A Streamlined Method for Sourcing Discourse-level Argumentation Annotations from the Crowd

Tristan Miller[*†] and **Maria Sukhareva**[‡] and **Iryna Gurevych**[*†]

[*]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
https://www.ukp.tu-darmstadt.de/

[†]Research Training Group AIPHES
Department of Computer Science, Technische Universität Darmstadt
https://www.aiphes.tu-darmstadt.de/

[‡]NLP Lab
BMW Group, Munich
Maria.Sukhareva@bmwgroup.com

Abstract

The study of argumentation and the development of argument mining tools depends on the availability of annotated data, which is challenging to obtain in sufficient quantity and quality. We present a method that breaks down a popular but relatively complex discourse-level argument annotation scheme into a simpler, iterative procedure that can be applied even by untrained annotators. We apply this method in a crowdsourcing setup and report on the reliability of the annotations obtained. The source code for a tool implementing our annotation method, as well as the sample data we obtained (4909 gold-standard annotations across 982 documents), are freely released to the research community. These are intended to serve the needs of qualitative research into argumentation, as well as of data-driven approaches to argument mining.

1 Introduction

Empirical study of argumentation requires examples drawn from authentic, human-authored text. Likewise, the applications of computational argumentation, such as argument mining, can require significant amounts of argument-annotated data to achieve reasonable performance. However, this data can be challenging to obtain in sufficient quantity and quality, particularly for discourse-level argumentation. This is because discourse-level annotation schemes are necessarily complex with respect to discrimination and delimitation (i.e., the variety of markable elements in the text and how to define their boundaries), expressiveness (i.e., the need to tag relationships between annotated elements), and context weighting (i.e., the amount of context around markable units that needs to be considered) (Fort et al., 2012). Successfully applying such schemes typically requires expensive and laborious work by expert-trained annotators.

In this paper, we present a method that facilitates the application of one such discourse-level argument annotation scheme (Stab and Gurevych, 2014). This scheme has been widely cited and used in argumentation studies (e.g., Lippi and Torroni, 2015; Persing and Ng, 2015; Nguyen and Litman, 2015; Persing and Ng, 2016; Ghosh et al., 2016; Eger et al., 2017; Nguyen and Litman, 2018), and while it is fairly coarse-grained, it is expensive to apply to new texts. Our method breaks down the annotation process into incremental, intuitive steps, each focusing on a small portion of the overall annotation scheme. We apply this method in a crowdsourcing setup with annotators who receive no training other than a brief set of annotation guidelines, as well as in a more traditional setup with extensively trained local annotators. We find that agreement between the two groups increases sublinearly with the number of crowd annotators, achieving up to $\alpha_U = 0.52$ when using ten crowd workers. We release not only our sample data set (consisting of 4909 gold-standard argument component and argument relation annotations over 982 product reviews), but also the source code for the annotation tool itself, which will allow others to produce their own quantity- and quality-controlled annotated data sets.

Proceedings of NAACL-HLT 2019, pages 1790–1796
Minneapolis, Minnesota, June 2 - June 7, 2019.

2 Background and Previous Mork

While there exists a great diversity of argumentation theories in philosophy and logic (e.g., Toulmin, 2003; Freeman, 2011; Walton et al., 2012), they tend to agree that an argument can be decomposed into various interrelated components. Inspired by Freeman's (2011) theory of the macro-structure of argumentation, Stab and Gurevych (2014) broadly categorize these components as *claims* (the conclusions that the audience is persuaded to accept or reject), *premises* (additional information offered to support or attack a given claim), and the *major claim* (the one central claim that relates all other claims in an argument). Taken together, this can be conceptualized as a graph or tree structure, with vertices representing the *argument components* (major claims, claims, and premises) and the directed edges representing the *argument relations* (support and attack).

Stab and Gurevych (2014) annotate a collection of persuasive texts with this scheme, associating each argument component they identify with a contiguous span of text from the document. They report that the annotation process involved "several training sessions" with their annotators, including collaborative annotation of eight example documents in order to obtain a common understanding of the task. This level of effort is in line with what has been reported for other discourse-level argumentation schemes. For example, annotation studies using the Freemanesque schemes of Peldszus and Stede (2013), Li et al. (2017), Haddadan et al. (2018), and Musi et al. (2018) all required one or more lengthy training sessions guided by argumentation experts and up to six pages of written instructions.

Using existing methods to alleviate the knowledge acquisition bottleneck, such as incidental supervision (Roth, 2017), or pre-annotation (Fort and Sagot, 2010), could speed the work of annotators—possibly at the risk of introducing a training bias—but would not obviate the need for expert training. (In any case, pre-annotation has never, to our knowledge, been successfully applied to hard discourse-level tasks such as annotating argumentation structures.) The complexity of the annotation scheme also seemingly rules out the use of crowdsourcing (Howe, 2006) and gamification (von Ahn, 2006), which are geared towards microtasks that are quick and easy for humans. Though one previous study has decomposed a discourse-level scheme for use with crowdsourcing (Kawahara et al., 2014), the constraints it imposes (fixed-size annotations, maximum document length of three sentences) are too restrictive for argumentation annotation.

By contrast, the crowdsourcing approach of Sukhareva et al. (2016), while not concerned with discourse-spanning annotations, employs a few mechanisms that are relevant for our own task. Their approach, intended for the labelling of semantic verb relations, breaks down the annotation work into a series of hierarchical, atomic microtasks. Only those parts of the annotation instructions relevant to the current microtask are shown to the annotator. Furthermore, annotators are encouraged to think of connecting words ("specifically", "generally speaking", "in other words", etc.) that justify their relation annotations. As described in the following section, we adapt and extend these mechanisms for our own annotation method.

3 Annotation Method

Our approach to mitigating the knowledge acquisition problem is an iterative procedure by which annotators apply a distinct subset of the annotation scheme at each step. In this manner, complex discourse-level annotations are built up piecemeal in simple steps. The iterative annotation process is supported by an online JavaScript-based interface. Taken together, this allows the Stab and Gurevych (2014) annotation scheme to be applied even by untrained annotators in a crowdsourcing setup.

In the first step of the annotation process, annotators are presented with the complete argumentative text and asked to select the one phrase (i.e., an arbitrary sequence of words) that best represents the major claim, or else to indicate that there is no such passage.[1] In the second step, annotators are presented once again with the full argument, but with its major claim marked.[2] The annotators then select the claims—that is, all phrases that directly speak to the major claim, as well as whether those passages support or attack the major claim—or else

[1] If the user indicates in any step that there is no text span corresponding to the argument component type, we ask them to perform a short alternative task. This is to prevent faithless workers from taking the easy way out of the annotation task, but also to collect further annotations of interest to us.

[2] In the second and third steps, the marked annotation is not necessarily the one applied by the annotator in the previous step. In fact, as we explain below, in our study we source all annotations from a given step simultaneously, distill them into a gold standard, and mark these gold-standard annotations for the next step. With this setup, there is no need for a given annotator to participate in all three steps.

indicate that there are no such text spans. In the third step, annotators see the full argument with one of its claims marked. As with the previous step, annotators select text spans corresponding to the premises of the claim and indicate each premises's stance; they also have the option of reporting that the claim has no premises.

The annotation tool automatically enforces the restrictions that annotations must be contiguous, must begin and end on a word boundary, and cannot overlap with their siblings or ancestors. Crucially, the instructions given to annotators at each step of the process do not attempt to explain the entire annotation scheme but rather describe only the immediate annotation task in layman's terms. Furthermore, the tool attempts to make this task more intuitive for users by framing the second and third steps as a sentence completion task. An example of this is the interface for annotating claims (see Fig. 1). The full argumentative text (in this case, a product review) is shown on the left half of the screen, with the major claim marked, and we separately show a copy of the major claim on the right half of the screen. The user is instructed to extend the major claim with additional supporting or attacking information by appending a "because" or "but" clause, respectively. The user does this by pressing the "but" or "because" button below the major claim and then highlighting a sentence or phrase from the review.

4 Annotation Study

To assess the suitability of our annotation procedure, we applied it in a crowdsourcing setup. Measuring interannotator agreement for crowdsourced annotations is problematic, however, because there are typically a huge number of annotators, most of whom annotate only a tiny fraction of the data set. To gauge the reliability of our crowdsourced annotations, we instead conducted an experiment that compared them to those produced by expert-trained annotators.

For the experiment, we randomly selected 40 Amazon product reviews from the McAuley et al. (2015) data set—four from each of ten product categories. Each review was annotated for major claims by ten crowd workers; all 40 reviews were also annotated for major claims by a fixed group of three locally recruited annotators trained by argumentation experts.[3] We then converted the annotated reviews to BIO tokens (Ramshaw and Marcus, 1995) and applied the annotation aggregation/denoising tool MACE (Hovy et al., 2013) to select at most two gold-standard major-claim annotations per review, one from the crowd (crowd) and one from the trained annotators (train).[4] We then compared the crowd and train gold standards, one review at a time, using Krippendorff's (1995) α_U, a unitizing measure that considers the token-level boundaries of the text spans marked by each annotator. We repeated this process to obtain and evaluate crowd and train claim annotations on the train major claims, and then again for crowd and train premise annotations on the train claims.

Note that in the gold standards for some reviews, there may be no major claim, no claims associated with the major claim, and/or no premises associated with a given claim. In many cases this is because the annotators generally agreed that such argument components were not present in the text. However, in other cases the various annotators *did* identify such argument components, but the agreement among them was too low for MACE to output a gold-standard annotation. A quandary therefore arises when deciding how to treat reviews where *neither* the crowd nor the train gold standard contains a given type of argument component. There is no (easy) way of determining from the MACE output whether missing annotations are due to agreement or disagreement, and even if this information were available, it is not clear how it could be incorporated into the calculation of α_U. For this reason, we apply two different strategies for handling missing annotations, and provide separate α_U calculations for each. The first strategy, skip, disregards missing annotations, excluding them from the mean agreement calculation. The second strategy, agree, treats missing annotations as total agreement ($\alpha_U = 1$).[5]

When using all ten crowdsourced annotations per review and the agree strategy, we achieved mean α_U scores of 0.4104, 0.5231, and 0.4385 for major claims, claims, and premises, respectively. With the skip strategy, the respective scores are 0.4104, 0.4845, and 0.2201. As expected, these scores

[3]We engaged US-based workers from Amazon Mechanical Turk at the US federal minimum wage of $7.25/hour. Our expert-trained annotators were salaried research staff whose equivalent hourly rate was three to five times higher.

[4]MACE accepts a threshold value that is used to discard instances that cannot be confidently assigned a gold label; we set this to 0.9.

[5]It is not possible to treat the missing annotations as "total disagreement" because per Krippendorff (1995), α_U has no concept of this; there is no lowest disagreement score.

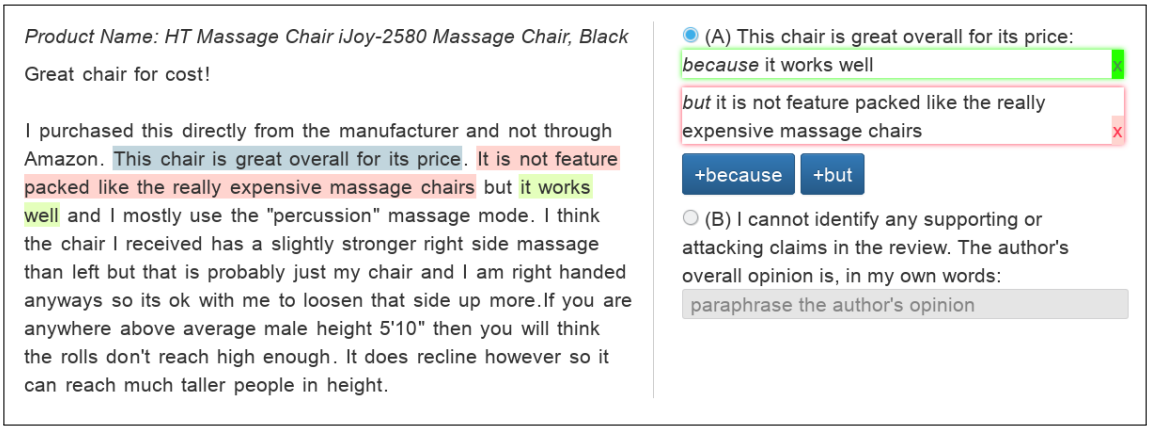

Figure 1: Annotation interface for the second step (claim annotation)

are lower than the agreement among extensively trained annotators reported by Stab and Gurevych (2014) ($\alpha_U = 0.7726, 0.6033, 0.7594$).[6] However, they are broadly comparable to interannotator agreement scores reported in similar (and in some cases, even simpler) discourse-level argument annotation studies with expert-trained annotators, such as Aharoni et al. (2014) ($\kappa = 0.4$), Musi et al. (2018) ($\kappa = 0.296$), and Li et al. (2017) ($\alpha_U = 0.2452$).

To measure how the number of crowd annotations impacts reliability, we performed an ablation study where we iteratively removed one crowd annotation at random from each review and repeated the MACE distillation and α_U calculation. The study was repeated 100 times and the resulting α_U scores averaged. The results are shown in Fig. 2, which plots the average α_U scores for major claims, claims, and premises when using one to ten crowd annotations per review. The plots are shown as error bars, where the top of the bar is the average agree score and the bottom is the average skip score. Reliability scores start to be uniformly positive with three annotations, with agreement for major claims and premises plateauing around seven annotations. The difference between the agree and skip scores is sizable only for premises.

5 Data Set and Software

Having satisfied ourselves that our method can produce reliable annotations via crowdsourcing, we applied it to a much larger subset of McAuley et al. (2015). The raw data consists of 982 English product reviews randomly sampled from the same ten product categories used in our evaluation study. For each argument component type in a review, we sourced annotations from five crowd workers, considering this to be an acceptable trade-off between annotation quality and cost. The MACE-produced gold standard contains 4909 annotations (937 major claims, 1134 claims, 852 premises, and 1986 argument relations). Our data set is distinguished from the review corpora of García Villalba and Saint-Dizier (2012) and Wyner et al. (2012) in that it is much larger, covers a broader range of product types, and is freely released under the CC BY 4.0 licence. It is comparable in size to but broader in scope than the Chinese-language hotel review corpus of Li et al. (2017).

Our data is distributed[7] as a set of XML Metadata Interchange (XMI) files, one per review, containing stand-off argument annotations that cross-reference the original texts from McAuley et al. (2015). (Because the original review texts are not available under a free licence, we do not include them in our distribution, but we provide a script for extracting them from the original corpus and merging them into our XMIs.) Also included is the JavaScript source for our annotation tool, as well as the Java source for preprocessing the raw data and postprocessing the annotations with MACE. This code can be used to crowdsource further annotated data sets using the McAuley et al. (2015) data at the desired level of quality. It could also be adapted to work with other raw corpora and other Freemanesque annotation schemes.

6 Conclusion

We have presented a scalable, simplified, iterative method for sourcing the discourse-level argumen-

[6] Apart from the fact that we used untrained annotators, the difference in agreement may also be due in part to our use of online user-generated content as opposed to student essays.

[7] https://github.com/UKPLab/naacl2019-argument-annotations

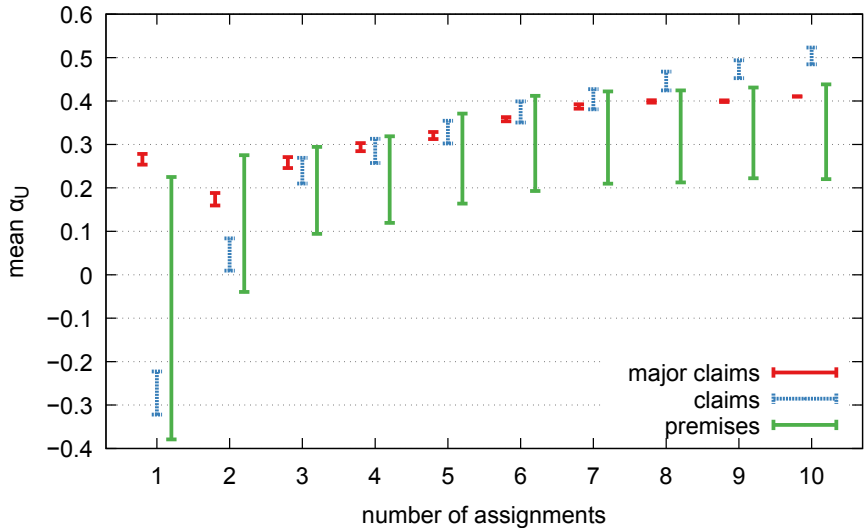

Figure 2: Results of the reliability study

tation annotations of Stab and Gurevych (2014), which may be adaptable to other annotation schemes based on Freeman's (2011) notion of argumentation. Our analysis shows that crowdsourced annotations obtained with our method yield substantial agreement with those obtained, with much greater effort, by expert-trained annotators. We have used our method to quickly and cheaply produce a large, argument-annotated data set of product reviews, which we freely release, along with the source code to our annotation interface and processing tools. Unlike with flat, context-free argument data such as that of Stab et al. (2018), training on our annotations would conceivably permit the identification not just of isolated argument components but of more complex argument structures. Our resources may also be of use for qualitative research on the linguistic features and rhetorical mechanisms of argumentative text (e.g., Peldszus and Stede, 2016).

For future work, we are investigating alternatives to MACE, which was designed for categorical annotations rather than the sequence labelling of our task. In particular, we are looking into the Bayesian method of Simpson and Gurevych (2018), which takes advantage of the sequential dependencies between BIO tags, and works more robustly with noisy, subjective data such as ours.

Acknowledgments

The authors thank Johannes Daxenberger and Christian Stab for many insightful discussions.

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional references 03VP02540 (ArgumenText) and 01UG1816B (CEDIFOR), and by the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1).

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the 1st Workshop on Argumentation Mining*, pages 64–68. Association for Computational Linguistics.

Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 11–22. Association for Computational Linguistics.

Karën Fort, Adeline Nazarenko, and Sophie Rosset. 2012. Modeling the complexity of manual annotation tasks: A grid of analysis. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 895–910.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the 4th Linguistic Annotation Workshop*, pages 56–63. Association for Computational Linguistics.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer Netherlands.

María Paz García Villalba and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *Computational Models of Argument*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 23–34. IOS Press.

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, pages 549–554. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2018. Annotation of argument components in political debates data. In *Proceedings of the Workshop on Annotation in Digital Humanities*, volume 2155, pages 12–16.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Ho vy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130. Association for Computational Linguistics.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired*, 14(6).

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 269–278. Dublin City University and Association for Computational Linguistics.

Klaus Krippendorff. 1995. On the reliability of unitizing continuous data. *Sociological Methodology*, 25:7–76.

Mengxue Li, Shiqiang Geng, Yang Gao, Shuhua Peng, Haijing Liu, and Hao Wang. 2017. Crowdsourcing argumentation structures in Chinese hotel reviews. In *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics*, pages 87–92.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 185–191. AAAI Press/International Joint Conferences on Artificial Intelligence.

Julian McAuley, Christopher Targett, Qinfeng "Javen" Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM.

Elena Musi, Manfred Stede, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. 2018. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*. European Language Resources Association.

Huy Nguyen and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28. Association for Computational Linguistics.

Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5892–5899. AAAI Press.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the 3rd Workshop on Argument Mining*, pages 103–112. 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, pages 543–552. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394. Association for Computational Linguistics.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 82–94. Association for Computational Linguistics.

Dan Roth. 2017. Incidental supervision: Moving beyond supervised learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 4885–4890. AAAI Press.

Edwin Simpson and Iryna Gurevych. 2018. Bayesian ensembles of crowds and deep learners for sequence tagging. *arXiv preprint*, 1811.00780.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1501–1510. Dublin City University and the Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674. Association for Computational Linguistics.

Maria Sukhareva, Judith Eckle-Kohler, Ivan Habernal, and Iryna Gurevych. 2016. Crowdsourcing a large dataset of domain-specific context-sensitive semantic verb relations. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2131–2137. European Language Resources Association.

Stephen E. Toulmin. 2003. *The Uses of Argument*, 2nd edition. Cambridge University Press.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2012. *Argumentation Schemes*, online edition. Cambridge University Press.

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *Computational Models of Argument*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 43–50. IOS Press.