CONTEXT-AWARE ARGUMENT MINING AND ITS APPLICATIONS IN EDUCATION

by

Huy V. Nguyen

Bachelor of Engineering

Hanoi University of Sciences and Technologies, Vietnam

2007

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Huy V. Nguyen

It was approved by

Diane J. Litman, Department of Computer Science

Rebecca Hwa, Department of Computer Science

Adriana I. Kovashka, Department of Computer Science

Kevin D. Ashey, School of Law

Dissertation Director: Diane J. Litman, Department of Computer Science

ABSTRACT

CONTEXT-AWARE ARGUMENT MINING AND ITS APPLICATIONS IN EDUCATION

Huy V. Nguyen, PhD

University of Pittsburgh, 2016

Context is crucial for identifying argument and argumentative relations in text, but existing argument studies have not addressed context dependence adequately. In this thesis, we propose *context-aware argument mining* that makes use of contextual features extracted from writing topics and context sentences to improve state-of-the-art argument component identification and argumentative relation classification. The effectiveness as well as generality of our proposed contextual features is proven through its application in different argument mining tasks in student essays. We further evaluate the applicability of our proposed argument mining models in an automated essay scoring task.

Keywords: argument mining, context segment, automated essay scoring.

TABLE OF CONTENTS

1.0 INTRODUCTION

Argumentation can be defined as a social, intellectual, verbal activity serving to justify or refute an opinion, consisting of statements directed towards obtaining the approbation of an audience. Originally proposed within the realms of Logic, Philosophy, and Law, computational argumentation has become an increasingly central core study within Artificial Intelligence (AI) which aims at representing components of arguments, and the interactions between components, evaluating arguments and distinguishing legitimate from invalid arguments [Bench-Capon and Dunne, 2007].

With the rapid growth of textual data and tremendous advances in text mining, argument (argumentation) mining in text[1] has apparently been an emerging research field that is to draw a bridge between formal argumentation theories and everyday life argumentative reasoning. Aiming at automatically identifying argument components (e.g., premises, claims, conclusions) in natural language text, and the argumentative relations (e.g., support, attack) between components, argument mining is found to promise novel opportunities for opinion mining, automated essay evaluation as well as offers great improvement for current legal information systems or policy modeling platforms. Argument mining has been studied in a variety of text genres like legal documents [Moens et al., 2007, Mochales and Moens, 2008, Palau and Moens, 2009], scientific papers [Teufel and Moens, 2002, Teufel et al., 2009, Liakata et al., 2012], news articles [Palau and Moens, 2009, Goudas et al., 2014, Sardianos et al., 2015], user-generated online comments [Cabrio and Villata, 2012, Boltužić and Šnajder, 2014], and student essays [Burstein et al., 2003, Stab and Gurevych, 2014b, Rahimi et al., 2014, Ong et al., 2014]. Problem formulations of argument mining have ranged from the separation of argumentative from non-argumentative text, the classification of argument components and

[1] Argument mining for short.

Essay 75: (0)Do arts and music improve the quality of life?

(1)My view is that the [*government should give priorities to invest more money on the basic social welfares such as education and housing instead of subsidizing arts relative programs*]$_{MajorClaim}$.

(2)[*Art is not the key determination of quality of life, but education is*]$_{Claim}$. (3)[*In order to make people better off, it is more urgent for governments to commit money to some fundamental help such as setting more scholarships in education section for all citizens*]$_{Premise}$. (4)This is simply because [*knowledge and wisdom is the guarantee of the enhancement of the quality of people's lives for a well-rounded social system*]$_{Premise}$.

(5)Admittedly, [*art, to some extent, serve a valuable function about enriching one's daily lives*]$_{Claim}$, for example, [*it could bring release one's heavy burden of study pressure and refresh human bodies through a hard day from work*]$_{Premise}$. (6)However, [*it is unrealistic to pursuit of this high standard of life in many developing countries, in which the basic housing supply has still been a huge problem with plenty of lower income family have squeezed in a small tight room*]$_{Premise}$. (7)By comparison to these issues, [*the pursuit of art seems unimportant at all*]$_{Premise}$.

(8)To conclude, [*art could play an active role in improving the quality of people's lives*]$_{Premise}$, but I think that [*governments should attach heavier weight to other social issues such as education and housing needs*]$_{Claim}$ because [*those are the most essential ways enable to make people a decent life*]$_{Premise}$.

Figure 1: A sample student essay taken from the corpus in [Stab and Gurevych, 2014a]. The essay has sentences numbered and argument components enclosed in tags for easy look-up.

argumentative relations, to the identification of argumentation structures/schemes.

To illustrate different tasks in argument mining, let us consider a sample student essay in Figure 1. The first sentence in the example is the writing prompt. The *MajorClaim* which states the author's stance towards the writing topic is placed at the first of the essay's body, i.e., sentence 1. The student author used different *Claims* (controversial statements) to validate/support and attack the major claim, e.g., claims in sentences {2, 5, 8}. Validity of the claims are underpinned/rebutted by *Premises* (reasons provided by the author), e.g., premises in sentences {5, 6, 7}. As the first task in argument mining, *Argument Component Identification* aims at recognizing argumentative portions in the text (Argumentative Discourse Units – ADUs [Peldszus and Stede, 2013]), e.g., a subordinate clause in sentence 1, or the whole sentence 2, and classifying those ADUs accordingly to their argumentative

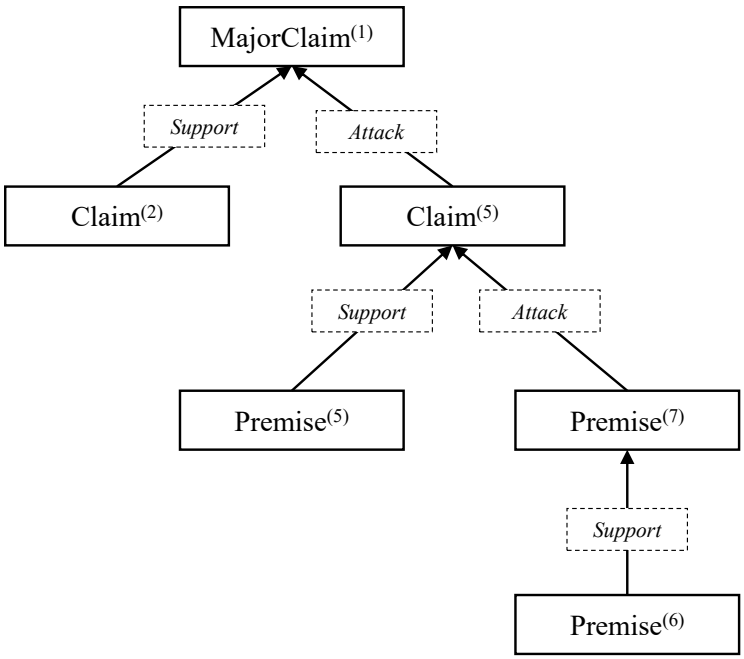

Figure 2: Graphical representation of a part of argumentation structure in the example essay. Argumentative relations are illustrated based on annotation by [Stab and Gurevych, 2014a].

roles, e.g., *MajorClaim*, *Claim*, and *Premise*. The two sub-tasks are often combined into a multi-way classification problem by introducing the *None* class. Thus, possible class labels for a candidate ADU are {*MajorClaim*, *Claim*, *Premise*, *None*}. However, determining boundaries of candidate ADUs to prepare input for argument mining models is a nontrivial preprocessing task. In order to simplify the main argument mining task, sentences are usually taken as primary units [Moens et al., 2007], or the gold-standard boundaries are assumed available [Stab and Gurevych, 2014b].

The second task, *Argumentative Relation Classification* [Stab and Gurevych, 2014b], considers possible pairs of argument components in a definite scope, e.g., paragraph,[2] or pairs of argument component and argument topic. For each pair, determines if a component supports or attacks the other component. As we have in the example essay, the *Claim* in

[2]The definite scope is necessary to make the distribution less skewed. In fact, the number of pairs that hold an argumentative relation is far smaller than the total number of possible pairs.

sentence 2 supports the *MajorClaim* in sentence 1: $Support(Claim^{(2)}, MajorClaim^{(1)})$. We also have $Attack(Claim^{(5)}, MajorClaim^{(1)})$, $Support(Premise^{(5)}, Claim^{(5)})$. Given the direct relations as in examples, one can infer $Attack(Premise^{(5)}, MajorClaim^{(1)})$ and so on.

While in argumentative relation classification one does not differentiate direct and inferred relations, *Argumentation Structure Identification* [Mochales and Moens, 2011] aims at constructing the graphical representation of argumentation in which edges are direct attachments between argument components. Attachment is an abstraction of support/attack relations, and is illustrated as arrowhead connectors in Figure 2. Attachment between argument components does not necessarily correspond to the components' relative positions in the text. For example, $Premise^{(6)}$ is placed between $Claim^{(5)}$ and $Premise^{(7)}$ in the essay, but $Premise^{(7)}$ is the direct premise of $Claim^{(5)}$ as shown in the figure.

1.1 AN OVERVIEW OF OUR THESIS WORK

In education, teaching argumentation and argumentative writing to student are in particular need of attention [Newell et al., 2011, Barstow et al., 2015]. Automated essay scoring (AES) systems have been proven effective to reduce teachers' workload and facilitate writing practices, especially in large-scale [Shermis and Burstein, 2013]. AES research has recently showed interest in automated assessment of different aspects of written arguments, e.g., evidence [Rahimi et al., 2014], thesis and argument strength [Persing and Ng, 2013, Persing and Ng, 2015]. However, the application of argument mining in automatically scoring argumentative essays has been studied limitedly [Ong et al., 2014, Song et al., 2014]. Motivated by the promising application of argument mining as well as the desire of automated support for argumentative writings in school, our research aims at building models that automatically mines arguments in natural language text, and applying argument mining outcome to automatically scoring argumentative essays. In particular, we propose *context-aware argument mining models* to improve state-of-the-art argument component identification and argumentative relation classification. In order to make the proposed approaches more applicable to the educational context, our research conducts both intrinsic and extrinsic evaluation when

comparing our proposed models to the prior work. Regarding intrinsic evaluation, we perform both random folding cross validation and cross-topic validation to assess the robustness of models. For extrinsic evaluation, our research investigates the uses of argument mining for automated essay scoring. Overall, our research on argument mining can be divided into three components with respect to their functional aspects.

1.1.1 Context-aware Argument Mining Models

The main focus of our research is building models for argument component identification and argumentative relation classification. As illustrated in [Stab and Gurevych, 2014a], context[3] is crucial for identifying argument components and argumentation structures. However, context dependence has not been addressed adequately in prior work [Stab et al., 2014]. Most of argument mining studies built prediction models that process each textual input[4] isolatedly from the surrounding text. To enrich the feature space of such models, history features such as argumentative roles of one or more preceding components, and features extracted separately from preceding and/or following text spans have been usually used [Teufel and Moens, 2002, Hirohata et al., 2008, Palau and Moens, 2009, Guo et al., 2010, Stab and Gurevych, 2014b]. However, the idea of using surrounding text as a context-rich representation of the prediction input for feature extraction was studied limitedly in few research [Biran and Rambow, 2011].

In many writing genres, e.g., debates, student essays, scientific articles, the availability of writing topics provides valuable information to help identify argumentative text as well as classify their argumentative roles [Teufel and Moens, 2002, Levy et al., 2014]. Especially, [Levy et al., 2014] defined the term Context Dependent Claim to emphasize the role of discussion topic in distinguishing claims relevant to the topic from the irrelevant statements. The idea of using topic and discourse information to help resolve ambiguities are commonly used in word sense disambiguation and sentiment analysis [Navigli, 2009, Liu,

[3]The thesis differentiates between global context and local context. While global context refers to the main topic/thesis of the document, the local context is instantiated by the actual text segment covering the textual unit of interest, e.g., preceding and following sentences.

[4]E.g., candidate ADU in argument component identification, or pair of argument components in argumentative relation classification.

2012]. Based on these observations, we hypothesize that argument component identification and argumentative relation classification can be improved with respect to prediction performance by considering contextual information at both local and global levels when developing prediction features.

Definition 1. *Context segment of a textual unit is a text segment formed by neighboring sentences and the unit itself. The neighboring sentences are called context sentences, and must be in the same paragraph with the textual unit.*

Instead of building prediction models that process each textual input isolatedly, our context-aware approach considers the input within its *context segment*[5] to enable advanced contextual features for argumentative relation classification. In particular, our approach aims at extracting discourse relations within the context segment to better characterize the rhetorical function of the unit in the entire text. Besides, the context segments instead of their units will be fed to textual entailment and semantic similarity scoring functions to extract semantic relation features. We expect that a score set by possible pairs extracted from two segments better represents the semantic relations of the two input units than their single score. As defining the context and identifying boundaries of context segment are not a focus of our research, we propose to use different heuristics, e.g., window-size, topic segmentation, to approximate the context segment given a textual unit, and evaluate contribution of such techniques to the final argument mining performance.

Definition 2. *Argument words are words that signal the argumentative content, and commonly used across different argument topics, e.g., 'believe', 'opinion'. In contrast, domain words are specific terminologies commonly used within the topic, e.g., 'art', 'education'. Domain words are a subset of content words that form the argumentative content.*

As of a use of global context, we propose an approach that uses writing topics to guide a semi-supervised process for separating *argument words* from *domain words*.[6] The extracted

[5]Term "context sentences" was used in [Qazvinian and Radev, 2010] to refer sentences surrounding a citation, that contain information about the cited source but do not explicitly cite it. In this thesis, we place no other constrains to context sentences than requiring them to be adjacent to the textual unit.

[6]Our definition of argument and domain words shares similarities with the idea of shell language and content in [Madnani et al., 2012] in that we aim to model the lexical signals of argumentative content. However while Madnani et al. emphasized the boundaries between argument shell and content, we do not require such a physical separation between the two aspects of an argument component.

vocabularies of argument words and domain words are then used to derive novel features and constraints for an argument component identification model.

1.1.2 Intrinsic Evaluation: Cross-validation

In educational settings, students can have writing assignments in a wide range of topics. Therefore a desired argument mining model that has practical application in student essays is the one that can yield good performance for new essays of different topic domains than those of the training essays. As a consequence, features which are less topic-specific will be more predictive when cross-topic evaluated. Given this inherent requirements to the argument mining tasks for student essays, our research emphasizes the evaluation of the robustness of argument mining models. In addition to random-fold cross-validation (i.e., training and testing data are randomly split from the corpus), we also conduct cross-topic validation (i.e., training and testing data are from essays of different writing topics [Burstein et al., 2003]) when comparing the proposed approaches with prior studies.

Beyond cross-topic evaluation, our research also uses different corpora to evaluate effectiveness of the proposed approaches. The first corpus consists of persuasive essays and the associated coding scheme specifies three different types of argument components: major claim, claim, and premise [Stab and Gurevych, 2014a]. The second corpus are academic writings collected from college Psychology classes and has sentences classified based on their argumentative roles: hypothesis, support finding, opposition finding, or non-argumentative [Barstow et al., 2015].

1.1.3 Extrinsic Evaluation: Automated Essay Scoring

Aiming at high performance and robust models of argument mining, the second goal of our research is to seek for an application of argument mining in automated argumentative essay evaluation. As proposed in the literature, an direct approach would be using prediction outcome (e.g., arguments identified by prediction models) to recall students' attention to not only the organization of their writings but also the plausibility of the provided arguments in the text [Burstein et al., 2004, Falakmasir et al., 2014]. Such feedback information also

helps teachers quickly evaluate writing performance of their students for better instructions. However, deploying an argument mining model to an existing computer-supported writing service, and evaluate it benefit to student learning would require a great amount of time and effort. Thus, it is set up as the long-term goal of our research. In the course of this thesis, we instead look for answers to the question whether the outcome of automated argument mining can predict essay scores.

For this goal, our research uses two corpora to conduct automated essay scoring experiments. The first corpus is the academic essays that were used for our argument mining experiments. Each essay in the corpus was reviewed by student peers, and was given both textual comments and numerical ratings by its peer reviewers. Therefore our research makes use of peer ratings as the gold standard for the essay scoring experiment. The second corpus is the Argument Strength Corpus, in which argumentative student essays were annotated with argument strength scores [Persing and Ng, 2015]. The argumentative essays of this corpus have certain similarities with the persuasive essays in the [Stab and Gurevych, 2014a] which are used for our argument mining study. Besides, both two corpora were originally used for automated essay scoring studies, thus the prior scoring models are perfect baselines to evaluate our proposed approach. In this research we employ two approaches for applying argument mining to automated essay scoring. The first approach simply uses statistics of argument components and argumentative relations identified by our argument mining models to train a scoring prediction model [Ong et al., 2014]. The second approach uses those statistics to augment the scoring model in [Persing and Ng, 2015].

1.2 THESIS STATEMENTS

Motivated by the benefit of contextual information from writing topics and context segments in argument mining, we propose context-aware argument mining that make use of additional context features derived from such contextual information. In this thesis, we aim to prove the following hypotheses of *the effectiveness of our proposed context features*:

- **H1**. Our proposed context features helps improve the argument mining performance.

This hypothesis is divided into two sub-hypotheses:

  - **H1-1**. Adding the context features improves the argument component identification in student essays in cross-fold and cross-topic validations. This hypothesis is proven in §3 and §4.
  - **H1-2**. Adding the context features improves the argumentative relation classification in student essays in cross-fold and cross-topic validations. This hypothesis will be tested in §5.

- **H2**. Prediction output of our proposed argument component identification and argumentative relation classification models for student essays improve automated argumentative essay scoring. This hypothesis will be tested in §6.

1.3 PROPOSAL OUTLINE

In the next chapter, we briefly discuss argument mining from its theoretical fundamentals to existing computational studies in different domains. Chapter 3 and 4 present our completed work on argument component identification. In Chapter 3, we present a novel algorithm to extract argument and domain words to use as new features and constraints for improving the argument component identification in student essays. Chapter 4 presents an evaluation of our proposed model for automated argument component identification in student essay using cross-topic validation. Chapter 5 and 6 describe our proposed work on argumentative relation classification in student essays and applying argument mining to automated argumentative essay scoring.

2.0 BACKGROUND

2.1 ARGUMENTATION THEORIES

From the ancient roots in dialectics and philosophy, models of argumentation have spread to core areas of AI including knowledge representation, non-monotonic reasoning, and multi-agent system research [Bench-Capon and Dunne, 2007]. This has given the rise of computational argumentation with two main approaches which are abstract argumentation and structured argumentation [Lippi and Torroni, 2015].[1] Abstract argumentation considers each argument as a primary element without internal structure, and focuses on the relation between arguments, or sets of them. In contrast, structured argumentation studies internal structure (i.e., argument components and their interaction) of argument that is described in terms of some knowledge representation formalism. Structured argumentation models are those typically employed in argument mining when the goal is to extract argument components from natural language. In this section, we describe two notable structured argumentation theories which are *Macro-structure of Argument* by [Freeman, 1991], and *Argumentation Scheme* by [Walton et al., 2008]. From the provided description of argumentation theories, we expect to give a concise yet sufficient introduction of related argument mining studies from a theoretical perspective.

Among a vast amount of structured argumentation theories have been proposed [Bentahar et al., 2010, Besnard et al., 2014], the *premise-conclusion* models of argument structure [Freeman, 1991, Walton et al., 2008] are the most commonly used in argument mining

[1] Abstract argumentation which is also called macro argumentation considers argumentation as a process. Structured argumentation, on the contrary, considers argumentation as a product and is also called micro argumentation [Mochales and Moens, 2011, Stab et al., 2014]

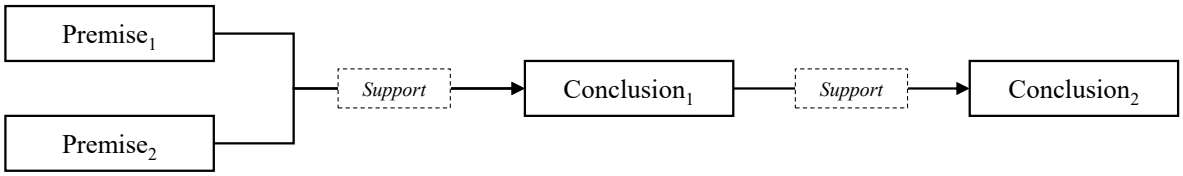

Figure 3: A complex macro-structure of argument consisting of linked structure (i.e., the support of Premise$_1$ and Premise$_2$ to Conclusion$_1$) ,and serial structure (i.e., the support of the two premises to Conclusion$_2$).

studies. In fact, the two corpora of argumentative writings that are studied in this thesis have coding schemes derived from the premise-conclusion structure of argument. [Walton et al., 2008] gave a simple and intuitive description of argument which specifies an argument as a set of statement consisting a conclusion, a set of premises, and an inference from the premises to the conclusion. In literature, claims are sometimes used as a replacement of conclusion, and premises are mentioned as evidences or reasons [Freeley and Steinberg, 2008]. The conclusion is the central component of the argument, and is what "we seek to establish by our argument" [Freeley and Steinberg, 2008]. The conclusion statement should not be accepted without additional reasons provided in premises. The second component of argument, i.e., premise, is therefore necessary to underpin the plausibility of the conclusion. Premises are "connected series of sentences, statements or propositions that are intended to give reason" for the conclusion [Freeley and Steinberg, 2008]. In a more general representation, premise can either support or attack the conclusion (i.e., giving reason or refutation) [Besnard and Hunter, 2008, Peldszus and Stede, 2013, Besnard et al., 2014]. Based on the premise-conclusion standard, argument mining studies have proposed different argumentative relation schemes to scope with the great diversity of argumentation in natural language text, for instances claim justification [Biran and Rambow, 2011], claim support vs. attack [Stab and Gurevych, 2014b], verifiability of support [Park and Cardie, 2014].

While premise-conclusion models do not differentiate functions of different premises[2], it

[2]Toulmin's argument structure theory [Toulmin, 1958] distinguishes the role of different types of premise, i.e., data, warrant, and backing, in the argument.

Argument from cause to effect

- *Major premise*: Generally, if A occurs, then B will (might) occur.
- *Minor premise*: In this case, A occurs (might occur).
- *Conclusion*: Therefore, in this case, B will (might) occur.

Critical questions

1. *Critique the major premise*: How strong is the causal generalization (if it is true at all)?
2. *Critique the minor premise*: Is the evidence cited (if there is any) strong enough to warrant to the generalization as stated?
3. *Critique the production*: Are there other factors that would or will interfere with or counteract the production of the effect in this case?

Figure 4: Argumentation scheme: Argument from Cause to Effect.

enables the *Macro-structure* of arguments which specifies the different ways that premises and conclusions combine to form larger complexes [Freeman, 1991].[3] For example, [Freeman, 1991] identified four main macro-structures of arguments: linked, serial, convergent, and divergent, to represent whether different premises contribute together, in sequence, or independently to one or multiple conclusions. An example of complex macro-structure of argument is shown in Figure 3. Based on Freeman's theory, [Peldszus and Stede, 2013] expand the macro-structure to cover more complex attack and counter-attack relations. In argument mining, the argumentation structure identification task aims at identifying the macro-structure of arguments in text [Palau and Moens, 2009, Peldszus and Stede, 2015].

Another notable construct of premise-conclusion abstraction is the *Argumentation Scheme Theory* [Walton et al., 2008]. The authors used the argumentation scheme notion to identify and evaluate reasoning patterns commonly used in everyday conversational argumentation, and other contexts, notably legal and scientific argumentation. In Argumentation Scheme Theory, arguments are instances of abstract argumentation schemes each of which requires premises, the assumption implicitly holding, and the exceptions that may undercut the argument. Each scheme has a set of critical questions matching the scheme and correspond to its

[3] In the *Macro-structure Structure of Argument Theory* the term 'argument' is thus not for premises, but for the complex of one or more premises put forward in favor of the conclusion.

premises, assumptions and exceptions, and such a set represents standard ways of critically probing into an argument to find aspects of it that are open criticism. Figure 4, illustrates the Argument-from-Cause-to-Effect scheme consisting of two premises and a conclusion. As we can realize argument schemes are distinguished by their content templates rather than their premise-conclusion structures. Identifying the argumentation scheme in the written argument has been considered to help recovering implicit premises and re-construct the full argument [Feng and Hirst, 2011]. On the other hand, research was also conducted to analyze the similarity and difference between argumentation schemes and discourse relations (i.e., Penn Discourse Treebank discourse relations [Prasad et al., 2008]) which is considered a fruitful support of automated argument classification and process [Cabrio et al., 2013].

2.2 ARGUMENT MINING IN DIFFERENT DOMAINS

Argument mining is a relatively novel research domain [Mochales and Moens, 2011, Peldszus and Stede, 2013, Lippi and Torroni, 2015] so its problem formulation is not well-defined but rather is considered potentially relevant to any text mining application that targets to argumentative text. Moreover, there is no consensus yet on an annotation scheme for argument components, or on the minimal textual units to be annotated. For these reasons, we follow [Peldszus and Stede, 2013] and consider in this study "argument mining as the automatic discovery of an argumentative text portion, and the identification of the relevant components of the argument presented there." We also borrow the term "argumentative discourse unit" [Peldszus and Stede, 2013] to refer the textual unit, e.g., text segment, sentences, clauses, which are considered as argument components.

In scientific domain, research has been long focusing on identifying the rhetorical status (i.e., the contribution to the overall text function of the article) of text segments, i.e., zone, to support summarization and information extraction of scientific publications [Teufel and Moens, 2002]. Different zone mining studies were also conducted for different scientific domains, e.g., chemistry, biology, and proposed different zone annotation schemes that targets the full-text or only abstract section of the articles [Lin et al., 2006, Hirohata et al.,

2008, Teufel et al., 2009, Guo et al., 2010, Liakata et al., 2012]. However, none of the zone mining models described local interactions across segments and thus the embedded argument structures in text are totally ignored. Despite this mismatch between zone mining and argument mining, the two areas solve a similar core problem which is text classification, which makes zone mining an inspiration of argument mining models.

Two other domains that have argument mining intensively studied are legal documents and user-generated comments. In legal domain, researchers seek for applications of automated recognition of arguments and argumentation structures in legal documents to support visualizing and qualifying arguments. A wide range of argument mining tasks have been studied including argumentative text identification [Moens et al., 2007], argument component classification (i.e., premise vs. conclusion), and argumentation structure identification [Mochales and Moens, 2008, Palau and Moens, 2009]. While the computational models for such argument mining tasks were evaluated using legal document corpora, those studies all employed the genre-independent premise-conclusion framework to represent the argument structure. Therefore many prediction features used in argument mining models for legal text, e.g., indicative keywords for argumentation, discourse connectives, are generally applicable to other argumentative text genres, e.g., student essays.

In user-generated comments, argument mining has been studied as a natural extension to opinion mining. While opinion mining answers what people think about for instance a product [Somasundaran and Wiebe, 2009], argument mining identifies reasons that explain the opinion. Among the first research on argument in user comments, [Cabrio and Villata, 2012] studied the acceptability of arguments in online debates by first determining whether two user comments support each other or not.[4] [Boltužić and Šnajder, 2014] extended the work by mining user comments for more fine-grained relations, i.e., {explicit, implicit} $\times$ {support, attack}. [Park and Cardie, 2014] addressed a different aspect of argumentative relation which is the verifiability of argumentative propositions in user comments. While the task does not solve whether the given proposition is a support or opposition of the debate topic, it provides a mean to analyze the arguments in terms of the adequacy of their support

[4]In their study, arguments are pros and cons user comments of the debate topic and were manually selected.

assuming support/attack propositions are labeled already.

Argument mining in student essays is rooted in argumentative discourse analysis for automated essay scoring [Burstein et al., 2003]. In argumentative[5] writing assignments, students are given a topic and asked to propose a thesis statement and justify support for the thesis. Oppositions are sometime required to make the thesis risky and nontrivial [Barstow et al., 2015]. Classifying argumentative elements in student essays has been used to support automated essay grading [Ong et al., 2014], peer review assistance [Falakmasir et al., 2014], and providing writing feedback [Burstein et al., 2004]. [Burstein et al., 2003] built a discourse analyzer for persuasive essays that aimed at identifying different discourse elements (i.e., sentence) such as for instance thesis, supporting idea, conclusion. Similarly, [Falakmasir et al., 2014] aimed at identifying thesis and conclusion statements in student writings, and used the prediction outcome to scaffold peer reviewers of an online peer review system. [Stab and Gurevych, 2014a] annotated persuasive essays using a domain-independent scheme specifying three types of argument components (major claim, claim, and premise) and two types of argumentative relations (support and attack). [Stab and Gurevych, 2014b] utilized the corpus for automated argument component and argumentative relation identification. [Ong et al., 2014] developed a rule-based system that labels each sentence in student writings in psychology classes an argumentative role, e.g., hypothesis, support, opposition, and found a strong relation between the presence of argumentative elements and essay scores. [Song et al., 2014] proposed to annotate argument analysis essays to identify responses of critical questions to judge the argument in writing prompts. The annotation were then used as novel features to improve an existing essay scoring model.

While studies in [Ong et al., 2014,Song et al., 2014] aimed at predicting the holistic score of the essays, research on automated essay scoring have recently investigated possibilities of grading essays on argument aspects, e.g., evidence [Rahimi et al., 2014], thesis clarity [Persing and Ng, 2013], and argument strength [Persing and Ng, 2015]. While these studies did not actually identified thesis statements or argument components in the essays, they provide strong baseline models as well as annotated data for research on application of argument mining on essay score prediction.

[5]The term "persuasive" was also used as an equivalent [Burstein et al., 2003,Stab and Gurevych, 2014a].

2.3 ARGUMENT MINING TASKS AND FEATURES

2.3.1 Argument Component Identification

To solve the argumentative label identification tasks (e.g., argumentative vs. not, premise vs. conclusion, rhetorical status of sentence), a wide variety of machine learning models has been applied ranging from classification models, e.g., Naive Bayes, Logistic Regression, Support Vector Machine (SVM), to sequence labeling models such as Hidden Markov Model (HMM), Conditional Random Field (CRF). Especially for zone mining in scientific articles, sequence labeling is a more natural approach given an observation that the flow of scientific writing exposes typical moves of rhetorical roles across sentences. Studies have been conducted to explore both HMM and CRF for automatically labeling rhetorical status of sentences in scientific publications using features derived from language models and relative sentence position [Lin et al., 2006, Hirohata et al., 2008, Liakata et al., 2012].

In the realm of argument mining, argument component identification studies have been focusing on deriving features that represent the argumentative discourse while being loyal to traditional classifiers such as SVM, Logistic Regression. Sequence labeling models were not used mostly due to the loose organization of natural language texts, e.g., student essays, user comments studied here. Prior studies have often used seed lexicons, e.g., indicative phrases for argumentation [Knott and Dale, 1994], discourse connectives [Prasad et al., 2008], to represent the organizational shell of argumentative content [Burstein et al., 2003, Palau and Moens, 2009, Stab and Gurevych, 2014b, Peldszus, 2014]. While the use of such lexicons shows effective, their coverage is far from efficient given the great diversity of argumentative writing in terms of both topic and style. Given the fact that the argumentative discourse consists of a language used to express claims, evidences and another language used to organize them, researchers have explored both supervised and unsupervised approaches to mine the organizational elements of argumentative text. [Madnani et al., 2012] used CRF to train a supervised sequence model using simple features like word frequency, word position, regular expression patterns. To leverage the availability of large amount of unprocessed data, [Séaghdha and Teufel, 2014] and [Du et al., 2014] built topic models based on LDA [Blei

et al., 2003] to learn two language models: topic language and shell language (rhetorical language, cf. [Séaghdha and Teufel, 2014]). While [Madnani et al., 2012] and [Du et al., 2014] used data which were annotated for shell boundaries to evaluate how well the proposed model separates shell from content, [Séaghdha and Teufel, 2014] showed that features extracted from the learned language models help improves a supervised zone mining model. In a similar vein, we post-process LDA output to extract argument and domain words which are used to improve the argument component identification.

In addition, contextual features were also applied to represent the dependency nature of argument components. The most popular are history features that indicate the argumentative label of preceding one or more components, and features extracted from preceding and following components [Teufel and Moens, 2002, Palau and Moens, 2009, Liakata et al., 2012, Stab and Gurevych, 2014b]. In many writing genres, e.g., debate, essay, scientific article, the availability of argumentative topics provide valuable information to help identify argumentative portions in text as well as classify their argumentative roles. [Levy et al., 2014] proposed the context-dependent claim detection task in which a claim is determine with respect to a given context - i.e., the input topic. To represent the contextual dependency, the authors made use of cosine similarity between the candidate sentence and the topic as a feature. For scientific writings, genre-specific contextual features were also considered including common words with headlines, section order [Teufel and Moens, 2002, Liakata et al., 2012]. As of context feature, we use writing topic to guide the separation of argument words from domain words. We also use common words with surrounding sentences and with writing topic as features.

2.3.2 Argumentative Relation Classification

The next step of identifying argument components is determining the argumentative relations, e.g., attack and support, between those components, or between arguments formed by those components. Research have explored different argumentative relation schemes that can be applied to pair of components, e.g., support vs. not [Biran and Rambow, 2011, Cabrio and Villata, 2012, Stab and Gurevych, 2014b], implicit and explicit support and attack [Boltužić

and Šnajder, 2014]. Because the instances being classified are pair of textual units, features usually involve information from both elements (i.e., source and target) of the pair (e.g., word pair, discourse indicators in source and target) and the relative position between them [Stab and Gurevych, 2014b]. Beyond features from superficial level, features were also extracted from semantic level of the relation including textual entailment and semantic similarity [Cabrio and Villata, 2012, Boltužić and Šnajder, 2014].

Unlike argument component identification where textual units are sentences or clauses, textual units in argumentative relation classification vary from clauses [Stab and Gurevych, 2014b] to multiple sentences [Biran and Rambow, 2011, Cabrio and Villata, 2012, Boltužić and Šnajder, 2014]. However, only few research has investigated the use of discourse relation within the text fragment to support the argumentative relation prediction. [Biran and Rambow, 2011] proposed that justifications of claim usually contain discourse structure which characterize the argumentation provided in the justification in support of the claim. However, their study made use of only discourse indicators but not the semantic relations. On the other hand, [Cabrio et al., 2013] studied the similarities and differences between Penn Discourse Treebank [Prasad et al., 2008] discourse relations and argumentation schemes [Walton et al., 2008]), and showed some PDTB discourse relations can be appropriate interpretations of particular argumentation schemes. Inspired by these pioneering studies, our thesis proposes to consider each argumentative unit in its relation with other surrounding text to enable advanced features extracted from the discourse context of the unit.

2.3.3 Argumentation Structure Identification

In contrast to the argumentative relation task, argumentation structure task emphasizes the attachment identification that is to determine if two argument components directly attach to each other, based on their rhetorical functions for the persuasion purpose of the text. Attachment is considered a generic argumentative relationship that abstracts both support and attack and is restricted to tree-structures in that a node attaches (has out-going edge) to only one other node, while can be attached (has in-coming edge) from one or more other nodes. [Palau and Moens, 2009] viewed legal argumentation as rooted at final decision that

is attached by conclusions which are further attached by premises. They manually examined a set of legal text and defined a context-free argumentative grammar to show a possibility of argumentative parsing for case law argumentation. [Peldszus and Stede, 2015] similarly assumed the tree-like representation of argumentation that have central claim be the root node to which pointed by claims (i.e., support or attack). Their data-driven approach took a fully-connected graph of all argument components as input and determined the edge weights based on features extracted from each component such as lemma, part-of-speech, dependency, as well the relative distance between the components. The minimum spanning tree of such weighted graph is returned as the output argumentation structure of the text.

Assuming that premises, conclusions and their attachment were already identified, [Feng and Hirst, 2011] aimed at determining the argumentation scheme [Walton et al., 2008] of the argument with the ultimate goal of recovering the implicit premises (enthymemes) of arguments. Besides the general features (relative position between conclusion and premises, number of premises) the study included scheme-specific features which are different for each target scheme (in one-vs-others classification) and based on pre-defined keywords and phrases.

A challenge to our context-aware argument mining model is determining the right context segment given the argument component. An ideal context segment is the minimal context segment that expresses a complete justification in a support of the argument component. Thus identifying the ideal context segment of an argument component requires to identify the argumentation structure. To make the context-aware argument mining idea more practical and easier to implement, our research does not require sentences in context segment must be semantically or topically related while some kind of relatedness among those sentences might be useful for the final argument mining tasks. In the course of this thesis, context segments are determined using simple heuristics such as window-size and topic segmentation output. In future, an use of argument structure identification for determining segment context is worth an investigation.

3.0 EXTRACTING ARGUMENT AND DOMAIN WORDS FOR IDENTIFYING ARGUMENT COMPONENTS IN TEXTS – COMPLETED WORK

3.1 INTRODUCTION

Argument component identification studies often use lexical (e.g., n-grams) and syntactic (e.g., grammatical production rules) features with all possible values [Burstein et al., 2003, Stab and Gurevych, 2014b]. However, such large and sparse feature spaces can cause difficulty for feature selection. In our study [Nguyen and Litman, 2015], we propose an innovative algorithm that post-processes the output of LDA topic model [Blei et al., 2003] to extract *argument words* (argument indicators, e.g. '*hypothesis*', '*reason*', '*think*') and *domain words* (specific terms commonly used within the topic's domain, e.g. '*bystander*', '*education*') which are used as novel features and constraints to improve the feature space. Particularly, we keep only argument words from unigram features, and remove higher order n-gram features (e.g., bigrams, trigrams). Instead of productions rules, we derive features from dependency parses which enable us to both retain syntactic structures and incorporate abstracted lexical constraints. Our lexicon extraction algorithm is semi-supervised in that we use manually-selected argument seed words to guide the process.

Different data-driven approaches for sublanguage identification in argumentative texts have been proposed to separate organizational content (shell) from topical content, e.g., supervised sequence modeling [Madnani et al., 2012], probabilistic topic models [Séaghdha and Teufel, 2014, Du et al., 2014]. Post-processing LDA [Blei et al., 2003] output was studied to identify topics of visual words [Louis and Nenkova, 2013] and representative words of topics [Brody and Elhadad, 2010, Funatsu et al., 2014]. Our algorithm has a similarity

with [Louis and Nenkova, 2013] in that we use seed words to guide the separation.

3.2 PERSUASIVE ESSAY CORPUS

The dataset for this study is an annotated corpus of persuasive essays [Stab and Gurevych, 2014a]. The essays are student writings in response to sample test questions of standardized English tests for foreign learners, and were posted online[1] for others' feedback. In the essays, the writers state their opinions (labeled as *MajorClaim*), towards the writing topics and validate those opinions with convincing arguments consisting of controversial statements (i.e., *Claim*) that support or attack the major claims, and evidences (i.e., *Premise*) that underpin the validity of the claims. Three experts identified possible argument components, i.e., *MajorClaim*, *Claim*, *Premise*, within each sentence, and connect the argument components using argumentative relations: *Support* and *Attack*. An example of persuasive essay in the corpus is given below.

Example essay 1: $^{(0)}$Effects of Globalization (Decrease in Global Tension)

$^{(1)}$During the history of the world, every change has its own positive and negative sides. $^{(2)}$Globalization as a gradual change affecting all over the world is not an exception. $^{(3)}$Although it has undeniable effects on the economics of the world; it has side effects which make it a controversial issue.

$^{(4)}$[*Some people prefer to recognize globalization as a threat to ethnic and religious values of people of their country*]$_{Claim}$. $^{(5)}$They think that [*the idea of globalization put their inherited culture in danger of uncontrolled change and make them vulnerable against the attack of imperialistic governments*]$_{Premise}$.

$^{(6)}$Those who disagree, believe that [*globalization contribute effectively to the global improvement of the world in many aspects*]$_{Claim}$. $^{(7)}$[*Developing globalization, people can have more access to many natural resources of the world*]$_{Premise}$ and [*it leads to increasing the pace of scientific and economic promotions of the entire world*]$_{Premise}$. $^{(8)}$In addition, they admit that [*globalization can be considered a chance for people of each country to promote their lifestyle through the stuffs and services imported from other countries*]$_{Premise}$.

$^{(9)}$Moreover, [*the proponents of globalization idea point out globalization results in considerable decrease in global tension*]$_{Claim}$ due to [*convergence of benefits of people of the world which is a natural consequence of globalization*]$_{Premise}$.

[1] *www.essayforum.com*

$^{(10)}$In conclusion, [*I would rather classify myself in the proponents of globalization as a speeding factor of global progress*]$_{MajorClaim}$. $^{(11)}$I think [*it is more likely to solve the problems of the world rather than intensifying them*]$_{Premise}$.

According to the coding scheme in [Stab and Gurevych, 2014a], each essay has one and only one MajorClaim. An essay sentence (e.g., sentence 9) can simultaneously have multiple argument components which are clauses of the sentence (Argumentative spans), and text spans that do not belong to any argument components (None spans). An argument component can be either a clause or a whole sentence (e.g., sentence 4). Sentences that do not contain any argument component are labeled *Non-argumentative* (e.g., sentences {1, 2, 3}). The three experts achieved inter-rater accuracy 0.88 for argument component labels and Krippendorff's α_U 0.72 for argument component boundaries.

Forming prediction inputs from Persuasive Essay Corpus is complicate due to the multiple-component sentences. For an illustration, let consider sentence 9 in the example. We have following text spans with their respective labels[2]:

Text span	Label
Moreover,	None
the proponents of globalization idea point out globalization results in considerable decrease in global tension	Claim
due to	None
convergence of benefits of people of the world which is a natural consequence of globalization	Premise
.	None

In this study, we use the model developed in [Stab and Gurevych, 2014b] as a baseline to evaluate our proposed approach. Following [Stab and Gurevych, 2014b], the None spans are not considered as prediction inputs. Therefore, a proper input of the prediction model is either a Non-argumentative sentence or an Argumentative span. Overall, the Persuasive Essay Corpus has 327 Non-argumentative sentences and 1346 Argumentative sentences. A distribution of argumentative labels is shown in the Table 1.

[2]A single punctuation is a proper span.

Argumentative label	#instances
Major-claim	90
Claim	429
Premise	1033
Non-argumentative	327
Total	1879

Table 1: Number of instances of each argumentative label in Persuasive Essay Corpus.

3.3 ARGUMENT AND DOMAIN WORD EXTRACTION

In this section we briefly describe the algorithm to extract argument and domain words from a development dataset using predefined argument keywords [Nguyen and Litman, 2015]. We recall that argument words are those playing a role of argument indicators and commonly used in different argument topics, e.g. '*reason*', '*opinion*', '*think*'. In contrast, domain words are specific terminologies commonly used within the topic, e.g. '*art*', '*education*'. Our notions of argument and domain languages share a similarity with the idea of shell language and content in [Madnani et al., 2012] in that we aim to model the lexical signals of argumentative content. However while [Madnani et al., 2012] emphasized the boundaries between argument shell and content, we emphasize more the lexical signals themselves and allow argument words to occur in the argument content. For example, the MajorClaim in Figure 1 has two argument words '*should*' and '*instead*' which make the statement controversial.

The development data for the Persuasive Essay Corpus are 6794 unlabeled essays (*Persuasive Set*) with titles collected from *www.essayforum.com*. We manually select 10 argument keywords/seeds that are the 10 most frequent words in the titles that seemed argument related: *agree, disagree, reason, support, advantage, disadvantage, think, conclusion, result, opinion*. We extract seeds of domain words as those in the titles but not argument keywords or stop words, and obtain 3077 domain seeds (with 136482 occurrences). Each domain seed

Topic 1 *reason exampl support agre think becaus disagre statement opinion believe therefor idea conclus ...*
Topic 2 *citi live big hous place area small apart town build communiti factori urban ...*
Topic 3 *children parent school educ teach kid adult grow childhood behavior taught ...*

Table 2: Samples of top argument words (topic 1), and top domain words (topics 2 and 3) extracted from the Persuasive Set. Words are stemmed.

is associated with an in-title occurrence frequency f.

All words in the development set including seed words are stemmed, and named entities are replaced with the corresponding NER labels by the Stanford parser. We run GibbsLDA++ implementation [Phan and Nguyen, 2007] of LDA [Blei et al., 2003] on the development set, and assign each identified LDA topic three weights: domain weight (DW) is the sum of domain seed frequencies; argument weight (AW) is the number of argument keywords[3]; and combined weight $CW = AW - DW$. For example, topic 2 in the LDA's output of Persuasive Set in Table 2 has $AW = 5$,[4] $DW = 0.15$, $CW = 4.85$, $f(citi) = 381/136482 = 0.0028$ given its 381 occurrences in the 136482 domain seed occurrences in the titles. LDA topics are ranked by CW with the top topic has highest CW value. We vary number of LDA topics k and select the k with the highest CW ratio of the top-2 topics ($k = 36$). The argument word list is the LDA topic with the largest combined weight given the best k. Domain words are the top words of other LDA topics but not argument or stop words.

Given 10 argument keywords, our algorithm returns a list of 263 argument words[5] which is a mixture of keyword variants (e.g. *think, believe, viewpoint, opinion, argument, claim*),

[3]Argument keywords are weighted more than domain seeds to reduce the size disparity of the two seed sets.

[4]Five argument keywords not shown in the table are: {*more, conclusion, advantage, who, which*}

[5]The complete list is shown in the APPENDIX A.

connectives (e.g. *therefore, however, despite*), and other stop words. 1582 domain words are extracted by the algorithm. We note that domain seeds are not necessarily present in the extracted domain words partially because words with occurrence less than 3 are removed from LDA topics.[6] On the other hand, the domain word list of Persuasive Set has 6% not in the domain seed set. Table 2 shows examples of top argument and domain words (stemmed) returned by the algorithm.

3.4 PREDICTION MODELS

3.4.1 Stab & Gurevych 2014

The model in [Stab and Gurevych, 2014b] (Stab14) uses following features extracted from the Persuasive Essay Corpus:

- *Structural features:* #tokens and #punctuations in argument component (AC)[7], in covering sentence, and preceding/following the AC in sentence; token ratio between covering sentence and AC. Two binary features indicate if the token ratio is 1 and if the sentence ends with a question mark. Five position features are covering sentence's position in essay, whether the AC is in the first/last paragraph, the first/last sentence of a paragraph.
- *Lexical features:* all n-grams of length 1-3 extracted from the text span that include the AC and its preceding text which is not covered by other AC's in sentence; verbs like '*believe*'; adverbs like '*also*'; and whether the AC has a modal verb.
- *Syntactic features:* #sub-clauses and depth of syntactic parse tree of the covering sentence of the AC; tense of main verb and grammatical production rules (VP $\rightarrow$ VBG NP) from the sub-tree that represent the AC.
- *Discourse markers:* discourse connectives of 3 relations: Comparison, Contingency, and

[6] Our implementation of [Stab and Gurevych, 2014b] model obtained performance improvement when removing rare n-grams, i.e., tokens with less than 3 occurrences. Thus, we applied the rare threshold of 3 to our pre-processing of the data.

[7] Gold-standard boundaries are used to identify Argumentative spans of the component.

Expansion[8] are extracted by the *addDiscourse* program [Pitler et al., 2009]. A binary feature indicates if the corresponding discourse connective precedes the AC.

- *First person pronouns:* Five binary features indicate whether each of *I*, *me*, *my*, *mine*, and *myself* is present in the covering sentence. An additional binary feature indicates if one of five first person pronouns is present in the covering sentence.
- *Contextual features:* #tokens, #punctuations, #sub-clauses, and presence of modal verb in preceding and following sentences of the AC.

In this study, we re-implement Stab14 to use as a baseline model. To evaluate our proposed model (described below) we compare its performance with the performance reported in [Stab and Gurevych, 2014b] as well as the performance of our implementation of Stab14.

3.4.2 Nguyen & Litman 2015

Our proposed model [Nguyen and Litman, 2015][9] (Nguyen15) improves Stab14 by using extracted argument and domain words as novel features and constraints to replace its n-gram and production rule features. Compared to n-grams in *lexical aspect*, argument words are believed to provide a much more compact representation of the argument indicators. As for the *structural aspect*, instead of production rules, e.g. "$S \rightarrow NP\ VP$", we use dependency parses to extract pairs of subject and main verb of sentences, e.g. "*I.think*", "*view.be*". Dependency relations are minimal syntactic structures compared to production rules. To further make the features topic-independent, we keep only dependency pairs that do not include domain words. In summary, our proposed model takes all features from the baseline except n-grams and production rules, and adds the following features: *argument words* as unigrams; *filtered dependency pairs* which are argumentative subject–verb pairs are used as skipped bigrams; and *numbers* of argument and domain words (see Figure 5). Our proposed model is compact with 956 original features compared to 5132 of the baseline.[10]

[8] Authors of [Stab and Gurevych, 2014b] manually collected 55 Penn Discourse Treebank markers after removing those that do not indicate argumentative discourse, e.g. markers of Temporal relations. Because the list of 55 discourse markers was not publicly available, we used a program to extract discourse connectives.

[9] In the paper, we named our model AD which stands for Argument and Domain word-based model.

[10] Counted in our implementation of Stab14. Because our implementation removes n-grams with less than 3 occurrences, it has smaller feature space than the original model in [Stab and Gurevych, 2014b].

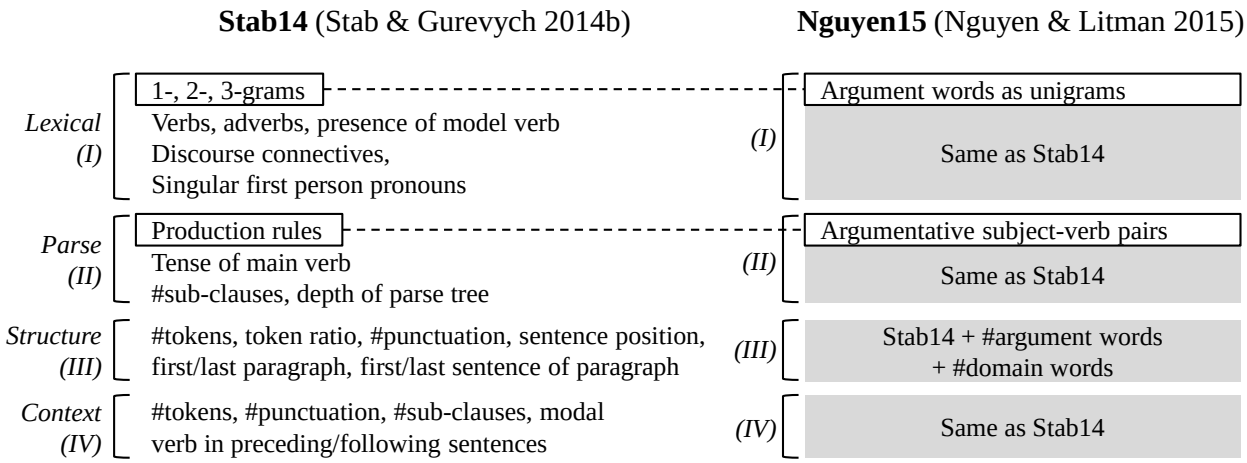

Figure 5: Feature illustration of Stab14 and Nguyen15. 1-, 2-, 3-grams and production rules in Stab14 are replaced by argument words and argumentative subject–verb pairs in Nguyen15.

3.5 EXPERIMENTAL RESULTS

3.5.1 Proposed vs. Baseline Models

This experiment replicates what was conducted in [Stab and Gurevych, 2014b]. We perform 10-fold cross validations and report the average results. In each run models are trained using LibLINEAR [Fan et al., 2008] algorithm with top 100 features returned by the InfoGain feature selection algorithm performed in the training folds. We use LightSIDE (lightside-labs.com) to extract n-grams and production rules, the Stanford parser [Klein and Manning, 2003] to parse the texts, and Weka [Hall et al., 2009] to conduct the machine learning experiments. Table 3 (left) shows the performances of three models: *BaseR* and *BaseI* are respectively the reported performance and our implementation of Stab14 [Stab and Gurevych, 2014b], and *Nguyen15* is our proposed model. Because of the skewed label distribution, all reported precision and recall are un-weighted average values from by-class performances.

	BaseR	BaseI	Nguyen15	BaseI	Nguyen15
#features	100	100	100	130	70
Accuracy	0.77	0.783	**0.794**+	0.803	**0.828***
Kappa	NA	0.626	**0.649***	0.640	**0.692***
Precision	**0.77**	0.760	0.756	0.763	**0.793**
Recall	0.68	0.687	**0.697**	0.680	**0.735**+

Table 3: Model performances with top 100 features (left) and best number of features (right). +, * indicate $p < 0.1, p < 0.05$ respectively in AD vs. BaseI comparison. Best values are in bold.

	AltAD	Nguyen15
Accuracy	0.770	**0.794***
Kappa	0.623	**0.649***
Precision	0.748	**0.756**
Recall	0.688	**0.697**
F1:MajorClaim	**0.558**	0.506
F1:Claim	0.468	**0.527***
F1:Premise	0.826	**0.844***
F1:None	1.000	1.000

Table 4: 10-fold performance with different argument words lists.

We note that there are performance disparities between BaseI (our implementation), and BaseR (reported performance in [Stab and Gurevych, 2014b]). The differences may mostly be due to dissimilar feature extraction methods and NLP/ML toolkits. Comparing BaseI and Nguyen15 shows that our proposed model Nguyen15 yields higher Kappa (significantly) and accuracy (trending).

To further analyze performance improvement by the Nguyen15 model, we use 75 randomly-selected essays to train and estimate the best numbers of features of BaseI and Nguyen15 (w.r.t F1 score) through a 9-fold cross validation, then test on 15 remaining essays. As shown in Table 3 (right), Nguyen15's test performance is consistently better with far smaller number of top features (70) than BaseI (130). Nguyen15 has 6 of 31 argument words not present in BaseI's 34 unigrams: *analyze, controversial, could, debate, discuss, ordinal*. Nguyen15 keeps only 5 dependency pairs: *I.agree, I.believe, I.conclude, I.think* and *people.believe* while BaseI keeps up to 31 bigrams and 13 trigrams in the top features. These indicate the dominance of our proposed features over generic n-grams and syntactic features.

3.5.2 Alternative Argument Word List

In this experiment, we study the prediction transfer of argument words when the development data to extract them is of a different genre than the test data. In a preliminary, we run the argument word extraction algorithm on a set of 254 academic writings (see §4.2 for a detailed description of this type of student essay) and extracted 429 argument keywords.[11]

To build an model based on the alternative argument word list (AltAD), we replace the argument words in Nguyen15 with those 429 argument words, re-filter the dependency pairs and update the number of argument words. We follow the same setting in the experiment above to train Nguyen15 and AltAD using top 100 features. As shown in Table 4, AltAD performs worse than Nguyen15, except a higher F1:MajorClaim but not significant. AltAD yields significantly lower accuracy, Kappa, F1:Claim and F1:Premise.

Comparing the two argument word lists gives us interesting insights. The two lists have 142 common words with 9 discourse connectives (e.g. '*therefore*', '*despite*'), 72 content words (e.g. '*result*', '*support*'), and 61 stop words. 30 of the common argument words appear in top 100 features of AltAD, but only 5 are content words: '*conclusion*', '*topic*', '*analyze*', '*show*', and '*reason*'. This shows that while the two argument word lists have a fair amount of common words, the transferable part is mostly limited to function words, e.g.

[11] The five argument keywords for this development set were *hypothesis, support, opposition, finding, study.* In that experiment, we did not consider each essay as an input document of LDA. Instead we broke essays into sections at citation sentences

discourse connectives, stop words. In contrast, 270 of the 285 unique words to AltAD are not selected for top 100 features, and most of those are popular terms in academic writings, e.g. '*research*', '*hypothesis*', '*variable*'. Moreover, Nguyen15's top 100 features have 20 argument words unique to the model, and 19 of those are content words, e.g. '*believe*', '*agree*', '*discuss*', '*view*'. These non-transferable parts suggest that argument words should be learned from appropriate seeds and development sets for best performance.

3.6 CONCLUSIONS

Our proposed features are shown to efficiently replace generic n-grams and production rules in argument mining tasks for significantly better performance. The core component of our feature extraction is a novel algorithm that post-processes LDA output to learn argument and domain words with a minimal seeding. These results proves our first sub-hypothesis (H1-1, §1.2) of effectiveness of context features in argument component identification. Moreover, our analysis gives insights into the lexical signals of argumentative content. While argument word lists extracted for different data can have parts in common, there are non-transferable parts which are genre-dependent and necessary for the best performance.

4.0 IMPROVING ARGUMENT MINING IN STUDENT ESSAYS USING ARGUMENT INDICATORS AND ESSAY TOPICS – COMPLETED WORK

4.1 INTRODUCTION

Argument mining systems for student essays need to be able to reliably identify argument components independently of particular writing topics. Prior argument mining studies have explored linguistic indicators of argument such as pre-defined indicative phrases for argumentation [Mochales and Moens, 2008], syntactic structures, discourse markers, first person pronouns [Burstein et al., 2003, Stab and Gurevych, 2014b], and words and linguistic constructs that express rhetorical function [Séaghdha and Teufel, 2014]. However only a few studies have attempted to abstract over the lexical items specific to argument topics for new features, e.g., common words with title [Teufel and Moens, 2002], cosine similarity with the topic [Levy et al., 2014], or to perform cross-topic evaluations [Burstein et al., 2003]. In a classroom, students can have writing assignments in a wide range of topics, thus features that work well when trained and tested on different topics (i.e., writing-topic independent features) are more desirable.

[Stab and Gurevych, 2014b] studied the argument component identification problem in persuasive essays, and used linguistic features like ngrams and production rules (e.g., *VP→VBG NP*, *NN→sign*) in their argument mining system. While their features were effective, their feature space was large and sparse. Our prior work [Nguyen and Litman, 2015] (see §3), addressed that issue by replacing n-grams with a set of argument words learned in a semi-supervised manner, and using dependency rather than constituent-based parsers, which were then filtered based on the learned argument versus domain word distinctions. While our new features were derived from a semi-automatically learned lexicon of argument and

domain words, the role of using such a lexicon was not quantitatively evaluated. Moreover, neither [Stab and Gurevych, 2014b] nor we used features that abstracted over topic lexicons, nor performed cross-topic evaluation.

In this chapter, we present our new study [Nguyen and Litman, 2016] that addresses the above limitations in four ways. First, in §4.2 we introduce a newly annotated corpus of academic essays from college classes and run all of our studies using both the new corpus and the prior persuasive essay corpus [Stab and Gurevych, 2014a] (see §3.2). Second, we present new features to model not only indicators of argument language but also to abstract over essay topics. Third, we build ablated models that do not use the extracted argument and domain words to derive new features and feature filters, so we can quantitatively evaluate the utility of extracting such word lists. Finally, in addition to 10-fold cross validation, we conduct cross-topic validation to evaluate model robustness when trained and tested on different writing topics.

Through experiments on two different corpora, we aim to provide support for the following three model-robustness hypotheses: *models enhanced with our new features will outperform baseline models* when evaluated using (h1) 10-fold cross validation and (h2) cross-topic validation; *our new models will demonstrate topic-robustness* in that (h3) their cross-topic and 10-fold cross validation performance levels will be comparable.

4.2 ACADEMIC ESSAY CORPUS

The Academic Essay Corpus consists of 115 student essays collected from a writing assignment of university introductory Psychology classes in 2014. The assignment requires each student to write an introduction of the observational study that she conducted. In the study, the student student proposes one or two hypotheses about the effects of different observational variables to a dependent variable, e.g., effect of gender to politeness. The student is asked to use relevant studies/theories to justify support for the hypotheses, and to present at least one theoretical opposition with a hypothesis. The students are required to write their introduction in form of an argumentative essay and follow the APA guideline that uses

Argumentative label	#sentences
Hypothesis	185
Finding	131
– *Support finding*	50
– *Opposition finding*	81
Non-argumentative	2998
Total	3314

Table 5: Number of sentences of each argumentative label in Academic Essay Corpus.

citations whenever they refer to prior studies. Compared to Persuasive Essay Corpus, while claims in the persuasive essays are mostly substantiated by personal experience, hypotheses in the academic essays are elaborated by findings from the literature. This makes the most distinguished difference between the two corpora.

We had two experts label each sentence of the essays whether it is a *Hypothesis* statement, *Support* finding, or *Opposition* finding (if so it is an *argumentative sentence*, no sentences have multiple labels). As the focus of this study is the identification of argument component without caring about the argumentative relation between components, *Support* and *Opposition* sentences are grouped into *Finding* category. The two annotators achieved inter-rater kappa 0.79 for the agreement on sentence labels for the coding scheme *Hypothesis-Finding*. For an example, two last paragraphs of an academic essay is given bellow. The essay's topic is "Amount of Bystanders Effect on Helping Behavior".

Example essay 2: (1)Several studies have been done in the past that also examine the ideas of the bystander effect and diffusion of responsibility, and their roles in social situations. (2)[Daniel M. Wegner conducted a study in 1978 that demonstrated the bystander effect on a college campus by comparing the ratio of bystanders to victim, which showed that *the more bystanders in comparison to the victims led to less people helping* (Wegner, 1983).]$_{Support}$ (3)[Another supporting study was conducted Rutkowski in 1983 that also demonstrated that *with larger groups comes less help for victims in non-emergency situations due to less social pressure* (Rutkowski, 1983).]$_{Support}$ (4)Although these studies demonstrate the bystander effect and diffusion of responsibility, other studies oppose these ideas. (5)[One

strong study that opposes the bystander effect was done in 1980 by Junji Harada that showed that *increase in group size, even in a face to face proximity, did not decrease the likelihood of being helped* (Harada, 1980).]$_{Opposition}$

(6)In order to find out specifically the effects that the bystander effect has in diverse settings, this study focuses on a non-emergency situation on a college campus. (7)[The hypothesis, based on the bystander effect demonstrated in Wegner's study (1978), is that *with more people around, less people will take the time to help the girl pick up her papers.*]$_{Hypothesis}$

In the example, the main content of argumentative sentences that express the argumentative role of the sentences (e.g., *hypothesis*, *support*, or *opposition*) are italicized. Given the annotation, *Finding* sentences are {2, 3, 5}. Table 5 shows the label distribution in the corpus. As we can see, the dataset is very skewed with Non-argumentative sentences are more than 90% of the data. Also while each essay has at least one Hypothesis statement, not all essays have Support and Opposition sentences.

4.3 PREDICTION MODELS

4.3.1 Stab14

As described in §3.4.1, Stab14 model was developed using Persuasive Essay Corpus. Despite the differences between persuasive essays and academic essays, the Stab14 model is also applicable to the Academic Essay Corpus. First, the two corpora share certain similarities in writing styles and coding schemes. Both corpora consist of student writings whose content is developed to elaborate a main hypothesis for a persuasion purpose. Regarding coding schemes, MajorClaims in persuasive essays correspond to Hypothesis statements in academic essays, and Claims match Support and Opposition findings. Premises in persuasive essays can be considered student writer's elaborations of previous studies in academic essay. Second, most of prediction features proposed in their study are generic and genre-independent, e.g., n-grams, grammatical production rules, and discourse connectives, which are expected to work for student writings in general. Therefore, we adapt [Stab and Gurevych, 2014b], Stab14, model to the Academic Essay Corpus for a baseline model to evaluate our approach. The version of Stab14 that works for Persuasive Essay is described in §3.4.1.

As the Academic Essay Corpus has annotation done at sentence-level and contains no information of argument component boundaries, all features of Stab14 that involve boundaries information are not applicable to Academic Essay Corpus. Therefore, Stab14 model is adapted to Academic Essay Corpus by simply extracting all features from the sentences, and removing features that require both argument component and covering sentence, e.g., token ratio.

4.3.2 Nguyen15v2

We implement two modified versions of the Nguyen15 model (§3.4.2) as the second baseline (Nguyen15v2),[1] one for each corpus. Additional experiments with Persuasive Essay Corpus showed that argument and domain word count features were not effective, so we decided to remove these two features from Nguyen15. For each version we re-implement the argument and domain word extraction algorithm (§3.3) to extract argument and domain words from a development dataset.

For the Academic Essay Corpus, we use 254 unannotated essays (*Academic Set*) with titles from Psychology classes in years 2011 and 2013 as the development data. We select 5 argument keywords which were specified in the writing assignments: *hypothesis, support, opposition, finding, study*. Filtering out argument keywords and stop words in essay titles of the academic set, we obtain 264 domain seeds (with 1588 occurrences). The argument and domain word extraction algorithm returns 11 LDA topics, 315 (stemmed) argument words,[2] and 1582 (stemmed) domain words. The learned argument words are a mixture of keyword variants (e.g. *research, result, predict*), methodology terms (e.g. *effect, observe, variable, experiment, interact*), connectives (e.g. *also, however, therefor*), and other stop words. Learned domain words have 86% not in the domain seed set. Table 6 shows examples of top argument and domain words (stemmed) returned by the algorithm.

[1] In the paper, we named this model Nguyen15 [Nguyen and Litman, 2016]. We do not use the original in this thesis because it might make people confused with Nguyen15 model described in §3.4.2.

[2] The complete list is shown in the APPENDIX A.

Topic 1 *studi research observ result hypothesi time find howev predict support expect oppos ...*
Topic 2 *respons stranger group greet confeder individu verbal social size peopl sneez ...*
Topic 3 *more gender women polit femal male men behavior differ prosoci express gratitud ...*

Table 6: Samples of top argument words (topic 1), and top domain words (topics 2 and 3) extracted from Academic Set. Words are stemmed.

4.3.3 wLDA+4

Our proposed model of this study, wLDA+4, is Nguyen15v2 (with the LDA supported features) expanded with 4 new feature sets extracted from the covering sentences of the associated argument components. A summary of features used in this model is given in Figure 6. To model the topic cohesion of essays, we include two common word counts:

1. *Numbers of common words* of the given sentence with the preceding one and with the essay title.

We also proposed new lexical features for better indicators of argument language. We observe that in argumentative essays students usually use comparison language to compare and contrast ideas. However not all comparison words are independent of the essay topics. For example, while adverbs (e.g., '*more*') are commonly used across essays, adjectives (e.g., '*cheaper*', '*richer*') seem specific to the particular topics. Thus, we introduce the following comparison features:

2. *Comparison words*: comparative and superlative adverbs. *Comparison POS*: two binary features indicating the presences of *RBR* and *RBS* part-of-speech tags.

We also see that student authors may use plural first person pronouns (*we*, *us*, *our*, *ours*, and *ourselves*) as a rhetorical device to make their statement sound more objec-

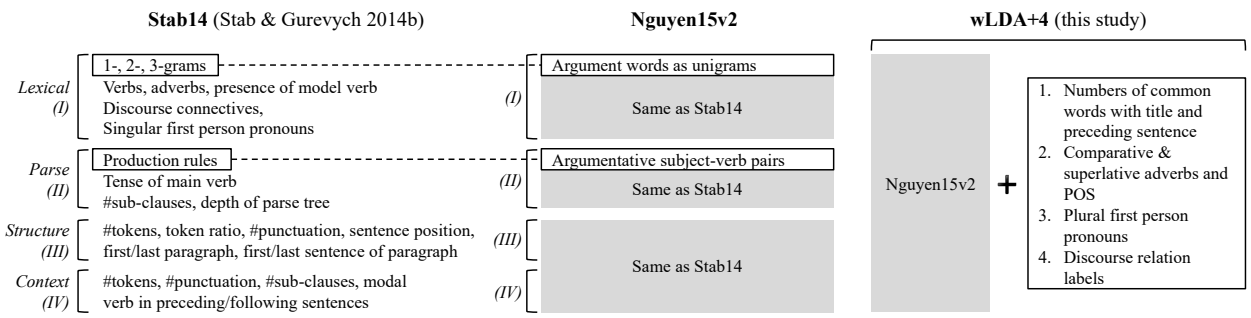

Figure 6: Feature illustration of Stab14, Nguyen15v2 and wLDA+4. 1-, 2-, 3-grams and production rules in Stab14 are replaced by argument words and argumentative subject–verb pairs in Nguyen15v2. wLDA+4 extends Nguyen15v2 with 4 new feature sets.

tive/persuasive, for instance "*we always find that we need the cooperation.*" We supplement the first person pronoun set in the baseline models with 5 plural first person pronouns:

3. Five binary features indicating whether each of 5 *plural first person pronouns* is present.

We notice that many discourse connectives used in baseline models are duplicates of our extracted argument words, e.g., '*however*'. Thus using both argument words and discourse connectives may inefficiently enlarge the feature space. To emphasize the discourse information, we include discourse relations as identified by addDiscourse program [Pitler et al., 2009] as new features:

4. Three binary features showing if each of *Comparison*, *Contingency*, *Expansion* discourse relations is present.[3]

4.3.4 wLDA+4 ablated models

We propose two simple alternatives to wLDA+4 to examine the role of argument and domain word lists in our argument mining task:

[3]The temporal discourse relation was not used in [Stab and Gurevych, 2014b] and thus is ignored in this study.

• **woLDA**: we disable the LDA-enabled features and constraints in wLDA+4 so that woLDA does not include argument words, but uses all possible subject–verb pairs. All other features of wLDA+4 are unaffectedly applied to woLDA. Comparing woLDA to wLDA+4 will show the contribution of the extracted argument and domain words to the model performance.

• **Seed**: extracted argument and domain word lists are replaced with only the seeds that were used to start the semi-supervised argument and domain word learning process (see next section). Comparing Seed to wLDA+4 will show whether it is necessary to use the semi-supervised approach for expanding the seeds to construct the larger/more comprehensive argument and domain word lexicons.

4.4 EXPERIMENTAL RESULT

4.4.1 10-fold Cross Validation

We first conduct 10-fold cross validations to evaluate our proposed model and the baseline models. All models are trained using the SMO (as in [Stab and Gurevych, 2014b]) implementation of SVM in Weka [Hall et al., 2009]. LightSIDE (*lightsidelabs.com*) and Stanford parser [Klein and Manning, 2003] are used to extract n-grams, parse trees and named entities. We follow [Stab and Gurevych, 2014b] and use top 100 features ranked by InfoGain algorithm on training folds to train the models. To obtain enough samples for a significance test when comparing model performance in 10-fold cross validation to cross-topic validation, we perform 10 runs of 10-fold cross validations (10×10 cross-validation) and report the average results over 10 runs.[4] We use T-tests to compare performance of models given that each model evaluation returns 10 samples of 10-fold cross validation performance. As the two corpora are very class-skewed, we report unweighted precision and recall. Also while accuracy is a common metric, kappa is a more meaningful value given our imbalanced data.

[4]From our prior study [Nguyen and Litman, 2015], and additional experiments, we also noticed that the skewed distributions of our corpora make stratified 10-fold cross validation performance notably affected by the random seeds. Thus, we decided to conduct multiple cross validations in this experiment to reduce any effect of random folding.

	Persuasive Essay Corpus				
Metric	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.787*	0.792*	0.780*	0.781*	**0.805**
Kappa	0.639*	0.649*	0.629*	0.632*	**0.673**
Precision	0.741*	0.745*	0.746*	0.740*	**0.763**
Recall	0.694*	0.698*	0.695*	0.695*	**0.720**
	Academic Essay Corpus				
Metric	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.934*	**0.942**+	0.933*	0.935*	0.941
Kappa	0.558*	**0.635**	0.528*	0.564*	0.629
Precision	0.804*	**0.830**+	0.829	0.826	0.825
Recall	0.628*	**0.695**	0.594*	0.637*	**0.695**

Table 7: 10×10-fold cross validation results. Best values in bold. +: $p < 0.1$, *: $p < 0.05$ by T-test when comparing with wLDA+4.

Model performances are reported in Table 7.

Our first analysis is about the performance improvement of our proposed model over the two baselines. We see that our model wLDA+4 significantly outperforms Stab14 in all reported metrics across both two corpora. However comparing wLDA+4 and Nguyen15v2 reveals inconsistent patterns. While wLDA+4 yields a significantly higher performances than Nguyen15v2 when evaluated in the persuasive corpus, our proposed model performs worse than that baseline in the academic corpus. Looking at individual metrics of these two models we see that Nguyen15v2 has trending higher accuracy ($p = 0.05$) and also trending higher precision ($p = 0.09$) than wLDA+4 in academic corpus. The differences on kappa and recall between the two models are not significant. These results partially support our first model-robustness hypothesis (h1) in that our proposed features improve over both baselines using 10-fold cross validation in the persuasive corpus only.

We now turn to our feature ablation results. Removing the LDA-enabled features from wLDA+4, we see that woLDA's performance figures are all significantly worse than wLDA+4 except for precision in the academic corpus. Furthermore, we find that argument keywords and domain seeds are poor substitutes for the full argument and domain word lists learned from these seeds. This is shown by the significantly lower performances of Seed compared to wLDA+4, except for precision in the academic corpus. Nonetheless, adding the features computed from just argument keywords and domain seeds still helps Seed perform better than woLDA (with higher accuracy, kappa and recall in both persuasive and academic corpora).

4.4.2 Cross-topic Validation

To better evaluate the models when predicting essays of unseen topics we conduct cross-topic validations where training and testing essays are from different topics [Burstein et al., 2003]. We examined 90 persuasive essays and categorized them into 12 groups including 11 single-topic groups, each corresponds to a major topics (have 4 to 11 essays), e.g., *Technologies* (11 essays), *National Issues* (10), *School* (8), *Policies* (7), and a mixed group of 17 essays of minor topics (each has less than 3 essays), e.g., *Prepared Food* (2 essays).

We manually split 115 academic essays into 5 topics accordingly to the studied variables. *Attractiveness* as a function of clothing color (20 essays), *Email-response rate* as a function of recipient size (22), *Helping-behavior* with effects of gender and group size (31), *Politeness* as a function of gender (23), *Self-description* and word choices with influences of gender and self-esteem (19).

Again all models are trained using the top 100 features selected in training folds. In each folding, we use essays of one topic for evaluation and all other essays to train the model. T-test is used to compare each two sets of by-fold performances.

We first evaluate the performance improvement of our model compared to the baselines. As shown in Table 8, wLDA+4 again yields higher performance than Stab14 in all metrics of both corpora, and the improvements are significant except for precision in the academic essay. Moreover we generally observe a larger performance gap between wLDA+4 and Stab14 in cross-topic validation than in 10-fold cross validation. More importantly, with cross-

	Persuasive Essay Corpus				
Metric	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.780*	0.796	0.774*	0.776*	**0.807**
Kappa	0.623*	0.654+	0.618*	0.623*	**0.675**
Precision	0.722*	0.757*	0.751	0.734	**0.771**
Recall	0.670*	0.695*	0.681*	0.686*	**0.722**
	Academic Essay Corpus				
Metric	Stab14	Nguyen15v2	woLDA	Seed	wLDA+4
Accuracy	0.928*	0.939+	0.931*	0.935*	**0.944**
Kappa	0.491*	0.598+	0.474*	0.547*	**0.630**
Precision	0.768	0.832	**0.866**	0.839*	0.851
Recall	0.565*	0.664	0.551*	0.617*	**0.686**

Table 8: Cross topic validation results. Best values in bold. +: $p < 0.1$, *: $p < 0.05$ by T-test when comparing with wLDA+4.

topic validation, wLDA+4 now yields better performance than Nguyen15v2 for all metrics in both persuasive and academic corpora. Especially, our proposed model now even has trending higher accuracy and kappa than Nguyen15v2 in academic corpus. This shows a clear contribution of our new features in the overall performance, and supports our second model-robustness hypothesis (h2) that *our new features improve the cross-topic performance in both corpora compared to the baselines.*

With respect to feature ablation results, our findings are consistent with the prior cross-fold results in that woLDA and Seed both have lower performance (often significantly) than wLDA+4 (with one exception). Seed again generally outperforms woLDA, indicating that deriving features from even impoverished argument and domain word lists is better than not using such lexicons at all.

Next, we compare wLDA+4 performance across the cross-fold and cross-topic experimen-

tal settings (using a T-test to compare the mean of 10 samples of 10-fold cross validation performance versus the mean of cross-topic validation performance). In both corpora we see that wLDA+4 yields higher performance for all metrics in cross-topic versus 10-fold cross validation, except for recall in the academic corpus. Of these cross-topic performance figures, wLDA+4 has significantly higher precision and trending higher accuracy in the persuasive corpus. In academic corpus, wLDA+4's cross-topic accuracy, precision and recall are all significantly better than the corresponding figures for 10-fold cross validation. These results support strongly our third model-robustness hypothesis (h3) that *our proposed model's cross-topic performance is as high as 10-fold cross validation performance.*

In contrast, Nguyen15v2's performance difference between cross-topic and random-folding validations does not hold a consistent direction. Stab14 returns significantly higher results in 10-fold cross validation than cross-topic validation in both persuasive and academic corpora. Also woLDA and Seed's cross-topic performances are largely worse than those of 10-fold cross validation. Overall, the cross-topic validation shows the ability of our proposed model to perform reliably when the testing essays are from new topics, and the essential contribution of our new features to this high performance.

To conclude this section, we give a qualitative analysis of the top features selected in our proposed model. In each folding we record the top 100 features with associated ranks. By the end of cross-topic validation, we have a pool of top features ($\approx$200 for each corpus), with an average rank for each. First we see that the proportion of argument words is about 49% of pooled features in both corpora, and the proportion of argumentative subject–verb pairs varies from 8% (in persuasive corpus) to 15% (in academic corpus). The new features introduced in wLDA+4 that are present in the top features include: two common word counts; *RBR* part-of-speech; person pronouns *We* and *Our*; discourse labels *Comparison, Expansion, Contingency.* All of those are in the top 50 except that *Comparison* label has average rank 79 in the persuasive corpus. This shows the utility of our new feature sets. Especially the effectiveness of common word counts encourages us to study advanced topic cohesion features in future work.

	Stab's test set		Nguyen's test set		
Metric	Stab best	Our SMO	Nguyen best	Our SMO	Our Lib-LINEAR
Accuracy	0.77	**0.816**	0.828	0.819	**0.837**
Kappa	–	**0.682**	0.692	0.679	**0.708**
Precision	0.77	**0.794**	0.793	0.762	**0.811**
Recall	0.68	**0.726**	0.735	0.703	**0.755**

Table 9: Model performance on test sets. Best values in bold.

4.4.3 Performance on Held-out Test Sets

The experiments above used 10×10-fold cross-validation and cross-topic validation to investigate the robustness of prediction features. Note that this required us to re-implement both baselines as neither had previously been evaluated using cross-topic validation.[5] However, since both baselines were evaluated on single held-out test sets of Persuasive Essay Copora, that were available to us, our last experiment compares wLDA+4's performance with the best *reported* results for the original baseline implementations [Stab and Gurevych, 2014b, Nguyen and Litman, 2015] using their exact same training/test set splits. That is, we train wLDA+4 trained using SMO classifier with top 100 features with the two training sets of 72 essays [Stab and Gurevych, 2014b] and 75 essays [Nguyen and Litman, 2015], and report the corresponding held-out test performances in Table 9.

While test performance of our model is higher than [Stab and Gurevych, 2014b], our model has worse test results than [Nguyen and Litman, 2015]. This is reasonable as our model was trained following the same configuration as in [Stab and Gurevych, 2014b][6], but was not optimized as in [Nguyen and Litman, 2015]. In fact, [Nguyen and Litman, 2015] obtained their best performing model using LibLINEAR classifier with top 70 features. If

[5]While Nguyen15v2 (but not Stab14) had been evaluated using 10-fold cross-validation, the random fold data cannot be replicate.

[6]With respect to the cross validations, while our chosen setting is in favor of Stab14, it still offers an acceptable evaluation as it is not the best configuration for either Nguyen15v2 or wLDA+4.

we keep our top 100 features but replace SMO with LibLINEAR, then wLDA+4 gains performance improvement with accuracy 0.84 and Kappa 0.71. Thus, the conclusions from our new cross fold/topic experiments also hold when wLDA+4 is directly compared with published baseline test set results.

4.5 CONCLUSIONS

Motivated by practical argument mining for student essays (where essays may be written in response to different assignments), we have presented new features that model argument indicators and abstract over essay topics, and introduced a new corpus of academic essays to better evaluate the robustness of our models. Our proposed model in this study shows robustness in that it yields performance improvement with both *cross-topic* and *10-fold cross* validations for different types of student essays, i.e., *academic* and *persuasive*. Moreover, our model's cross-topic performance is even higher than cross-fold performances for almost all metrics.

Experimental results also show that while our model makes use of effective baseline features that are derived from extracted argument and domain words, the high performance of our model, especially in cross-topic validation, is also due to our new features which are generic and independent of essay topics. That is, to achieve the best performance, the new features are a necessary supplement to the learned and noisy argument and domain words. These results along with the results obtained in Chapter 3 strongly prove our first sub-hypothesis (H1-1, §1.2) of the effectiveness of contextual features in argument component identification.

5.0 EXTRACTING CONTEXTUAL INFORMATION FOR IMPROVING ARGUMENTATIVE RELATION CLASSIFICATION – PROPOSED WORK

5.1 INTRODUCTION

Research on classifying argumentative relation between pairs of arguments or argument components has proposed a variety of features ranging from superficial level, e.g., word pair, relative position, to semantic level, e.g., semantic similarity, textual entailment. [Cabrio and Villata, 2012, Boltužić and Šnajder, 2014] studied online debate corpora and aimed at identifying whether user comments support or attack the debate topic.[1] They proposed to use content-rich features including semantic similarity and textual entailment. In principle, they expect the comment text (which is usually longer) to entail the topic phrase (which is usually shorter). [Boltužić and Šnajder, 2014] calculated semantic similarity between each comment sentence and the topic phrase, and returned the max and mean of sentence-level similarity score. Despite the fact that user comments are usually long with multiple sentences, both [Cabrio and Villata, 2012] and [Boltužić and Šnajder, 2014] did not consider the discourse structure of the comment as an auxiliary information to support the prediction. It has been proposed in [Biran and Rambow, 2011] that justifications (e.g., user comment) usually contain discourse structure that characterize the argumentation provided in the justification. We believe that identifying the discourse structures of justification will give insights on argumentation patterns used by writers to show their stances towards the argument topic.

To illustrate our idea, let consider the following excerpt from a persuasive essay in Persuasive Essay Corpus:

[1] Both user comments and debate topics are consider argument in the studies.

(1)In addition, cooking is one of arts humans create. (2)The more cooked food we chosen, the more cooking skills we lose. (3)At the increasing living pace, the majority of people tend to choose microwave as their unique cooker that help them prepare a dish in five minutes. (4)But rare people have been aware that this has contributed to a modification of cooking habits, which may cause the loss of our custom and culture about cooking.

(5)In conclusion, although the invention of prepared foods definitely satisfies the demand of some people who are busy in their work, it is not a good thing.

The excerpt consist of a justification in sentences {1, 2, 3, 4} which support a claim in sentence 5. Analyzing the discourse structure of the justification, we can see that the writer wanted to prove that *losing cooking skills* is a bad thing, which causes *losing custom and culture*, which consequently shows a stance against the *prepared foods*.

Differently from [Cabrio and Villata, 2012,Boltužić and Šnajder, 2014], [Stab and Gurevych, 2014b] aimed at classifying the argumentative relations (i.e., support vs. non-support[2]) between argument components. An argument components in [Stab and Gurevych, 2014b] is a sentence or a sentence clause so it is less content-rich than user comments in [Cabrio and Villata, 2012, Boltužić and Šnajder, 2014]. [Stab and Gurevych, 2014b] proposed a diverse feature set including features involving information from both components of the pair. e.g., word pairs, common words, relative positions. However, a limitation of their model is the lack of contextual information as mentioned in their paper [Stab and Gurevych, 2014b]. For example, it is hard to determine the support relation between these two argument component: "*It helps relieve tension and stress*" and "*Exercising improves self-esteem and confidence*" without knowing that "*it*" refers to "*Exercising*". Another example is given in the following excerpt:

(1)However, there are some serious problems springing from modern technology. (2)First, deadly and powerful weapons can be a huge threat to the world's peace. (3)Second, a lot of people spend too much time using hi-tech devices nowadays. (4)They abuse them so severely that they feel they can hardly live without them. (5)This can have a detrimental effect on their health, since they are likely to develop many dangerous diseases, including obesity, heart attack and high blood-pressure.

To support the claim in sentence 1, the writer provides two justifications. The first justification in sentence 2 mentions "*weapons*" and "*threat*" which give a clear signal of

[2]Non-support relations include attacks and no-relations.

support for "*serious problems*" mentioned in the claim. However, the second justification is a series of premises in sentences {3, 4, 5} which together prove a main point of *health issue*. Without considering the context given in sentences {3, 4}, one cannot easily see that the *health issues* listed in sentence 5 are caused by modern technology, and thus cannot decide if the premise in sentence 5 is a support of the claim.

Given these issues of existing work on argumentative relation classification, we proposed a general framework that exploiting contextual information to tackle the problems. First, instead of considering argument components isolatedly as in [Stab and Gurevych, 2014b], our approach put each argument component in its context segment (see Definition 1, §1.1) to enrich the justification and enable contextual features. Second, we extract discourse relations, textual entailment, and semantic similarity from the context segments to use as contextual features. We consider both two discourse structure framework which are Penn Discourse Treebank [Prasad et al., 2008], and Rhetorical Structure Theory [Carlson et al., 2001] and use available toolkits for discourse relation extraction. To evaluate the contribution of contextual features, we augment the prediction models in [Stab et al., 2014] with these contextual features, and evaluate the enhanced model using two corpora.

5.2 DATA

The first corpus used in this study is the Persuasive Essay Corpus [Stab and Gurevych, 2014a] (see §3.2). Accoring to the coding scheme in [Stab and Gurevych, 2014a], after identifying possible argument components (i.e., MajorClaim, Claim, Premise) in an essay, annotators were asked to identify the relation (i.e., Support, Attack) between pairs of argument components. Constraints are applied to relation identification. First, argumentative relations are directed and can hold between a Premise and another Premise, a Premise and a (Major-) Claim, or a Claim and a MajorClaim. Except for the relation from Claim to MajorClaim, an argumentative relation does not cross paragraph boundaries. Three annotators achieved Krippendorff's $\alpha = 0.81$ for argumentative relations. Of 429 Claims, 365 support the associated MajorClaim, and 64 attack. In all annotated argumentative relations, 1312 are Support

	#instances
Major-claim	90
Claim	429
Claim (support)	365
Claim (attack)	64
Support relation	1312
Attack relation	161

Table 10: Statistics of the Persuasive Essay Corpus.

and 161 are Attack. Statistics of argumentative relations in Persuasive Essay Corpus is given in Table 10.

The second corpus we use for this study is an expanded version of Academic Essay Corpus (see §4.2). A problem with the current version of Academic Essay Corpus is the small numbers of Support and Opposition sentences. In 115 essays, we obtained only 50 Support and 81 Opposition sentences which contribute 1.5% and 2.4% of the whole corpus, respectively. This highly skewed distribution of argumentative labels would cause great difficulty to prediction models. Therefore, we are finalizing another annotation with the same coding scheme, but on a second set of academic essays. This second set consists of 91 academic essays collected from the the same Psychology course as in Academic Essay Corpus but in years 2011 and 2013. The writing assignment required student to provide at least 5 supports and one opposition so we expect to obtain more positive argumentative labels than in the original Academic Essay Corpus. The second set of academic essays includes 2290 sentences and were previously coded by 2 experts. Detailed inter-rater agreement is shown in Table 11. We are now consolidating the argumentative labels for sentences where the two annotators disagreed to create the final corpus.[3] For the sentences where the two annotators agreed, we set what was annotated as final labels. A statistic of label distribution computed

[3]Once the disagreements are resolved, we will also replicate the argument component identification experiments on this dataset.

Label	kappa
Hypothesis	0.86
Support	0.58
Opposition	0.61
Non-argumentative	0.65
4-way	0.67

Table 11: Inter-rater kappa's for the second set of academic essay.

Argumentative label	#sentences
Hypothesis	113
Finding	178
– *Support finding*	116
– *Opposition finding*	62
Total sentences	2290

Table 12: Number of sentences of each argumentative label, where the two annotated agreed in the second set of academic assays.

using agreed sentences are shown in Table 12.

5.3 TWO PROBLEM FORMULATIONS AND BASELINE MODELS

5.3.1 Relation with Argument Topic

For the argumentative relation classification task, we form two prediction problems. The first problem formulation adapts the problem statement in [Biran and Rambow, 2011, Cabrio and

Villata, 2012,Boltužić and Šnajder, 2014]. That is given an argument topic and an argumentative content, identify whether the argumentative content is for or against the argument topic. In persuasive essay, argument topic is the MajorClaim. We can only use Claim components as argumentative content because not all Premise components were annotated for a relation with the MajorClaim [Stab and Gurevych, 2014a]. In academic essay, argument topics are Hypothesis sentences, and argumentative content are Support/Opposition sentences.

To evaluate our proposed approach in this problem formulation, we implement two baseline models. The first baseline model follows the approach in [Boltužić and Šnajder, 2014] in that makes use of only semantic similarity and textual entailment features. Features are computed on the argument topic sentence and context segment of the argumentative content. The second baseline model re-implements the approach in [Stab and Gurevych, 2014b].

5.3.2 Pair of Argument Components

The second problem formulation follows the problem statement in [Stab and Gurevych, 2014b] that is to identify argumentative relation between possible pairs of argument components in the same paragraph. For this problem setting, only Persuasive Essay Corpus is usable because the academic essays do not have Support/Opposition sentences annotated for argumentative relation between them. Both models proposed in [Boltužić and Šnajder, 2014] and [Stab and Gurevych, 2014b] are used as the baseline models.

5.3.3 Baseline Models

The first baseline model re-implements the model proposed in [Boltužić and Šnajder, 2014]. [Boltužić and Šnajder, 2014] studied the argumentative relations between user comments and the arguments in online debates, and built a prediction model using textual entailment and semantic text similarity features. Following their work, we apply 7 pre-trained textual entailment algorithms for each pair of texts and used two output from each algorithm, i.e., a binary decision (Entailment vs. Not) and the degree of confidence, to form 14 features. Regarding semantic similarity features, we compute similarity score for each possible pair of

sentences between user comment and the argument, and use the set of individual score as well as the mean score as features.

Our second baseline models adapts the work by [Stab and Gurevych, 2014b] for classifying argumentative relations between argument components. Given a pair of argument components (one is considered the source and the other is the target) [Stab and Gurevych, 2014b] extracted 4 feature sets for their classification model.

- Structural features: number of tokens and punctuations in source and target, and the absolute difference of each two counts; sentence position of source and target, and sentence distance between them; whether the source and the target are in the first or last sentence of the paragraph, or in the same sentence.
- Lexical features: pairs of words, and pair of first words from source and target; number of common words and the presence of modal verb in the source and the target.
- Syntactic features: syntactic production rules extracted from the source and the target.
- Indicators: discourse connectives that are present in the source and the target.
- Predicted type: the predicted argument component label of the source and the target.

5.3.4 Evaluations

Following our work on argument component identification, we conduct both 10-fold cross validation and cross-topic validation. We use the topic information collected in §4 to separate essays into groups. Besides, because [Stab and Gurevych, 2014b] used a fixed data split to train and test their model, we will use their data split to validate our implementation of their model as well as compare performance of our proposed model with their reported results.

5.4 SOFTWARE SUPPORT

We use the following software for different processing tasks in extracting contextual features:

- PDTB discourse parser by [Wang and Lan, 2015].
- RST discourse parser by [Xue et al., 2015].

- Excitement Open Platform [Pado et al., 2013] for textual entailment between two texts.
- SEMILAR [Rus et al., 2013], TakeLab [Šarić et al., 2012], and Sent2Vec [Huang et al., 2013] for semantic similarity between short text, i.e., sentences or phrases.

5.5 PILOT STUDY

We conduct a preliminary experiment with the Persuasive Essay Corpus to evaluate the effectiveness of the discourse relation features in the argumentative relation classification task. Following [Stab and Gurevych, 2014b], we extract all possible ordered pairs of argument components in the same paragraph.[4] In 6330 pairs obtained, 989 (15.6%) have support relation. The rest either have attack relations or no relations, and are grouped into Non-support class. In this study, we aim at predicting if a given ordered pair has support or non-support relation.

We implement a simple baseline model (*Lexical*) that uses only lexical features in the model by [Stab and Gurevych, 2014b]. In fact, lexical features are reported the most effective for the argumentative relation classification [Stab and Gurevych, 2014b]. For our proposed model, we first extract following *Discourse* features from the context segments of the source and target components:

- Discourse connectives: we extract connectives in context sentences preceding and following the source and the target.
- Discourse relations: we extract PDTB and RST discourse relations within the source and target sentences, between source context sentences, between target context sentences, and between a sentence in source segment and a sentence in target segment.

Our proposed model uses the discourse features and keeps the following features from the baseline model: first word pair, modal verb, and common word. Because our discourse features are supposed to represent a different aspect of argumentative relations than the word pair features do, comparing our proposed model with the baseline will reveal if the

[4]Each two components form two ordered pairs of {source, target}.

	Lexical	Discourse	Proposed
Accuracy	0.849	**0.857**	0.848
Kappa	0.320	0.317	**0.365**
Precision	0.707	**0.733**	0.706
Recall	0.634	0.626	**0.666**
F1:Support	0.401	0.388	**0.452**
F1:Non-support	0.914	**0.919**	0.912

Table 13: 10-fold performance. Best values are in bold.

discourse features can help predict the argumentative relations. In this first experiment, we use the window-size 2 heuristic[5] to create the context segment. In particular, context segment of a component consists of at most two preceding and two following sentences, and the covering sentence of the component; all sentences must be in the same paragraph. If the source and target segments overlap, overlapping sentences are kept for the source segment, and removed from the target segment.

We train both proposed and baseline models using LibLINEAR algorithm [Fan et al., 2008], and evaluate them using 10-fold cross validation. As suggested in [Stab and Gurevych, 2014b], no feature selection is performed. Results are show in Table 13. We first see that the Lexical model obtains F1:Support 0.401 and F1:Non-support 0.911 which are close to the 10-fold performances reported in [Stab and Gurevych, 2014b]. This validates our implementation of the lexical features. Second, the Discourse model that uses only discourse feature is shown comparable to the Lexical model. While Lexical model has higher recall, Discourse model prioritizes precision. However, Lexical model is better at identifying Support relations. Finally, the data shows that our proposed model obtains the best performance. Especially, our proposed model yields significantly higher kappa, recall, and F1:Support than the Lexical. This proves the effectiveness of our discourse features compared to the word

[5]Window-size 2 was chosen because paragraphs in Persuasive Essay Corpus have 3 sentences in average. Our next experiment will test the effect of window's size to prediction performance.

pair features.

5.6 SUMMARY

In order to improve argumentative relation classification, we propose to consider input unit in relations with surrounding sentences to enable advanced context features. Our pilot study has shown that discourse features extracted from the context segments are more efficient than word pair features for the argumentative relation classification task. Our next step will investigate more features from the context segment, e.g., common word with context sentences, textual entailment, similarity score set. Performance improvement by adding context features if happens will prove our second sub-hypothesis (H1-2), and along with our prior results (§3, §4) prove our first hypothesis (H1) about the effectiveness of contextual information in argument mining.

6.0 IDENTIFYING ARGUMENT COMPONENT AND ARGUMENTATIVE RELATION FOR AUTOMATED ARGUMENTATIVE ESSAY SCORING – PROPOSED WORK

6.1 INTRODUCTION

Application of argument mining in automated essay scoring has been actively investigated recently. [Ong et al., 2014] developed a rule-based model for identifying argument components in academic essays, and found a relation between a statistic of argument components and essay score. [Song et al., 2014] annotated student essays for critical responses to the argument provided in the writing prompt and used features extracted from the annotation to improve an existing essay scoring system. [Persing and Ng, 2015] developed a scoring model for the argument strength dimension on student essays and used features derived from output of an argument component identification model [Stab and Gurevych, 2014b].

Due to the availability of the Argument Strength Corpus [Persing and Ng, 2015], the first part of our study focuses on applying argument mining to the task of automatically scoring argument strength of essays. In the second part, we conduct a similar study for predicting holistic score of academic essays. In this study, we first use our trained argument mining models (§4, §5) to identify argument components and argumentative relations in the essays. [Persing and Ng, 2015] included only statistics of argument components as features for their scoring model. We however hypothesize that in addition to argument components, argumentative relations provide valuable information for determining the argument strength. We conduct different experiments to explore the use of argument mining output to support argumentative essay scoring.

6.2 ARGUMENT STRENGTH CORPUS

The Argument Strength Corpus [Persing and Ng, 2015] consists of 1000 argumentative essays collected from International Corpus of Learner English (ICLE) [Granger et al., 2009]. Each essay was scored for the strength of argument in the essay, using a numerical score from one to four at half-point increments. A summary of the scoring scoring rubric is given below:

Description of Argument Strength	Score
Essay makes a *strong argument* for its thesis and would convince most readers.	4
Essay makes a *decent argument* for its thesis and could convince some readers.	3
Essay makes a *weak argument* for its thesis or sometimes even *argues against it.*	2
Essay *does not make an argument* or it is often *unclear what the argument is.*	1

To evaluate the annotation accuracy, [Persing and Ng, 2015] selected 846 essays for multiple-graded by different annotators. They achieved inter-rater accuracy up to 0.89 when allowing annotators to agree on argument strength score within 1.0-point ranges. Table 14 shows the number of essays that receive each of the 7 scores for argument strength.

6.3 ARGUMENT MINING FEATURES FOR AUTOMATED ARGUMENT STRENGTH SCORING

In their scoring model, along with non-argument mining features (e.g., POS ngrams, semantic frames...) [Persing and Ng, 2015] included 7 *argument component features* based on the identification of major claim, claim and premise in the essay using the model developed in [Stab and Gurevych, 2014b]:

1. Number of major claims.

Score	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Number of essays	2	21	116	342	372	132	15

Table 14: Essay score distribution.

2. Number of claims.
3. Number of premises.
4. Fraction of paragraphs that contain either a claim or a major claim.
5. Fraction of paragraphs that contain at least one argument component of any kind.
6. Whether more than half of the essay's paragraphs contain no claims or major claims.
7. Whether more than one of the essay's paragraphs contain no components.

6.3.1 First experiment: impact of performance of argument component identification

Our first experiment re-implements the full model in [Persing and Ng, 2015], and replaces the 7 argument component features by the alternative features calculated based on output of our trained model for argument component identification. We test whether more accurate argument mining results yield more reliable scores.

6.3.2 Second experiment: impact of performance of argumentative relation identification

Our second experiment increments the first experiment by using our alternative argument component features, and adding argumentative relation features by our trained model for argumentative relation classification. In this experiment, we test whether adding argumentative relation information improves the scoring model. We train a prediction model to determine a Claim or Premise supports/attacks the MajorClaim. We extract following *argumentative relation features*:

1. Number of Claims that support the MajorClaim.
2. Number of Claims that attack the MajorClaim.
3. Number of Premises that support the MajorClaim.
4. Number of Premises that attack the MajorClaim.
5. Whether the first Claim or Premise supports the MajorClaim.
6. Whether the last Claim or Premise supports the MajorClaim.
7. Fraction of paragraphs that contain all support Claims/Premises.
8. Fraction of paragraphs that contain all attack Claims/Premises.
9. Whether more than half of the essay's paragraphs contain no support Claims/Premises.
10. Whether more than one of the essay's paragraphs contain no support Claims/Premises.
11. Sequence of two consecutive argumentative relations of the same paragraph. This feature captures the argumentation patterns in the essays. We expect that good essays may reveal different argumentation patterns than the bad essays.

In this experiment, we also use the argumentative relation classification model in [Stab and Gurevych, 2014b] as a baseline, so we can evaluate the impact of the argumentative relation identification performance to the scoring task.

6.3.3 Third experiment: only argument mining features

Our third experiment develops a scoring model using only our *argument mining features*, i.e., argument component features and argumentative relation features. This experiment evaluates whether argument mining output can predict argument strength score reliably, and comparably w.r.t the full model in [Persing and Ng, 2015].

6.4 ARGUMENT MINING FEATURES FOR PREDICTING PEER RATINGS OF ACADEMIC ESSAYS

The second part of our study aims at predicting the holistic score of the academic essays in our expanded Academic Essay Corpus. Each student essay in the second set of the

Essay set	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, 6)	[6, 7]
2011	0	0	0	6	11	10
2013	2	4	2	17	16	20

Table 15: Number of essays in each peer rating bin.

expanded Academic Essay Corpus (§5.2) was reviewed by student peers, and was given textual comments as well as numerical ratings. This set of academic essays consists of 61 essays collected in 2013, and 30 essays in 2011. Each essay in 2011 was rated by 3 student peers, and essays in 2013 were rated by 4 peers each.[1] We use the weighted average ratings as the final score of essays.[2] While the two rating rubrics have different descriptions,[3] both ask peer reviewers to evaluate the assigned writings using the same set of criteria and a 7-point scale. The point 1 means this essay needs work (did not achieve goals and failed to meet criteria) and point 7 means this essay was excellent (accomplished all goals and met all criteria). The criteria set is given below:

- Research question and background information
- Study design and hypothesis statements
- Convincing evidence-based justification for each research hypothesis
- Appropriate integration of conflicting research for at least one hypothesis

Three essays of the 2011 subset do not have peer ratings, thus we have total 88 essays with peer ratings for this study. The histogram of peer ratings for each subset is given in the Table 15.

[1]While each essay was assigned to a number of peer reviewers which was specified by the instructors, the actual number of reviewers for individual essays could be less because not all student reviewers completed their peer review assignment.

[2]The peer review practices were conducted using the SWoRD peer review system [Cho and Schunn, 2007]. The final peer rating of each essay is the weighted average score of individual ratings given by the peer reviewers, in which each peer rating is weighted by the rating accuracy determined automatically by SWoRD.

[3]The original description of rubrics are provided in the APPENDIX B

Given this data set of academic essays, we conduct experiments to test if argument mining features can predict the peer ratings. First, we follow the idea proposed in [Ong et al., 2014] and our experiment in §6.3.3 to build a regression model using statistics of argument components and argumentative relations to predict the peer rating of essays. Our second experiment applies our full model in §6.3.2 for this prediction task. Because our main goal is to investigate an application of argument mining features for essay score prediction, we do not tailor our proposed model previously designed for Argument Strength Corpus to make it better fit with the peer rating data. Therefore, the baseline model in this experiment is the scoring model in [Persing and Ng, 2015].

6.5 SUMMARY

By incorporating argument mining features, i.e., features derived from the identified argument components and argumentative relations, to an existing essay scoring model for argument strength, we explore a possibility of using argument mining outcome to predict argumentative essay scores. We expect that adding argument mining features, especially features computed based on the identified argumentative relations, improves essay scoring performance in both Argument Strength Corpus and peer rating datasets, and consequently proves our second hypothesis H2 about the usability of argument mining models in automated argumentative essay scoring.

7.0 SUMMARY

In this thesis, we propose context-aware argument mining models that use global and local contextual information to improve state-of-the-art argument mining performance. Our completed work on argument component identification (§3, §4) has shown that context features that exploit argument indicators and writing topic significantly improve the prediction performance. This proves our first sub-hypothesis of the effectiveness of context features in argument mining. Our first proposed work investigates features extracted from context segments to improve argumentative relation classification (§5). We plan to use proposed context features to replace generic linguistic features such as word pairs, syntactic production rules which are not generalized well across topic domains. Performance improvement by our incorporation of context features into the prior model for argumentative relation classification will prove our second sub-hypothesis of the effectiveness of context features in argument mining. Our second proposed work explores a possibility of using argument mining outcome to predict argumentative essay scoring (§6). We hypothesize that argument mining features derived from identified argument components and argumentative relations help build more accurate scoring model for argumentative essays. In order to prove this hypothesis, we will enhance an existing argumentative essay scoring model with argument mining features, and the improvement in score prediction will prove our second hypothesis of the usability of argument mining in automated argumentative essay scoring.

Upon the success in proving our hypotheses, the contributions of our research in this thesis are of two fold. First, we present robust argument mining models that identify argument component and argumentative relations in student essays, and work well in cross-essay and cross-topic settings. To the Computational Linguistic and Computational Argumentation communities, we offer state-of-the-art argument mining models, and put a step toward mak-

ing argument mining more practical and ready for real tasks. We believe that our proposed approaches are general and widely applicable to different text genres. In future we plan to apply context-aware argument mining to the context-dependent claim identification task in Wikipedia articles [Levy et al., 2014], and argumentative relation recognition in online debates [Boltužić and Šnajder, 2014].

Second, we develop an improved model for argumentative essay scoring that directly models written arguments in essay. This result promises new opportunities in Intelligent Tutoring System and Automated Essay Scoring areas for a system that is able to evaluate argumentative essays and give feedback of arguments used in the texts. Our long-term goal is to incorporate the argument mining and argumentative essay scoring models into an online peer review system to help both student authors and student reviewers. We expect this will be an essential support for promoting writing practice in school, especially for argumentative writing.

8.0 TIMELINE OF PROPOSED WORK

Time	Work	Deliverables
Mar–May, 2016	Build a context-aware argumentative relation classification model and evaluate the model using two student essay corpora (§5).	A conference submission about the proposed model on identifying argumentative relation between pair of argument components.
Jun–Aug	Extract argument component and argumentative relations from the Argument Strength Corpus to build a scoring model for argument strength of essays (§6).	A conference submission about argument mining in student essays and application in automated essay scoring.
Aug–Oct	Experiment with argument mining features to predict peer rating of essays (§6).	A journal submission about context-aware argument mining and application in argumentative essay scoring.
Oct–Jan, 2017	Thesis writing	Thesis ready to defend
Jan–Mar	Thesis revising	Complete thesis

APPENDIX A

LISTS OF ARGUMENT WORDS

List of 263 argument words extracted from the persuasive development set (6794 essays). Words are stemmed, named entities are replaced by their NER labels. Words are sorted in descending order of their probabilities returned by the LDA topic model.

> that the is of it peopl some be to other in are a on as this there for more believ view opinion both howev can with NUMBER than discuss not while have own think an or benefit would should argu may give no conclus advantag agre hand point who which issu could has reason do side argument differ from consid by such way certain fact those topic better say when individu instanc whether exampl abov been posit negat therefor effect much disagre societi clear sinc extent claim disadvantag result will rather moreov obvious far regard drawback nevertheless tend aspect concern still onli seem thus take well consider furthermor might number support strong controversi perspect becom bring hold outweigh case signific lead although benefici experi debat even alway import idea admit impact due base undeni second merit consequ group matter word into addit first come essenti compar henc sever especi wide convinc firm term one major particular doubt sum great evid despit approach up method deni these favor con out role begin anoth obtain each abl mention pros belief wherea influenc besid sens usual varieti phenomenon nowaday less inevit necessari former trend illustr contrari prefer viewpoint often seen rang main conclud befor critic possibl various greater numer plenti assert suitabl encourag oppon valuabl practic potenti vital mean latter opposit analyz crucial meanwhil same advoc accept relat contrast though capabl instead examin aforement enhanc put depend said harm easili turn acquir stand divers definit further accord worth general attent appropri undoubt total pivot effici regardless oppos known appar contend deal remain maintain nonetheless inde absolut

List of 315 argument words extracted from the academic development set (254 essays):

> the to of a in and that studi DATE PERSON is this more be on are it as with or NUMBER was by ORGANIZATION they an not will for were research like have would found than when their if also differ there from 's which other at these has result becaus hypothesi

observ how find could show been but support can howev anoth whether between what increas import less LOCATION previous may such those then mani predict both suggest conduct look them had hypothes while done base variabl way into all rate about did some question examin focus similar therefor test see determin so specif well compar general expect signific same oppos doe measur often due onli even believ understand order seem consid either set evid high better ani whi lead state possibl rather idea act much ask work given investig although sinc amount shown indic larg actual prior correl thus among say conclud depend come further addit exampl includ SET still play data purpos certain literatur explain involv attempt fact independ life regard overal made common assum natur part though sever design particular opposit form defin frequent main potenti creat just consist build topic answer strong psycholog across relev problem aim turn alway conflict befor tendenc littl great mention simpl evalu own off respect new appear within refer regardless avoid implic chanc exist assess reveal benefit knowledg yet down again she long conclus attribut various normal behind frequenc along necessari appli insight least whole extrem kind one e.g. ad must despit seek manner essenti wide instanc effici propos distinct equal start describ unlik goal probabl sourc combin categori remain obtain enough everyon analyz quick comparison move success confound circumst event impli real togeth limit open util taken statist absenc came reduc infer accur assumpt inclin extens contrari went slight divid ultim perhap inde difficult proven separ final contrast end half too last replic demograph

APPENDIX B

PEER RATING RUBRICS FOR ACADEMIC ESSAYS

Peer rating rubrics for the academic essays in 2011:

Phrased as questions the rubric criteria for the writing include:

- Was the research question described as important?
- Was the study contextualized and distinguished from prior research?
- Did the introduction include a brief high-level overview of study design and a clear statement of the hypotheses?
- Was there a convincing evidence-based justification for each research hypothesis?
- Did the introduction appropriately integrate conflicting research findings into a convincing argument for at least one hypothesis?

Rate the degree to which the writing accomplished the goals described in the Parts of the Paper document and met the criteria from the rubric. Use a seven-point scale going from 1 meaning this section needs work (did not achieve goals and failed to meet criteria) to 7 meaning this section was excellent (accomplished all goals and met all criteria) and with a 4 meaning this section was adequate (partially accomplished goals and met some criteria).

Peer rating rubrics for the academic essays in 2013:

Consider the following points when giving your rating:

- Central topic introduced and background information provided?
- Brief high-level overview of study design and clear statement of hypotheses?
- Appropriate integration of conflicting research findings into a convincing argument for at least one hypothesis?

BIBLIOGRAPHY

[Šarić et al., 2012] Šarić, F., Glavaš, G., Karan, M., Šnajder, J., and Dalbelo Bašić, B. (2012). TakeLab: Systems for Measuring Semantic Text Similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 441–448, Montréal, Canada. Association for Computational Linguistics.

[Barstow et al., 2015] Barstow, B., Schunn, C., Fazio, L., Falakmasir, M., and Ashley, K. (2015). Improving Science Writing in Research Methods Classes Through Computerized Argument Diagramming. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Pasadena, California.

[Bench-Capon and Dunne, 2007] Bench-Capon, T. J. and Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial intelligence*, 171(10-15):619–641.

[Bentahar et al., 2010] Bentahar, J., Moulin, B., and Bélanger, M. (2010). A Taxonomy of Argumentation Models Used for Knowledge Representation. *Artif. Intell. Rev.*, 33(3):211–259.

[Besnard et al., 2014] Besnard, P., Garcia, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., and Toni, F. (2014). Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4.

[Besnard and Hunter, 2008] Besnard, P. and Hunter, A. (2008). *Elements of Argumentation*. MIT Press.

[Biran and Rambow, 2011] Biran, O. and Rambow, O. (2011). Identifying Justifications in Written Dialogs by Classifying Text as Argumentative. *International Journal of Semantic Computing*, 5(4):363–381.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

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

[Brody and Elhadad, 2010] Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics.

[Burstein et al., 2004] Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, 25:27–36.

[Burstein et al., 2003] Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18(1):32–39.

[Cabrio et al., 2013] Cabrio, E., Tonelli, S., and Villata, S. (2013). From Discourse Analysis to Argumentation Schemes and Back: Relations and Differences. In *Computational Logic in Multi-Agent Systems*, volume 8143 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg.

[Cabrio and Villata, 2012] Cabrio, E. and Villata, S. (2012). Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 208–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Carlson et al., 2001] Carlson, L., Marcu, D., and Okurowski, M. E. (2001). Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Cho and Schunn, 2007] Cho, K. and Schunn, C. D. (2007). Scaffolded Writing and Rewriting in the Discipline: A Web-based Reciprocal Peer Review System. *Computers & Education*, 48(3):409–426.

[Du et al., 2014] Du, J., Jiang, J., Yang, L., Song, D., and Liao, L. (2014). Shell Miner: Mining Organizational Phrases in Argumentative Texts in Social Media. In *Proceedings of the 2014 IEEE International Conference on Data Mining*, ICDM '14, pages 797–802, Washington, DC, USA. IEEE Computer Society.

[Falakmasir et al., 2014] Falakmasir, M. H., Ashley, K., Schunn, C., and Litman, D. (2014). Identifying Thesis and Conclusion Statements in Student Essays to Scaffold Peer Review. In *Intelligent Tutoring Systems*, volume 8474 of *Lecture Notes in Computer Science*, pages 254–259. Springer International Publishing.

[Fan et al., 2008] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research*, 9:1871–1874.

[Feng and Hirst, 2011] Feng, V. W. and Hirst, G. (2011). Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics.

[Freeley and Steinberg, 2008] Freeley, A. and Steinberg, D. (2008). *Argumentation and Debate*. Cengage Learning.

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

[Funatsu et al., 2014] Funatsu, T., Tomiura, Y., Ishita, E., and Furusawa, K. (2014). Extracting Representative Words of a Topic Determined by Latent Dirichlet Allocation. In *eKNOW 2014, The Sixth International Conference on Information, Process, and Knowledge Management*, pages 112–117.

[Goudas et al., 2014] Goudas, T., Louizos, C., Petasis, G., and Karkaletsis, V. (2014). Argument Extraction from News, Blogs, and Social Media. In *Artificial Intelligence: Methods and Applications*, volume 8445 of *Lecture Notes in Computer Science*, pages 287–299. Springer International Publishing.

[Granger et al., 2009] Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). *International Corpus of Learner English v2*. Presses universitaires de Louvain, Louvain-la-Neuve.

[Guo et al., 2010] Guo, Y., Korhonen, A., Liakata, M., Silins, I., Sun, L., and Stenius, U. (2010). Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107, Uppsala, Sweden. Association for Computational Linguistics.

[Hall et al., 2009] 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.

[Hirohata et al., 2008] Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M. (2008). Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 381–388.

[Huang et al., 2013] Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. (2013). Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 2333–2338, New York, NY, USA. ACM.

[Klein and Manning, 2003] Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

[Knott and Dale, 1994] Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

[Levy et al., 2014] 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*, pages 1489–1500, Dublin, Ireland.

[Liakata et al., 2012] Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

[Lin et al., 2006] Lin, J., Karakos, D., Demner-Fushman, D., and Khudanpur, S. (2006). Generative Content Models for Structural Analysis of Medical Abstracts. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, BioNLP '06, pages 65–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lippi and Torroni, 2015] Lippi, M. and Torroni, P. (2015). Argument mining: a machine learning perspective. Buenos Aires, Argentina.

[Liu, 2012] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

[Louis and Nenkova, 2013] Louis, A. and Nenkova, A. (2013). What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association of Computational Linguistics*, 1:341–352.

[Madnani et al., 2012] Madnani, N., Heilman, M., Tetreault, J., and Chodorow, M. (2012). Identifying High-Level Organizational Elements in Argumentative Discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montrèal, Canada. Association for Computational Linguistics.

[Mochales and Moens, 2008] Mochales, R. and Moens, M.-F. (2008). Study on the Structure of Argumentation in Case Law. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 11–20, Amsterdam, The Netherlands, The Netherlands. IOS Press.

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

[Moens et al., 2007] Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic Detection of Arguments in Legal Texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA. ACM.

[Navigli, 2009] Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, 41(2):10:1–10:69.

[Newell et al., 2011] Newell, G. E., Beach, R., Smith, J., and VanDerHeide, J. (2011). Teaching and Learning Argumentative Reading and Writing: A Review of Research. *Reading Research Quarterly*, 46(3):273–304.

[Nguyen and Litman, 2015] 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*, pages 22–28, Denver, CO. Association for Computational Linguistics.

[Nguyen and Litman, 2016] Nguyen, H. and Litman, D. (2016). Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics. In *Proceedings 29th International FLAIRS Conference*, Key Largo, FL.

[Ong et al., 2014] Ong, N., Litman, D., and Brusilovsky, A. (2014). Ontology-Based Argument Mining and Automatic Essay Scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, Maryland. Association for Computational Linguistics.

[Pado et al., 2013] Pado, S., Noh, G., Stern, A., Wang, R., and Zanol, R. (2013). Design and Realization of a Modular Architecture for Textual Entailment. *Journal of Natural Language Engineering*, 1:1–34.

[Palau and Moens, 2009] Palau, R. M. and Moens, M.-F. (2009). Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

[Park and Cardie, 2014] Park, J. and Cardie, C. (2014). Identifying Appropriate Support for Propositions in Online User Comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

[Peldszus, 2014] Peldszus, A. (2014). Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, Baltimore, Maryland. Association for Computational Linguistics.

[Peldszus and Stede, 2013] 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 and Stede, 2015] 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)*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

[Persing and Ng, 2013] Persing, I. and Ng, V. (2013). Modeling Thesis Clarity in Student Essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

[Persing and Ng, 2015] 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*, pages 543–552, Beijing, China. Association for Computational Linguistics.

[Phan and Nguyen, 2007] Phan, X.-H. and Nguyen, C.-T. (2007). GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA). Technical report, Technical report.

[Pitler et al., 2009] Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics.

[Prasad et al., 2008] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco. European Language Resources Association (ELRA). ACL Anthology Identifier: L08-1093.

[Qazvinian and Radev, 2010] Qazvinian, V. and Radev, D. R. (2010). Identifying Non-explicit Citing Sentences for Citation-based Summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 555–564, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Rahimi et al., 2014] Rahimi, Z., Litman, D., Correnti, R., Matsumura, L., Wang, E., and Kisa, Z. (2014). Automatic Scoring of an Analytical Response-To-Text Assessment. In *Intelligent Tutoring Systems*, volume 8474 of *Lecture Notes in Computer Science*, pages 601–610. Springer International Publishing.

[Rus et al., 2013] Rus, V., Lintean, M., Banjade, R., Niraula, N., and Stefanescu, D. (2013). SEMILAR: The Semantic Similarity Toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 163–168, Sofia, Bulgaria. Association for Computational Linguistics.

[Séaghdha and Teufel, 2014] Séaghdha, D. . and Teufel, S. (2014). Unsupervised learning of rhetorical structure with un-topic models. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-14)*, Dublin, Ireland.

[Sardianos et al., 2015] Sardianos, C., Katakis, I. M., Petasis, G., and Karkaletsis, V. (2015). Argument Extraction from News. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO. Association for Computational Linguistics.

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

[Somasundaran and Wiebe, 2009] Somasundaran, S. and Wiebe, J. (2009). Recognizing Stances in Online Debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore. Association for Computational Linguistics.

[Song et al., 2014] Song, Y., Heilman, M., Beigman Klebanov, B., and Deane, P. (2014). Applying Argumentation Schemes for Essay Scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.

[Stab and Gurevych, 2014a] Stab, C. and Gurevych, I. (2014a). Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

[Stab et al., 2014] Stab, C., Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective. In Cabrio, E., Villata, S., and Wyner, A., editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 40–49, Bertinoro, Italy. CEUR-WS.

[Teufel and Moens, 2002] Teufel, S. and Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4).

[Teufel et al., 2009] Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards Discipline-independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, pages 1493–1502, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Toulmin, 1958] Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press Cambridge.

[Walton et al., 2008] Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

[Wang and Lan, 2015] Wang, J. and Lan, M. (2015). A Refined End-to-End Discourse Parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China. Association for Computational Linguistics.

[Xue et al., 2015] Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., and Rutherford, A. (2015). The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.