Argument Mining for Understanding Peer Reviews

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, Lu Wang
Khoury College of Computer Sciences
Northeastern University
Boston, MA 02115
{hua.x, nikolov.m, badugu.n}@husky.neu.edu
luwang@ccs.neu.edu

Abstract

Peer-review plays a critical role in the scientific writing and publication ecosystem. To assess the efficiency and efficacy of the reviewing process, one essential element is to understand and evaluate the reviews themselves. In this work, we study the content and structure of peer reviews under the argument mining framework, through automatically detecting (1) argumentative propositions put forward by reviewers, and (2) their types (e.g., evaluating the work or making suggestions for improvement). We first collect 14.2K reviews from major machine learning and natural language processing venues. 400 reviews are annotated with 10,386 propositions and corresponding types of EVALUATION, REQUEST, FACT, REFERENCE, or QUOTE. We then train state-of-the-art proposition segmentation and classification models on the data to evaluate their utilities and identify new challenges for this new domain, motivating future directions for argument mining. Further experiments show that proposition usage varies across venues in amount, type, and topic.

1 Introduction

Peer review is a process where domain experts scrutinize the quality of research work in their field, and it is a cornerstone of scientific discovery (Hettich and Pazzani, 2006; Kelly et al., 2014; Price and Flach, 2017). In 2015 alone, approximately 63.4 million hours were spent on peer reviews (Kovanis et al., 2016). To maximize their benefit to the scientific community, it is crucial to understand and evaluate the construction and limitation of reviews themselves. However, minimal work has been done to analyze reviews' content and structure, let alone to evaluate their qualities.

As seen in Figure 1, peer reviews resemble arguments: they contain **argumentative propositions** (henceforth propositions) that convey reviewers' interpretation and evaluation of the research. Constructive reviews, e.g., review #2, often contain in-depth analysis as well as concrete suggestions. As a result, automatically identifying propositions and their types would be useful to understand the composition of peer reviews.

Review #1 (*rating*: 5, *# sentences*: 11)
[Quality: This paper demonstrates that convolutional and relational neural networks fail to solve visual relation problems ...]FACT [This points at important limitations of current neural network architectures where architectures depend mainly on rote memorization.]EVAL ... [Significance: This work demonstrates failures of relational networks on relational tasks. . .]FACT [Pros: Important message about network limitations.]EVAL [Cons: Straightforward testing of network performance on specific visual relation tasks.]EVAL . . .
Review #2 (*rating*: 5, *# sentences*: 10)
[The authors present two autoregressive models . . .]FACT. . . [In that context , this work can be viewed as applying deep autoregressive density estimators to policy gradient methods.]EVAL. . . [At least one of those papers ought to be cited.]REQ [It also seems like a simple, obvious baseline is missing from their experiments . . .]EVAL. . . [The method could even be made to capture dependencies between different actions by adding a latent probabilistic layer . . .]EVAL. . . [A direct comparison against one of the related methods in the discussion section would help]REQ. . .

Figure 1: Sample ICLR review excerpts. Propositions are annotated with types, such as FACT (fact), EVAL (evaluation), and REQ (request). Review #2 contains in-depth evaluation and actionable suggestion, thus is perceived to be of a higher quality.

Therefore, we propose *an argument mining-based approach to understand the content and structure of peer reviews*. Argument mining studies the automatic detection of argumentative components and structure within discourse (Peldszus and Stede, 2013). Specifically, argument types (e.g. evidence and reasoning) and their arrangement are indicative of argument quality (Habernal and Gurevych, 2016; Wachsmuth et al., 2017). In this work, we focus on two specific tasks: (1) **proposition segmentation**—detecting elementary argumentative discourse units that are propositions, and (2) **proposition classification**—labeling the propositions according to their types (e.g., evaluation vs. request).

Since there was no annotated dataset for peer reviews, as part of this study, we first collect 14.2K reviews from major machine learning (ML) and natural language processing (NLP) venues. We create a dataset, **AMPERE** (Argument Mining for PEer REviews), by annotating 400 reviews with $10,386$ propositions and labeling each proposition with the type of EVALUATION, REQUEST, FACT, REFERENCE, QUOTE, or NON-ARG.[1] Significant inter-annotator agreement is achieved for proposition segmentation (Cohen's $\kappa = 0.93$), with good consensus level for type annotation (Krippendorf's $\alpha_U = 0.61$).

We benchmark our new dataset with state-of-the-art and popular argument mining models to better understand the challenges posed in this new domain. We observe a significant drop of performance for proposition segmentation on AMPERE, mainly due to its different argument structure. For instance, 25% of the sentences contain more than one proposition, compared to that of 8% for essays (Stab and Gurevych, 2017), motivating new solutions for segmentation and classification.

We further investigate review structure difference across venues based on proposition usage, and uncover several patterns. For instance, ACL reviews tend to contain more propositions than those in ML venues, especially with more requests but fewer facts. We further find that reviews with extreme ratings, i.e., strong reject or accept, tend to be shorter and make much fewer requests. Moreover, we probe the salient words for different proposition types. For example, ACL reviewers ask for more "examples" when making requests, while ICLR reviews contain more evaluation of "network" and how models are "trained".

2 AMPERE Dataset

We collect review data from three sources: (1) openreview.net—an online peer reviewing platform for ICLR 2017, ICLR 2018, and UAI 2018 [2]; (2) reviews released for accepted papers at NeurIPS from 2013 to 2017; and (3) opted-in reviews for ACL 2017 from Kang et al. (2018).

[1] Dataset and annotation guideline can be found at http://xinyuhua.github.io/Resources/naacl19/.

[2] ICLR reviews are downloaded from the public API: https://github.com/iesl/openreview-py. UAI reviews are collected by the OpenReview team.

EVALUATION: Subjective statements, often containing qualitative judgment. Ex: *"This paper shows nice results on a number of small tasks."*
REQUEST: Statements suggesting a course of action. Ex: *"The authors should compare with the following methods."*
FACT: Objective information of the paper or commonsense knowledge. Ex: *"Existing works on multi-task neural networks typically use hand-tuned weights..."*
REFERENCE: Citations and URLs. Ex: *"see MuseGAN (Dong et al), MidiNet (Yang et al), etc "*
QUOTE: Quotations from the paper. Ex: *"The author wrote 'where r is lower bound of feature norm'."*
NON-ARG: Non-argumentative statements. Ex: *"Aha, now I understand."*

Table 1: Proposition types and examples.

Dataset	#Doc	#Sent	#Prop
Comments (Park and Cardie, 2018)	731	3,994	4,931
Essays (Stab and Gurevych, 2017)	402	7,116	6,089
News (Al Khatib et al., 2016)	300	11,754	14,313
Web (Habernal and Gurevych, 2017)	340	3,899	1,882
AMPERE	400	8,030	10,386

Table 2: Statistics for AMPERE and some argument mining corpora, including # of annotated propositions.

In total, $14,202$ reviews are collected (ICLR: $4,057$; UAI: 718; ACL: 275; and NeurIPS: $9,152$). All venues except NeurIPS have paper rating scores attached to the reviews.

Annotation Process. For proposition segmentation, we adopt the concepts from Park et al. (2015) and instruct the annotators to identify elementary argumentative discourse units on sentence or sub-sentence level, based on their discourse functions and topics. They then classify the propositions into five types with an additional non-argument category, as explained in Table 1.

400 ICLR 2018 reviews are sampled for annotation, with similar distributions of length and rating to those of the full dataset. Two annotators who are fluent English speakers first label the 400 reviews with proposition segments and types, and a third annotator then resolves disagreements.

We calculate the inter-annotator agreement between the two annotators. A Cohen's κ of 0.93 is achieved for proposition segmentation, with each review treated as a BIO sequence. For classification, unitized Krippendorf's α_U (Krippendorff, 2004), which considers disagreements among segmentation, is calculated per review and then averaged over all samples, and the value is 0.61. Among the exactly matched proposition segments, we report a Cohen's κ of 0.64.

Statistics. Table 2 shows comparison between AMPERE and some other argument mining datasets of different genres. We also show the number of propositions in each category in Table 3. The most frequent types are evaluation (38.3%) and fact (36.5%).

EVAL	REQ	FACT	REF	QUOT	NON-A	Total
3,982	1,911	3,786	207	161	339	10,386

Table 3: Number of propositions per type in AMPERE.

3 Experiments with Existing Models

We benchmark AMPERE with popular and state-of-the-art models for proposition segmentation and classification. Both tasks can be treated as sequence tagging problems with the setup similar to Schulz et al. (2018). For experiments, 320 reviews ($7,999$ propositions) are used for training and 80 reviews ($2,387$ propositions) are used for testing. Following Niculae et al. (2017), 5-fold cross validation on the training set is used for hyperparameter tuning. To improve the accuracy of tokenization, we manually replace mathematical formulas, variables, URL links, and formatted citation with special tokens such as <EQN>, <VAR>, <URL>, and <CIT>. Parameters, lexicons, and features used for the models are described in the supplementary material.

3.1 Task I: Proposition Segmentation

We consider three baselines. **FullSent**: treating each sentence as a proposition. **PDTB-conn**: further segmenting sentences when any discourse connective (collected from Penn Discourse Treebank (Prasad et al., 2007)) is observed. **RST-parser**: segmenting discourse units by the RST parser in Feng and Hirst (2014).

For learning-based methods, we start with Conditional Random Field (**CRF**) (Lafferty et al., 2001) with features proposed by Stab and Gurevych ((2017), Table 7), and **BiLSTM-CRF**, a bidirectional Long Short-Term Memory network (BiLSTM) connected to a CRF output layer and further enhanced with ELMo representation (Peters et al., 2018). We adopt the BIO scheme for sequential tagging (Ramshaw and Marcus, 1999), with O corresponding to NON-ARG. Finally, we consider **jointly modeling** segmentation and classification by appending the proposition types to BI tags, e.g., B-fact, with CRF (**CRF-joint**) and BiLSTM-CRF (**BiLSTM-CRF-joint**).

Table 4 shows that BiLSTM-CRF outperforms other methods in F1. More importantly, the performance on reviews is lower than those reached on existing datasets, e.g., an F1 of 86.7 is obtained by CRF for essays (Stab and Gurevych, 2017). This is mostly due to essays' better structure, with frequent use of discourse connectives.

	Prec.	**Rec.**	**F1**
FullSent	73.68	56.00	63.64
PDTB-conn	51.11	49.71	50.40
RST-parser	30.28	43.00	35.54
CRF	66.53	52.92	58.95
BiLSTM-CRF	**82.25**	**79.96**	**81.09***
CRF-joint	74.99	63.33	68.67
BiLSTM-CRF-joint	81.12	78.42	79.75

Table 4: Proposition segmentation results. Result that is significantly better than all comparisons is marked with $*$ ($p < 10^{-6}$, McNemar test).

	Overall	EVAL	REQ	FACT	REF	QUOT
With Gold-Standard Segments						
Majority	40.75	57.90	–	–	–	–
PropLexicon	36.83	40.42	36.07	32.23	59.57	31.28
SVM	60.98	63.88	**69.02**	54.74	**69.47**	7.69
CNN	**66.56***	**69.02**	63.26	**66.17**	67.44	**52.94**
With Predicted Segments						
Majority	33.30	47.60	–	–	–	–
PropLexicon	23.21	22.45	23.97	23.73	35.96	16.67
SVM	51.46	54.05	48.16	52.77	52.27	4.71
CNN	55.48	57.75	53.71	55.19	48.78	33.33
CRF-joint	50.69	46.78	55.74	52.27	**55.77**	26.47
BiLSTM-CRF-joint	**62.64***	**62.36***	**67.31***	**61.86**	54.74	**37.36**

Table 5: Proposition classification F1 scores. Results that are significant better than other methods are marked with $*$ ($p < 10^{-6}$, McNemar test).

3.2 Task II: Proposition Classification

With given proposition segments, predicted or gold-standard, we experiment with proposition-level models to label proposition types.

We utilize two baselines. **Majority** simply assigns the majority type in the training set. **PropLexicon** matches the following lexicons for different proposition types in order, and returns the first corresponding type with a match; if no lexicon is matched, the proposition is labeled as NON-ARG:

- REFERENCE: <URL>, <CIT>
- QUOTE: “, ”, ’
- REQUEST: *should, would be nice, why, please, would like to, need*
- EVALUATION: *highly, very, unclear, clear, interesting, novel, well, important, similar, clearly, quite, good*
- FACT: *author, authors, propose, present, method, parameters, example, dataset, same, incorrect, correct*

For supervised models, we employ linear **SVM** with a squared hinge loss and group Lasso regularizer (Yuan and Lin, 2006). It is trained with the top 500 features selected from Table 9 in (Stab and Gurevych, 2017) by χ^2 test. We also train a convolutional neural network (**CNN**) proposed by Kim (2014), with the same setup and pre-trained word embeddings from word2vec (Mikolov et al., 2013). Finally, results by joint models of CRF and BiLSTM-CRF are also reported.

F1 scores for all propositions and each type are reported in Table 5. A prediction is correct when both segment and type are matched with the true labels. CNN performs better for types with significantly more training samples, i.e., evaluation and fact, indicating the effect of data size on neural model's performance. Joint models (CRF-joint and BiLSTM-CRF-joint) yield the best F1 scores for all categories when gold-standard segmentation is unavailable.

4 Proposition Analysis by Venues

Here we leverage the BiLSTM-CRF-joint model trained on the annotated AMPERE data to identify propositions and their types in unlabeled reviews from the four venues (ICLR, UAI, ACL, and NeurIPS), to understand the content and structure of peer reviews at a larger scale.

Proposition Usage by Venue and Rating. Figure 2 shows the average number of propositions per review, grouped by venue and rating. Scores in $1-10$ are scaled to $1-5$ by $\lceil x/2 \rceil$, with 1 as strong reject and 5 as strong accept. ACL and NeurIPS have significantly more propositions than ICLR and UAI. Ratings, which reflect a reviewer's judgment of paper quality, also affect proposition usage. We find that reviews with extreme ratings, i.e., 1 and 5, tend to have fewer propositions.

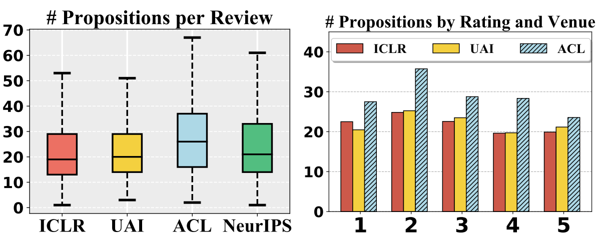

Figure 2: Proposition number in reviews. Differences among venues are all significant except UAI vs. ICLR and ACL vs. NeurIPS ($p < 10^{-6}$, unpaired t-test).

We further study the distribution of proposition type in each venue. As observed in Figure 3, ACL

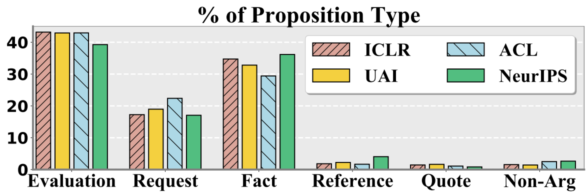

Figure 3: Distribution of proposition type per venue.

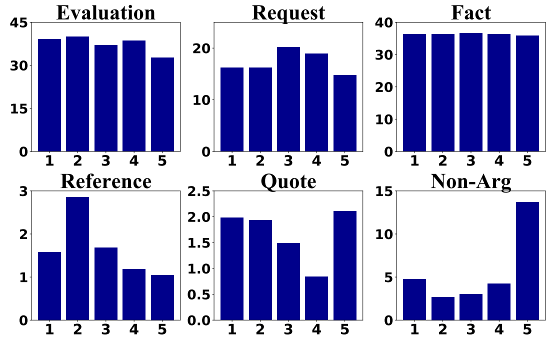

Figure 4: Distribution of proposition type per rating (in %) on AMPERE.

reviews contain more requests but fewer facts than other venues. Specifically, we find that 94.6% of ACL reviews have at least one REQUEST proposition, compared to 81.5% for ICLR and 84.7% for UAI. We also show proposition type distribution based on ratings in Figure 4. Reviews with the highest rating tend to use fewer evaluation and reference, while reviews with ratings of $3-4$ (borderline or weak accept) contain more requests. We further observe a sharp decrease of QUOTE usage in rating group 4, and a surge of NON-ARG for rating group 5, while FACT remains consistent across rating ranges.

Proposition Structure. Argumentative structure, which is usually studied as support and attack relations, reveals how propositions are organized into coherent text. According to Park and Cardie (2018), 75% of support relations happen between adjacent propositions in user comments. We thus plot the proposition transition probability matrix in Figure 5, to show the argument structure in AMPERE. The high probabilities along the diagonal line imply that propositions of the same type are often constructed consecutively, with the exception of quote, which is more likely to be followed by evaluation.

Proposition Type and Content. We also probe the salient words used for each proposition type, and the difference of their usage across venues. For each venue, we utilize log-likelihood ratio test (Lin and Hovy, 2000) to identify the represen-

	Evaluation	Request	Fact	Reference	Quote
All Venues	overall, unclear, not, contribution, seem, interesting	please, could, should, if, why, would, more, suggest	think, each, some, data, useful, written, proposes	<URL>, et, al., conference, paper, proceedings, arxiv	", paper, we, :, our
ICLR	network, general, acceptance, convinced, trained	network, appendix, recommend, because, novelty	training, results, work, then, image	deep, ;, nips, pp., speech	not, section, 4, 5, agent
UAI	quality, relevant, found, presentation, major	<VAR>, model, method, nice, column	stochastic, called, considers, sense, writing	artificial, discovery, etc., via, systems	–, second, column, processes, connections
ACL	weaknesses, strengths, so, word, main	consider, examples, further, models, proposed	word, method, words, proposed, embeddings	language, extraction, emnlp, computational, linguistics	
NeurIPS	theoretical, <EQN>, interest, practical, nips	following, clarity, address, significance, quality	<EQN>, maximum, may, comments, characters	for, see, class, detailed, guidelines	of, in, which, <EQN>, reviewer

Table 6: Salient words ($\alpha = 0.001$, χ^2 test) per proposition type. Top 5 frequent words that are unique for each venue are shown. "<EQN>", "<URL>", and "<VAR>" are equations, URL links, and variables.

	Eval	Req	Fact	Ref	Quot	Non-A
Eval	50.3	17.2	27.3	1.0	1.4	2.9
Req	32.2	41.6	19.4	1.8	2.3	2.8
Fact	33.5	11.0	51.2	1.3	0.9	2.0
Ref	15.0	10.8	18.0	50.9	3.6	1.8
Quot	31.2	23.6	25.5	1.3	12.1	6.4
Non-A	31.9	15.5	22.7	1.3	2.8	25.9

Figure 5: Proposition transition prob. on AMPERE.

tative words in each proposition type compared to other types. Table 6 shows both the commonly used salient words across venues and the unique words with top frequencies for each venue ($\alpha = 0.001$, χ^2 test). For evaluation, all venues tend to focus on clarity and contribution, with ICLR discussing more about "network" and NeurIPS often mentioning equations. ACL reviews then frequently request for "examples".

5 Related Work

There is a growing interest in understanding the content and assessing the quality of peer reviews. Authors' feedback such as satisfaction and helpfulness have been adopted as quality indicators (Latu and Everett, 2000; Hart-Davidson et al., 2010; Xiong and Litman, 2011). Nonetheless, they suffer from author subjectivity and are often influenced by acceptance decisions (Weber et al., 2002). Evaluation by experts or editors proves to be more reliable and informative (van Rooyen et al., 1999), but requires substantial work and knowledge of the field. Shallow linguistic features, e.g., sentiment words, are studied in Bornmann et al. (2012) for analyzing languages in peer reviews. To the best of our knowledge, our work is the first to understand the content and structure of peer reviews via argument usage.

Our work is also in line with the growing body of research in argument mining (Teufel et al., 1999; Palau and Moens, 2009). Most of the work focuses on arguments in social media posts (Park and Cardie, 2014; Wei et al., 2016; Habernal and Gurevych, 2016), online debate portals or Oxford-style debates (Wachsmuth et al., 2017; Hua and Wang, 2017; Wang et al., 2017), and student essays (Persing and Ng, 2015; Ghosh et al., 2016). We study a new domain of peer reviews, and identify new challenges for existing models.

6 Conclusion

We study the content and structure of peer reviews under the argument mining framework. AMPERE, a new dataset of peer reviews, is collected and annotated with propositions and their types. We benchmark AMPERE with state-of-the-art argument mining models for proposition segmentation and classification. We leverage the classifiers to analyze the proposition usage in reviews across ML and NLP venues, showing interesting patterns in proposition types and content.

Acknowledgements

This research is based upon work supported in part by National Science Foundation through Grants IIS-1566382 and IIS-1813341. We are grateful to the OpenReview team, especially Michael Spector, for setting up the API to facilitate review data collection. We also thank three anonymous reviewers for their constructive suggestions.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

Mathieu Blondel and Fabian Pedregosa. 2016. Lightning: large-scale linear classification, regression and ranking in python.

Lutz Bornmann, Markus Wolf, and Hans-Dieter Daniel. 2012. Closed versus open reviewing of journal manuscripts: how far do comments differ in language use? *Scientometrics*, 91(3):843–856.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. 2014. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521. Association for Computational Linguistics.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554. Association for Computational Linguistics.

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

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

William Hart-Davidson, Michael McLeod, Christopher Klerkx, and Michael Wojcik. 2010. A method for measuring helpfulness in online peer review. In *Proceedings of the 28th ACM international conference on design of communication*, pages 115–121. ACM.

Seth Hettich and Michael J Pazzani. 2006. Mining for proposal reviewers: lessons learned at the national science foundation. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 862–871. ACM.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208, Vancouver, Canada. Association for Computational Linguistics.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661. Association for Computational Linguistics.

Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. 2014. Peer review in scientific publications: benefits, critiques, & a survival guide. *EJIFCC*, 25(3):227.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Alexandros Komninos and Suresh Manandhar. 2016. Dependency based embeddings for sentence classification tasks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1490–1500. Association for Computational Linguistics.

Michail Kovanis, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. 2016. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS One*, 11(11):e0166387.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, 38:787–800.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Tavite M Latu and André M Everett. 2000. *Review of satisfaction research and measurement approaches*. Citeseer.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995. Association for Computational Linguistics.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210. ACM.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 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 (Volume 1: Long Papers)*, pages 543–552. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Simon Price and Peter A Flach. 2017. Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Susan van Rooyen, Nick Black, and Fiona Godlee. 1999. Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts. *Journal of clinical epidemiology*, 52(7):625–629.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255. Association for Computational Linguistics.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Ellen J Weber, Patricia P Katz, Joseph F Waeckerle, and Michael L Callaham. 2002. Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA*, 287(21):2790–2793.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200. Association for Computational Linguistics.

Stephen J Wright. 2015. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507. Association for Computational Linguistics.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

A Annotation Details

Data Selection. We select 400 reviews from the ICLR 2018 dataset for the annotation study. To ensure the subset is representative of the full dataset, samples are drawn based on two aspects: review length and rating score.

Table 7 shows the distribution of reviews with regard to their length in the full ICLR 2018 dataset and the subset we sampled for annotation (AMPERE). As can be seen, the distribution over five bins are consistent between AMPERE and full dataset. A similar trend is observed on rating distribution in Table 8.

A subset of the reviews also have revision history, which can be used as a proxy for opinion change and review quality in future work. To that end, we manually set the ratio of revised reviews vs. unrevised ones to 3:1 (c.f. 9:1 on the full ICLR2018 dataset), to ensure that enough revised reviews are being annotated. Notice that, in this study, we only consider the initial version of a review if any revision exists.

Length	(0,200]	(200,400]	(400,600]	(600,800]	(800,∞)
AMPERE	14.8%	35.5%	25.3%	10.0%	14.6%
ICLR2018	17.6%	39.3%	23.8%	11.4%	7.9%

Table 7: Review length distribution of the full ICLR 2018 dataset and AMPERE, which consists of 400 sampled reviews.

Rating	**1**	**2**	**3**	**4**	**5**
AMPERE	3.0%	32.5%	43.8%	19.3%	1.5%
ICLR2018	2.6%	32.5%	42.4%	20.6%	1.8%

Table 8: Review rating distribution of AMPERE and the full ICLR 2018 dataset.

Inter-annotator Agreement (IAA). To measure IAA, we first follow Stab and Gurevych (2017) to calculate the unitized Krippendorf's α_U (Krippendorff, 2004) for each review, and report the average for each type.

We further consider agreement on the proposition level. However, since the segmented proposition boundaries by two annotators do not always match, we only consider the exact matched segments for Cohen's κ. The agreement scores for each type are listed in Table 9.

	EVAL	REQ	FACT	REF	QUOT	NON-A	overall
α_U	0.51	0.64	0.60	0.63	0.41	0.18	0.61
κ	0.60	0.68	0.64	0.88	0.59	0.27	0.64

Table 9: Inter-annotator agreement for all categories.

Sample Annotations. We show examples of annotated propositions in Table 10.

B Experiments

B.1 Data Preprocessing

For preprocessing, we tokenize and split reviews into sentences with the Stanford CoreNLP toolkit (Manning et al., 2014). We manually substitute special tokens for mathematical equations, URLs, and citations or references. In total, 302 variables (`<VAR>`), 125 equations (`<EQN>`), 62 URL links (`<URL>`), and 97 citations (`<CIT>`) are identified in 400 reviews.

B.2 Training Details

For all models except CNN, we conduct 5-fold cross validation on training set to select hyperparameters.

CRF. We utilize the CRFSuite (Okazaki, 2007) implementation and tune coefficients C_1 and C_2 for ℓ_1 and ℓ_2 regularizer. For segmentation task the optimal setup is $C_1 = 0.0$ and $C_2 = 1.0$; for joint prediction, $C_1 = 1.0$ and $C_2 = 0.01$ is used.

BiLSTM-CRF. We experiment with implementation by Reimers and Gurevych (2017) with an extra ELMo embedding. Based on the cross

EVALUATION	The paper shows nice results on a number of small tasks.
	With its poor exposition of the technique, it is difficult to recommend this paper for publication.
	I like the general approach of explicitly putting desired equivariance in the convolutional networks.
	The paper covers a very interesting topic and presents some though-provoking ideas.
	I'm not sure this strong language can be justified here.
REQUEST	I would really like to see how the method performs without this hack.
	can the authors motivate this aspect better?
	I suggest using [hidelinks] for hyperref.
	More explanation needed here.
	In addtion -> In addition
FACT	Existing works on multi-task neural networks typically use hand-tuned weights for weighing losses across different tasks
	This work proposes a dynamic weight update scheme that updates weights for different task losses during training time by making use of the loss ratios of different tasks
	In this paper, the authors trains a large number of MNIST classifier networks with differing attributes (batch-size, activation function, no. layers etc.)
	This paper is based on the theory of group equivariant CNNs (G-CNNs), proposed by Cohen and Welling ICML'16.
REFERENCE	[1] Burnetas, A. N., & Katehakis, M. N. (1997). Optimal adaptive policies for Markov decision processes. Mathematics of Operations Research, 22(1) , 222-255
	VARIANCE-BASED GRADIENT COMPRESSION FOR EFFICIENT DISTRIBUTED DEEP LEARNING
	see MuseGAN (Dong et al), MidiNet (Yang et al), etc
	e.g. Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis, Yang et al.
QUOTE	The author wrote "where r is lower bound of feature norm"
	"In a probabilistic context-free grammar (PCFG), all production rules are independent"
	Quoting from its abstract: "Using commodity hardware, our implementation achieves $\sim$ 90% scaling efficiency when moving from 8 to 256 GPUs."
NON-ARG	Did I miss something here?
	Below, I give some examples
	are all the test images resized before hand?
	How was this chosen?

Table 10: Sample annotated propositions.

validation for both segmentation and joint learning, the optimal network architecture selected has two layers with 100 dimensional hidden states each, with dropout probabilities of 0.5 for both layers. The word embedding pre-trained by Komninos and Manandhar (2016) is chosen, as it outperforms GloVe embeddings (Pennington et al., 2014) trained either on Google News or Wikipedia.

SVM. We utilize SAGA (Defazio et al., 2014) implemented in the Lightning library (Blondel and Pedregosa, 2016) to learn a linear SVM optimized with Coordinate Descent (Wright, 2015). The coefficient for a group Lasso regularizer (Yuan and Lin, 2006) is set to 0.001 by cross validation.

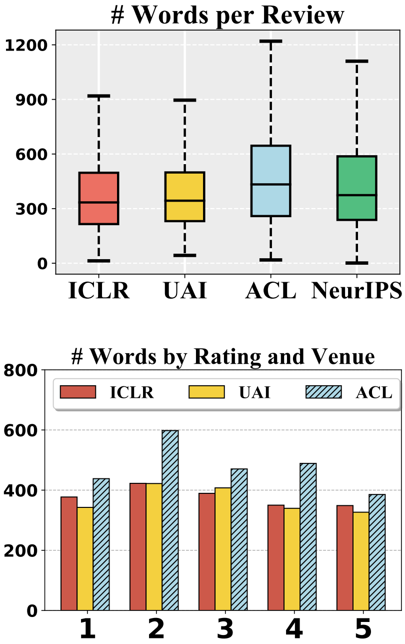

Figure 6: Word count in reviews by venue and rating. The word counts are significantly different between all venue pairs except UAI vs. ICLR and ACL vs. NeurIPS ($p < 10^{-6}$, unpaired t-test).

CNN. We implement the CNN-non-static variant as described in Kim (2014), with the following configuration: filter window sizes of $\{3,4,5\}$, with 128 feature maps each. Dropout probability is 0.5. 300 dimensional word embeddings are initiated with the pre-trained word2vec on 100 billion Google News (Mikolov et al., 2013).

C Further Analysis

Review Length by Venue and Rating. We compare review length of different venues in the top row of Figure 6. Unpaired t-test shows that ACL and NeurIPS have significantly longer reviews than UAI and ICLR ($p < 10^{-6}$), which is consistent with the trend for proposition counts, as described in Figure 2 in the paper.

We further group reviews by their ratings and display the average length per category in Figure 6. Again, we observe similar trends for the distribution of proposition count, where reviews with extreme ratings tend to be shorter.

Proposition Structure. We calculate the proposition type transition matrix as a proxy to uncover the local argumentative structure information. As is shown in Figure 7, propositions are more likely to be followed by propositions of the same type, while for NeurIPS the transition from reference to

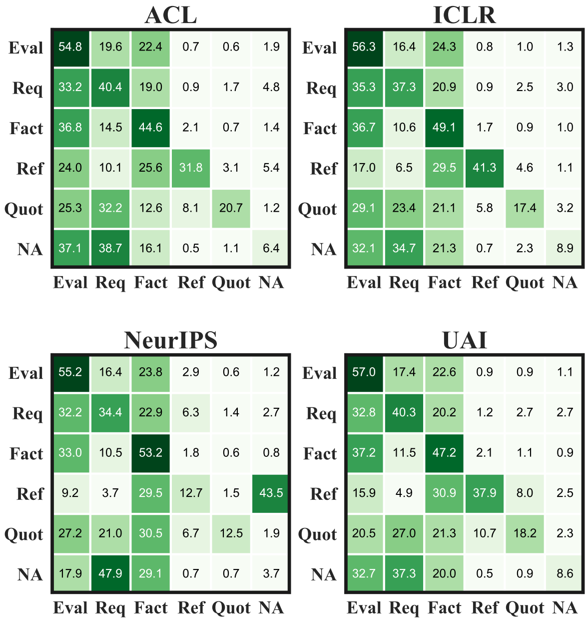

Figure 7: Proposition type transition matrix in different venues.

non-argument is much more prominent than other venues. A closer look at the dataset indicates that this might be because many formatted headers are mistakenly predicted as reference, e.g. "For detailed reviewing guidelines, see `<URL>`". They are usually followed by text such as "Comments to the author", which is predicted correctly as NON-ARG.