argumentative features (All) improves AES performance. Due to only a single test set, there is no significance analysis. In the ablation tests, the best qwk is obtained when removing RL features. However, removing any of AC, CL, and AF feature sets decreases qwk. Experimenting with other combinations of argumentative feature sets, we could further improve qwk up to 0.649 when adding AC, CL, and TS features to BASE. Both argument component and argumentative relation features (needed to compute structure) are present in the best set, which shows the necessity of complete argument mining.

As with our cross-fold experiment, we observe that the best sets of argumentative features do not generalize across AES and TOEFL11 data. We hypothesize that argument mining accuracy and interactions between argumentative and baseline features determine which classes of argumentative features are more effective. This suggests that feature selection is a necessary task-specific practice when deploying argument mining for AES.

In sum, both the cross-fold and cross-prompt validations give clear evidence of performance gain by argumentative features for AES tasks. Our finding is stronger support for automated persuasive essay scoring in practice than prior studies because it confirms that performance improvements are significant even when the base AES model is competitive and the argument mining is fully automated.

Our cross validation experiments did not do an exhaustive feature selection, but aimed for evaluating argumentative features by groups to get an insight of how possible outputs of argument mining can help improve AES. When comparing results in n-fold cross validation and cross-prompt validation (not included in the tables), argument typology features (TS) performed the worst when used alone, and contributed least when added to the base model. In future work, we plan to improve argumentative relation mining with joint prediction and study if relation-based features (i.e., RL, TS) can be more effective.

Test performance on TOEFL11 data

Our last experiment follows the procedure in (Klebanov et al. 2016) in which AES models are evaluated using TOEFL11 training and test sets. This allows us to directly compare our results with the prior study. For the best performance of the base AES model (BASE), we conduct 10-fold cross validation in the training set to compare different learning algorithms. The result shows that Random Forest algorithm works the best.

Our BASE model achieved qwk 0.604 in the test set, which is higher than the best model in (Klebanov et al. 2016) which combined word count with 9 argument structure features and obtained qwk 0.540. Adding all argumentative features sightly improves qwk to 0.607. When using the best combination of features (i.e., {AC, CL, RL, AF}) which was determined by ablation test with cross-fold validation (Table 5), we obtained the second best result in this experiment: $qwk = 0.618$. Indeed, the best qwk = 0.622 is achieved when adding all argumentative features except RL to BASE model.

Overall, the test results again confirm our prior findings of the value of argumentative features for AES, and that the best set of features is an open problem and may need extensive studies to determine for different use cases.

Conclusions and Future Work

With recent achievements of argument mining in text, using argument mining to improve automated essay assessment is becoming more realistic. In this paper, we investigated the value of argument mining for automated persuasive essay scoring by addressing important limitations of prior work. First, by building an end-to-end argument mining pipeline, we made argumentative feature extraction fully-automated. In fact, our argument mining models were trained using an argument mining corpus different than the corpora used for our AES tasks. Second, across cross-fold, cross-prompt, and train-test experiments with two holistic AES corpora, our results demonstrated that argument mining could improve AES performance compared to very competitive baselines. We hypothesize that argumentative features explore an additional space than the generic statistical and lexical features traditionally used in AES. Third, we provided insights on the robustness of different argumentative feature groups.

While we have built a novel end-to-end argument mining system to support this study, argument mining performance is not our main focus here. In the future, we want to expand our AES research with other argument mining systems and compare these systems in terms of how they benefit AES. We expect that with more accurate argument mining models, argumentative features can be even more effective. In addition, we would like to move from automated essay scoring to an automated writing evaluation system that can provide feedback, where argumentative features are used to explain the scores of argumentative essays.

References

Barker, E., and Gaizauskas, R. 2016. Summarizing Multi-Party Argumentative Conversations in Reader Comment on News. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 12–20. Berlin, Germany: Association for Computational Linguistics.

Blanchard, D.; Tetreault, J.; Higgins, D.; Cahill, A.; and Chodorow, M. 2013. TOEFL11: A Corpus of Non-native English. *ETS Research Report Series* 2013(2):i–15.

Boltužić, F., and Šnajder, J. 2014. Back up your Stance: Recognizing Arguments in Online Discussions. In *Proceedings of the First Workshop on Argumentation Mining*, 49–58. Baltimore, Maryland: Association for Computational Linguistics.

Dikli, S. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment* 5(1).

Dong, F., and Zhang, Y. 2016. Automatic Features for Essay Scoring – An Empirical Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1072–1077. Austin, Texas: Association for Computational Linguistics.

Egan, C.; Siddharthan, A.; and Wyner, A. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 134–143. Berlin, Germany: Association for Computational Linguistics.

Freeman, J. B. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. Foris Publications.

Ghosh, D.; Khanam, A.; Han, Y.; and Muresan, S. 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)*, 549–554. Berlin, Germany: Association for Computational Linguistics.

Granger, S.; Dagneaux, E.; Meunier, F.; and Paquot, M. 2009. *International Corpus of Learner English v2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter* 11(1):10–18.

Klebanov, B. B.; Stab, C.; Burstein, J.; Song, Y.; Gyawali, B.; and Gurevych, I. 2016. Argumentation: Content, Structure, and Relationship with Essay Quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 70–75. Berlin, Germany: Association for Computational Linguistics.

Levy, R.; Bilu, Y.; Hershcovich, D.; Aharoni, E.; and Slonim, N. 2014. Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1489–1500.

Mochales, R., and Moens, M.-F. 2011. Argumentation mining. *Artificial Intelligence and Law* 19(1):1–22.

Nguyen, H., and Litman, D. 2015. Extracting Argument and Domain Words for Identifying Argument Components in Texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, 22–28. Denver, CO: Association for Computational Linguistics.

Nguyen, H., and Litman, D. 2016a. Context-aware Argumentative Relation Mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1127–1137. Berlin, Germany: Association for Computational Linguistics.

Nguyen, H., and Litman, D. 2016b. Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics. In *Proceedings 29th International FLAIRS Conference*.

Peldszus, A., and Stede, M. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7(1):1–31.

Peldszus, A., and Stede, M. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 938–948. Lisbon, Portugal: Association for Computational Linguistics.

Persing, I., and Ng, V. 2015. 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*, 543–552. Beijing, China: Association for Computational Linguistics.

Persing, I., and Ng, V. 2016. End-to-End Argumentation Mining in Student Essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1384–1394. San Diego, California: Association for Computational Linguistics.

Phandi, P.; Chai, K. M. A.; and Ng, H. T. 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*, 431–439. Lisbon, Portugal: Association for Computational Linguistics.

Shermis, M. D., and Burstein, J. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Stab, C., and Gurevych, I. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 46–56. Doha, Qatar: Association for Computational Linguistics.

Stab, C., and Gurevych, I. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43(3):619–659.

Taghipour, K., and Ng, H. T. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891. Austin, Texas: Association for Computational Linguistics.

Wachsmuth, H.; Al Khatib, K.; and Stein, B. 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*, 1680–1691. Osaka, Japan: The COLING 2016 Organizing Committee.

Wachsmuth, H.; Naderi, N.; Hou, Y.; Bilu, Y.; Prabhakaran, V.; Thijm, T. A.; Hirst, G.; and Stein, B. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 176–187. Valencia, Spain: Association for Computational Linguistics.