Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics

Huy Nguyen[1] and **Diane Litman**[2]
[1,2]Computer Science Department and [2]Learning Research Development Center
University of Pittsburgh, PA 15260
E-mail: [1]hvn3@pitt.edu, [2]dlitman@pitt.edu

Abstract

Argument mining systems for student essays need to be able to reliably identify argument components independently of particular essay topics. Thus in addition to features that model argumentation through topic-independent linguistic indicators such as discourse markers, features that can abstract over lexical signals of particular essay topics might also be helpful to improve performance. Prior argument mining studies have focused on persuasive essays and proposed a variety of largely lexicalized features. Our current study examines the utility of such features, proposes new features to abstract over the domain topics of essays, and conducts evaluations using both 10-fold cross validation as well as cross-topic validation. Experimental results show that our proposed features significantly improve argument mining performance in both types of cross-fold evaluation settings. Feature ablation studies further shed light on relative feature utility.

Introduction

Argument mining in text involves automatically identifying argument components[1] (e.g., *thesis, claim*) as well as argumentative relations between them (e.g., *support, attack*). Argument mining has been studied in a variety of text genres like legal documents (Moens et al. 2007), scientific papers (Teufel and Moens 2002; Liakata et al. 2012), and online comments and debates (Park and Cardie 2014; Boltužić and Šnajder 2014). In education, teaching argumentation and argumentative writing to students are in particular need of attention (Newell et al. 2011; Barstow et al. 2015), and argument mining in student essays is believed to promise novel opportunities for automated argumentative essay evaluation and feedback (Burstein, Chodorow, and Leacock 2004; Ong, Litman, and Brusilovsky 2014; Rahimi et al. 2014; Song et al. 2014).

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, Marcu, and Knight 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, Marcu, and Knight 2003; Lippi and Torroni 2015). 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 are desirable (Burstein, Marcu, and Knight 2003).

Stab and Gurevych (2014b) studied the argument component identification problem in persuasive essays, and used linguistic features like unigrams, bigrams, trigrams 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. Nguyen and Litman (2015) addressed that issue by replacing n-grams with a set of argument words[2] 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[3]. While their 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 Nguyen and Litman (2015) used features that abstracted over topic lexicons, nor performed cross-topic evaluation.

Our current study addresses the above limitations in three ways. First, we present new features to model not only indicators of argument language but also to abstract over essay topics. Second, 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 essay topics.

Through experiments, we aim to provide support for the

[1]An argument component is a text portion that has a specific role in forming the arguments in the text (Peldszus and Stede 2013).

[2]**Argument words** are commonly used in papers on different topics, e.g., '*think*', '*reason*', as opposed to **domain words** that are only used in same-topic papers, e.g., '*bystander*', '*education*'.

[3]Dependency parses were used to extract pairs of subject and main verb, and only those that did not contain domain words were kept, e.g., "*we-predict*", "*I-think*".

following hypotheses: *models enhanced with our new features will outperform baseline models* when evaluated using (1) 10-fold cross validation and (2) cross-topic validation; *our new models will demonstrate topic-robustness* in that (3) their cross-topic and 10-fold cross validation performance levels will be comparable.

Related Work

Argument mining studies of professional text, e.g., scientific articles, have often taken advantage of the presence of citations and structural information such as section headings for feature development (Teufel and Moens 2002; Liakata et al. 2012). Since student essays often do not have such information, associated argument mining studies have typically used more generic linguistic indicators of argument such as discourse connectives, n-grams and production rules (Burstein, Marcu, and Knight 2003; Stab and Gurevych 2014a), and argument words and argumentative subject–verb pairs (Nguyen and Litman 2015). No prior argument mining studies of student essays use features that abstract over essay topics to the best of our knowledge, although in scientific articles common words with title have been used (Teufel and Moens 2002).

Argument mining studies have often used seed lexicons, e.g., indicative terms for argumentation (Knott and Dale 1994), discourse connectives (Prasad et al. 2008), to represent the organizational shell of argumentative content. Recently, 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). In a similar vein, Nguyen and Litman (2015) post-processed LDA (Blei, Ng, and Jordan 2003) output to extract argument and domain words. We use Nguyen and Litman's algorithm to create one baseline for our evaluation.

Topic abstraction with lexical chains – sequences of related words that contribute to the continuity of meaning, have been successfully applied to areas such as summarization (Barzilay and Elhadad 1997) and topic detection and tracking (Hatch, Stokes, and Carthy 2000). In this study we only compute common words between essay sentences and with essay titles to model topic abstraction. In the future we will consider advanced methods such as using WordNet or thesauruses for tracking semantically similar topic words.

Student Essay Corpus

We use the student essay corpus compiled by Stab and Gurevych (2014a), which is available online, to evaluate our approach. The corpus contains 90 persuasive essays which are writing responses to the test questions on standardized tests, and are posted to *www.essayforum.com* by the authors for reviews. In the essays, the authors state their stances, i.e., *major claims*, towards the writing topics and validate those stances with convincing arguments consisting of *claims* (controversial statements) that support or attack the major claims, and *premises* that underpin the validity of the claims. While major claims in the data are related to the argument topics, i.e., context-dependent (Levy et al. 2014), the claims can be context-independent, i.e., not related directly to the topics (Lippi and Torroni 2015). Three experts annotated possible argument components at clause-level, i.e., *major claim, claim* and *premise*, within each sentence and achieved inter-rater accuracy 0.88 for argument component labels and Krippendorff's α_U 0.72 for argument component boundaries.

Major claim	Claim	Premise	None
90	429	1033	327

Table 1: Number of argument components of each type.

In the excerpt below from a persuasive essay, sentences are numbered for easy look-up and argument components are enclosed by tags.

> (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) 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*]$_{premise}$.

In this study, we aim to classify the argument components in persuasive essays as *Major Claim*, *Claim*, *Premise* or *None*. Numbers of argument components of different types are reported in Table 1.

Prediction Models

Stab14: We implement the model in (Stab and Gurevych 2014b) as the first baseline. A summary of features used in this baseline is given in Figure 1.

Nguyen15: Our second baseline is the model in (Nguyen and Litman 2015). We first reimplement the algorithm in (Nguyen and Litman 2015) to extract argument and domain words from a development dataset, which are used to derive two LDA-enabled lexical sets (**LDAlex**). *Argument words*: we use only unigrams that are argument words. *Argumentative subject–verb pairs*: we keep only subject–verb pairs that do not contain domain words. We create the Nguyen15 model by replacing n-grams and production rules in Stab14 using LDAlex features, as illustrated in Figure 1.

wLDA+4: Our proposed model (*wLDA+4*) is Nguyen15 (with the LDA supported features) expanded with 4 new feature sets extracted from the covering sentences of the associated argument components. To model the topic cohesion of essays, we include two common word counts:

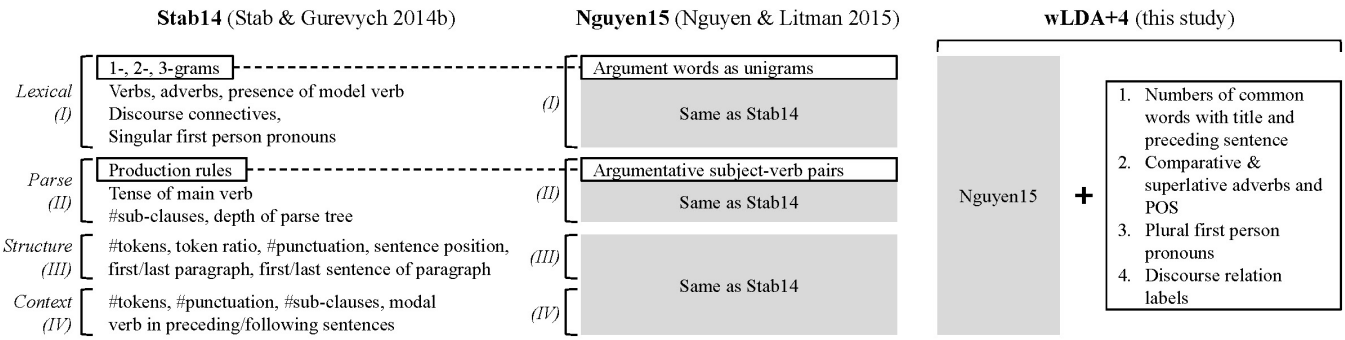

Figure 1: Feature illustration of **Stab14**, **Nguyen15** and **wLDA+4**. 1-, 2-, 3-grams and production rules in Stab14 are replaced by argument words and argumentative subject–verb pairs in Nguyen15. wLDA+4 extends Nguyen15 with 4 new feature sets.

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 objective/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, Louis, and Nenkova 2009) as new features:

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

• **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**: the 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]The temporal discourse relation was not used in (Stab and Gurevych 2014b) and thus is ignored in this study.

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). Words are stemmed.

Argument and Domain Word Extraction

In this section we briefly describe the algorithm to extract argument and domain words from development datasets using predefined argument keywords (see (Nguyen and Litman 2015) for more details). We use 6794 persuasive essays with post titles collected from *www.essayforum.com* by (Nguyen and Litman 2015) as the development data. The 10 most frequent words in the post titles that seemed argument related are used as argument keywords/seeds: *agree, disagree, reason, support, advantage, disadvantage, think, conclusion, result, opinion* (Nguyen and Litman 2015). Seeds of domain words are words in the titles but not argument keywords or stop words. There are 3077 domain seeds with 136482 occurrences. Each domain seed is associated with an in-title occurrence frequency f. All words in the development sets including seed words are stemmed, and named entities are replaced with the corresponding NER labels.

We run GibbsLDA++[5] implementation of LDA (Blei, Ng,

[5]http://gibbslda.sourceforge.net

and Jordan 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;[6] and combined weight $CW = AW - DW$. For example, topic 2 in Table 2 has AW = 5 (five argument keywords are not shown), DW = 0.15, CW = 4.85, $f(citi)$ = 0.0028 given its 381 occurrences in the 136482 domain seed occurrences in the titles. In each run we rank LDA topics by CW with the top topic has largest CW. 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.

Finally, we obtain 263 (stemmed) argument words and 1806 (stemmed) domain words. The argument words consist of keyword variants (e.g., *believe, viewpoint, argument, claim*), connectives (e.g., *therefore, however, despite*), and other stop words (Nguyen and Litman 2015). While the argument word list is greatly expanded from the argument keywords, the domain word list has only 6% not in the domain seed set. We note that many domain seeds are not present in the extracted domain words because words with occurrence less than 3 are removed from LDA topics.

Experimental Results

10-fold Cross Validation

We first conduct 10-fold cross validations to evaluate our proposed model and the baseline models. 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 train all models using SMO implementation of SVM in Weka (Hall et al. 2009) with top 100 features ranked by InfoGain algorithm on training folds. To reduce any effect of random folding, we perform 10 runs of 10-fold cross validations (10×10 cross-validation) and report the average results over 10 runs in Table 3 left section. 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 corpus is 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.

First we see that our model wLDA+4 significantly outperforms both Stab14 and Nguyen15 in all reported metrics. These results support our first hypothesis in that *our proposed features improve over both baselines using 10-fold cross validation*.

Regarding feature ablation results, we see that without LDA-enabled features woLDA's performance figures are all significantly worse than wLDA+4. 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. Nonetheless, adding the features computed from just argument keywords and domain seeds still helps Seed perform better than woLDA (except a lower precision). Finally, the fact that woLDA does not have ngram features or argument words makes it generally obtain lower performance than Nguyen15 and Stab14.

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

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. 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).

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 sets of by-fold performances.

As shown in the right section of Table 3, wLDA+4 again yields significantly higher performance than Stab14 in all metrics. Moreover we generally observe a larger performance gap between wLDA+4 and Stab14 in cross-topic validation than in 10-fold cross validation. WLDA+4 yields better performance than Nguyen15 for all metrics, with precision and recall improvements being significant. This shows a clear contribution of our new features in the overall performance, and supports our second hypothesis that *our new features improve the cross-topic performance compared to the baselines*.

With respect to feature ablation, our findings are consistent with the prior cross-fold results in that woLDA and Seed both have lower performance (often significantly) than wLDA+4, and Seed again generally outperforms woLDA.

Our next analysis compares wLDA+4 performance across the cross-fold and cross-topic experimental 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). We see that wLDA+4 yields higher performance for in cross-topic versus 10-fold cross validation with significantly higher precision and trending higher accuracy. These results support strongly our third robustness hypothesis that *our proposed model's cross-topic performance is as high as 10-fold cross validation performance*.

In contrast, Nguyen15'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. 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

	10-fold cross validation					Cross-topic validation				
Metric	Stab14	Nguyen15	woLDA	Seed	wLDA+4	Stab14	Nguyen15	woLDA	Seed	wLDA+4
Accuracy	0.787*	0.792*	0.780*	0.781*	**0.805**	0.780*	0.796	0.774*	0.776*	**0.807**
Kappa	0.639*	0.649*	0.629*	0.632*	**0.673**	0.623*	0.654+	0.618*	0.623*	**0.675**
Precision	0.741*	0.745*	0.746*	0.740*	**0.763**	0.722*	0.757*	0.751	0.734	**0.771**
Recall	0.694*	0.698*	0.695*	0.695*	**0.720**	0.670*	0.695*	0.681*	0.686*	**0.722**

Table 3: Cross validation results. Best values in bold. +: $p < 0.1$, *: $p < 0.05$ by T-test when comparing with wLDA+4.

a pool of top features (≈200), with an average rank for each. First we see that the proportion of argument words is 49% of pooled features, and the proportion of argumentative subject–verb pairs is 8%. 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. This shows the utility of our new feature sets. Especially the effectiveness of common word counts encourages us to study advanced topic-related features (Levy et al. 2014) in future work.

Performance on Held-out Test Sets

The experiments above used 10-fold cross-validation and cross-topic validation to investigate feature robustness. Note that this required us to reimplement both baselines as neither had previously been evaluated using cross-topic validation.[7] However, since both baselines were originally evaluated on single held-out test sets 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 4.

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)[8], 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 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.

[7]While Nguyen15 (but not Stab14) had been evaluated using 10-fold cross-validation, the data folds were not available.

[8]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 Nguyen15 or wLDA+4.

Metric	Stab best	Our SMO	Nguyen best	Our SMO	Our Lib-LINEAR
Acc.	0.77	**0.82**	0.83	0.82	**0.84**
Kappa	–	**0.68**	0.69	0.68	**0.71**
F1	0.73	**0.75**	0.76	0.73	**0.78**
Prec.	0.77	**0.79**	0.79	0.76	**0.81**
Recl.	0.68	**0.73**	0.74	0.70	**0.76**

Table 4: Model performance on test sets. Best values in bold.

Conclusions and Future Work

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. Our proposed model shows generality in that it yields performance improvement with both *cross-topic* and *10-fold cross* validations. Moreover, our model's cross-topic performance is even higher than cross-fold performances. Ablation 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 motivated by the student argumentative writing genre. That is, to achieve the best performance, the new features are a necessary supplement to the learned and noisy argument and domain words. Our next study will focus on argumentative relation classification, i.e., support vs. attack. We plan to use semantic relations and lexical cohesion to predict similarities and conflicts in the content.

Acknowledgments

This research is supported by NSF Grant 1122504. We thank the reviewers for their helpful feedback.

References

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*.

Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization*, 10–17.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

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

Burstein, J.; Chodorow, M.; and Leacock, C. 2004. Automated essay evaluation: The Criterion online writing service. *AI Magazine* 25:27–36.

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.

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, 797–802. Washington, DC, USA: IEEE Computer Society.

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.

Hatch, P.; Stokes, N.; and Carthy, J. 2000. Topic Detection, a new application for lexical chaining. In *The 22nd BCS IRSG Colloquium on Information Retrieval*, 94–103.

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

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

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

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.

Lippi, M., and Torroni, P. 2015. Context-independent Claim Detection for Argument Mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 185–191. Buenos Aires, Argentina: AAAI Press.

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*, 20–28. Montrèal, Canada: Association for Computational Linguistics.

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*, 11–20. Amsterdam, The Netherlands, The Netherlands: IOS Press.

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, 225–230. New York, NY, USA: ACM.

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

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*, 24–28. Baltimore, Maryland: Association for Computational Linguistics.

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

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.

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*, 683–691. Association for Computational Linguistics.

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.

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*, 601–610. Springer International Publishing.

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)*.

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*, 69–78. Baltimore, Maryland: Association for Computational Linguistics.

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*, 1501–1510. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

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)*, 46–56. Doha, Qatar: Association for Computational Linguistics.

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