Deep Argumentative Structure Analysis as an Explanation to Argumentative Relations

Paul Reisert† Naoya Inoue† Naoaki Okazaki† Kentaro Inui†
Tohoku University†
{preisert,naoya-i,okazaki,inui}@tohoku.ac.jp

1 Introduction

In recent years, there has been growing interest in the automatic analysis of argumentative texts [11], such as the identification of a *shallow* discourse structure for such texts by way of argumentative relation detection [1, 24, 13, etc.] and argumentative zoning [26, 10, 8, etc.]. Argumentative relation detection is the task of identifying argumentative relations (typically, support and attack relations) between discourse segments. Argumentative zoning is the task of identifying argumentative zones such as premise and claim.

Suppose we are analyzing the following debate text discussing the topic "*Should shopping malls be allowed to be open on holidays?*":

(1) *[I as an employee find it practical to be able to shop on weekends.]*$_{S_1}$ *[Sure, other people have to work in the shops on the weekend,]*$_{S_2}$ *[but they can have days off during the week]*$_{S_3}$

In this text, segment S_2 attacks segment S_1 and segment S_3 attacks, or undercuts, the relationship between S_2 and S_1 (argumentative relation detection). In another view, segments S_2 and S_3 serve as premises and segment S_1 as a claim (argumentative zoning).

The design of these shallow discourse analysis tasks has an advantage in their simplicity, which makes human annotation simple and reliable, achieving relatively high inter-annotator agreement [20]. Previous annotation studies thus have mainly focused on creating corpora for the identification of shallow discourse structures.

In this work, we propose a task design for going beyond shallow discourse structure by analyzing argumentative texts at a deeper level. For this, we consider a task of explaining *why* it makes sense to interpret each support/attack relation. For instance, in Example 1, a reasonable explanation why S_2 can be interpreted as an attack to S_1 is the following:

(2) (i) S_1 states that *(to) be able to shop on weekends* (relevant to the topic) is a positive thing. (ii) S_2 presupposes that *(to) be able to shop on weekends* will make *other people to work in the shops on the weekends* (consequence), and (iii) states that the consequence is an undesirable thing. (iv) Namely, S_1 states a positive aspect of one thing whereas S_2 states a negative consequence of the same thing; therefore, S_2 attacks S_1.

We consider the task of producing such an explanation (i.e. the author's logical reasoning) for each argumentative relation underlying a given argumentative text.

This direction of task design has several advantages. First, understanding the logical reasoning behind an argumentative text contributes toward determining implicit argumentative relations not indicated with an explicit discourse marker. Analysis of implicit discourse relations is a long-standing open problem in discourse analysis [9, 2, etc.]. We expect that this direction of research will provide a new approach to it. Second, identifying the logical reasoning will be useful for a range of argumentation mining applications. One obvious example is to aggregate multiple arguments and produce a logic-based abstractive summary. It will also be required in automatically assessing the quality of the logic structure of a given argumentation (cf. automatic essay scoring [23, 27]). Furthermore, it will be useful for generation of rebuttals in application contexts where a human and machine are cooperatively engaged in a debate (for decision support or education). Shallow discourse structure analysis, as assumed in previous work, suffers a large gap between what it produces and what is required in these useful potential applications.

2 Rhetorical patterns

2.1 Key idea

The key challenge of defining a computationally feasible explanation generation task is to establish the appropriate concept of *explanation of argumentative relation* (EAR) in a machine-friendly representation. In Argumentation Theory, a number of formal theories to describe an argumentative structure of a debate have been studied [7, 15, 28]. One prominent formalism is [28]'s *Argumentation Schemes*, which is composed of 65 common reasoning patterns together with a set of critical questions that assess an argument's acceptability. These theories are suggestive; however, it is not trivial how to operationalize such theories as a computational task. The main focus of the theories are purely in organizing the "nature" of human argumentation, where the level of machine-friendliness is not necessary.

To address this issue, we formalize the explanation generation task as the task of identifying a pattern of explanation (henceforth, *rhetorical pattern*) coupled with a slot-filling problem, where the slot is linked with an arbitrary phrase in an input text. Suppose we are generating an explanation to the rebuttal relation between S_2 to S_1 in Example 1. In-

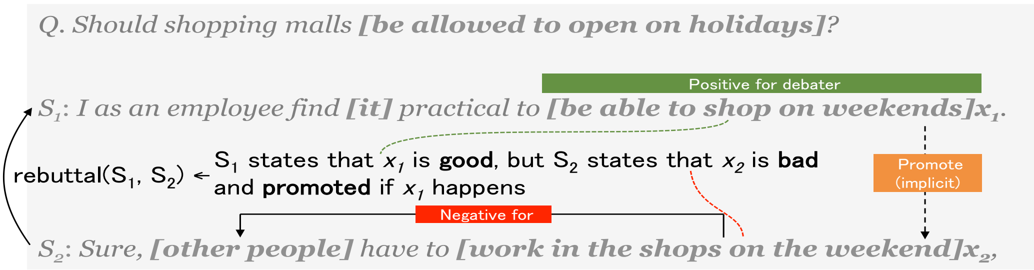

Figure 1: Pattern instantiation example

tuitively, as illustrated in Figure 1 (see *rebuttal*$(S_1, S_2) \leftarrow$ ...), the first step of this task is to identify a rhetorical pattern "*S_1 states that x_1 is good. However, S_2 states that if x_1 is brought about, consequence x_2, a bad thing, will happen. Therefore, S_2 is the rebuttal to S_1.*" The second task is to fill the slots x_1, x_2 in the pattern with a phrase from the text: x_1, x_2 needs to be filled with *be able to shop on weekends* and *work in the shops on the weekend*, respectively.

What should be noted is that such rhetorical patterns are not arbitrary but highly skewed. In fact, as we report in Section 4, the variety of logical reasoning underlying argumentative relations can be largely captured by only a small number of predefined patterns. We thus create an inventory of such major rhetorical patterns and annotate only typical patterns of logical reasoning with those predefined patterns, leaving uncommon patterns to be labeled simply as OTHER. These design decisions make our task of identifying deep argumentative structure considerably simple while going beyond previous task settings of shallow analysis.

2.2 Rhetorical patterns

We consider building our inventory of rhetorical patterns based on Walton's Argumentation Schemes. Among his 60 schemes, we create our first rhetorical patterns from the *Argument from Consequences* scheme: "[Premise] If a is brought about, *good* consequences will occur. [Conclusion] a should be brought about." We analyze the argumentative microtext corpus [13], a small, highly reliable corpus consisting of important ingredients for computational argumentation (see Section 4.1 for more details), and find that the *Argument from Consequences* scheme can be commonly used in debate argumentation.

As described in Section 4, our corpus study revealed that the majority of EARs can be represented by two factors: namely, sentiment polarity and bi-polar causality, where bi-polar causality includes *promote* and *suppress* relations. In terms of Walton's Argumentation Schemes, this is a variant of *Argument from Consequences*: "[Premise] If a is brought about, *good* consequences will occur. [Conclusion] a should be brought about." Bi-polar causality is especially useful for this scheme, where the *promotion* of something good can be considered a good consequence, or the *suppression* of something good can be considered a bad consequence. To create a set of rhetorical patterns, it is crucial to keep the framework as general as possible so that it can represent other types of EARs, as well as to cover the majority of EARs. This paper is the first step towards such a general theory.

Below, we list the generalized patterns for the *rebuttal* (S_1, S_2) relation in Figure 1, where the first pattern is ideal with the following arbitrary segments: x_1=*be able to shop on weekends* and x_2=*work in the shops on the weekend*.

1. S_1 states that x_1 is **good**, but S_2 states that x_2 is **bad** and was or will be **promoted** if x_1 happens or happened or does not happen or did not happen.
2. S_1 states that x_1 is **bad**, but S_2 states that x_2 is **good** and was or will be **promoted** if x_1 happens or happened or does not happen or did not happen.
3. S_1 states that x_1 is **good**, but S_2 states that x_2 is **good** and was or will be **suppressed** if x_1 happens or happened or does not happen or did not happen.
4. S_1 states that x_1 is **bad**, but S_2 states that x_2 is **bad** and was or will be **suppressed** if x_1 happens or happened or does not happen or did not happen.
5. S_1 states that x_1 is **good**, but S_2 states that x_2 is **good** and was or will be **not be promoted** if x_1 happens or happened or does not happen or did not happen.
6. S_1 states that x_1 is **bad**, but S_2 states that x_2 is **bad** and was or will **not be promoted** if x_1 happens or happened or does not happen or did not happen.
7. S_1 states that x_1 is **good**, but S_2 states that x_2 is **bad** and was or will be **not be suppressed** if x_1 happens or happened or does not happen or did not happen.
8. S_1 states that x_1 is **bad**, but S_2 states that x_2 is **good** and was or will be **not be suppressed** if x_1 happens or happened or does not happen or did not happen.

In this work, we also annotate whether the causality is explicitly stated with a linguistic expression (because, due to, etc) in the corresponding segment. For example, in Figure 1, it is implicit that X_2 is promoted by X_1. This implicitness comes from the debater's own knowledge, which is strongly assumed to be shared or inferred with the reader of the text. For the remaining argumentative relations (*support* and *undercut*), the sentiment in the above patterns are modified to fit the appropriate relation. For example, the first pattern in the above list would be modified to the following for the relation *support* (S_1, S_2): S_1 states that x_1 is good, and S_2 states that x_2 is good and was or will be promoted if x_1 happens or happened or does not happen or did not happen.

The rhetorical patterns for support correspond to Premise and Conclusion in *Argument from Consequences*. In addition, Critical Questions (CQs), which are questions to assess the quality of associated argumentation, correspond to rhetorical patterns of rebuttal and undercuts (e.g. CQ: *Are* *there other consequences of the opposite value that should be taken into account?*).

In total, we define eight variety of rhetorical patterns for each argumentative relation: $8 \times 3 = 24$ patterns in total.

3 Related Work

Conventionally, discourse structure analysis has been studied in the context of discourse relation identification. Earlier work includes discourse theories such as rhetorical structure theory which aim at creating coherent, tree-like structures for describing texts, where text units are typically adjacent [12]. Other theories such as cross-document structure theory focus on the identification of discourse relations spanned across multiple documents [18]. The Penn Discourse TreeBank, the largest manually annotated corpus for discourse relations, targeted both implicit and explicit relation detection for either adjacent sentences or clauses [17]. In the field of Argumentation mining, previous work has proposed several kinds of tasks such as structure identification task (e.g., support-attack relation detection) [1, 14]. In addition, a wide variety of corpora have been created in several domains including scientific articles, essays, and online discussions [5, 24]. These studies aim to capture the shallow structures of debates and do not try to explain a debater's reasoning.

One may think that stance classification is closely related to our task in terms that identifying the stance of a debate participant towards a discussion topic at a document or sentence level, and several corpora have been created for the task [6, 16, etc.]. However, since this direction of research focuses only on the classification of the stance polarity of a given paragraph or sentence, generating the explanation between two argumentative components has been out of scope.

Several argumentative corpora have been created for argumentation mining fields such as argument component identification, argument component classification, and structure identification [21, 22, 24], but none of them are like our current task setting. Reed et al. [21] annotated AraucariaDB corpus [19] with Walton's Argumentation Scheme, and the successive work [3] creates a machine learning-model to classify an argument into five sets of schemes. However, they do not annotate instantiations of variables in Argumentation Scheme, and do not report the inter-annotator agreement. Green [4] conducted preliminary work on identifying a set of argumentation schemes used in scientific articles based on Argumentation Scheme. However, they do not actually create a corpus.

4 Annotation Study

To examine the task feasibility, we conduct an annotation study and observe the coverage of patterns to create an annotated corpus. Our annotators consist of two fluent-English speakers.

4.1 Dataset

We explicitly annotate rhetorical patterns on top of the argumentative microtext corpus composed by [13]. The dataset includes a total of 112 texts,[1], each consisting of roughly five argumentatively relevant segments composed of a main claim and support and attack segments (see Figure 1 for a partial example taken from the corpus). This "layer-cake" style annotation facilitates the collaboration between our annotation and the traditional shallow discourse structure annotation (e.g., comparison, joint task definition). The argumentative function of each segment has been manually annotated as either support, rebuttal, or undercut. Prior to our final annotations reported for coverage and reliability, we conduct several trial annotations and improve our guidelines accordingly. In order to analyze EARs which are not covered by the present rhetorical patterns, we asked annotators to annotate them with a special pattern "OTHER".

We use brat [25], the general-purpose annotation tool for NLP, as an annotation interface. We provide the annotators with the original, segmented debate text. For each text, we provide its set of argumentative relations. For each relation, we list the appropriate, predefined list of rhetorical patterns. The annotators then select an arbitrary phrases from the segmented text when filling in the pattern slots (see x_1 and x_2 in Section 2.2).

4.2 Results and discussion

For testing the coverage of our proposed rhetorical patterns, we utilize 20 argumentative texts consisting of 87 relations. In total, 34 were agreed to be represented with at least one of the rhetorical patterns, whereas 17 were agreed not to be representable. This indicates a 67% coverage of agreed patterns. For the 17 relations which could not be captured by our proposed patterns, we found that non-representable relations can be mainly categorized into two patterns (frequency): *Presupposition* (10), *Analogical reasoning* (3).

Presupposition. One frequent non-representable relation is that one segment claims the truth value of presupposition of another segment. Consider the following example where S_3 supports S_1:

(3) *[No, the retirement age should be raised to 65 again.]*$_{S_1}$ *[People are getting older on average,]*$_{S_2}$ *[but they are not sicker and not duller because of it.]*$_{S_3}$ (b022)

S_3 can be interpreted as support to S_1 because S_3 presupposes that a precondition of *raising the retirement age* is that elderly people around the retirement age are well and states that the precondition holds.

Analogical reasoning. We found that analogical reasoning is used in some non-representable relations. In below, S_3 supports S_1:

(4) *[There should be a higher fine for dog dirt on the pavement.]*$_{S_1}$ *[Dog dirt is disgusting and a hygiene problem.]*$_{S_2}$ *[Also children, adults and other animals aren't allowed to leave droppings on the pavement.]*$_{S_3}$ (b032)

[1] We ignore 23 of the texts that did not include a debate-oriented topic question.

If we analyze the support relation closely, we can see that an analogy is stated; analogously to the fact that *children, adults and other animals aren't allowed to leave droppings on the pavement*, *dog dirt* should be punished. It is important in our future work that we expand our rhetorical patterns to allow for such representations.

Although the purpose of this study was to determine the coverage captured by our patterns, we were also interested in calculating the reliability of our current annotation scheme. Specifically, we calculate annotator agreement for each relation. Although our annotators arbitrarily selected phrases from the original debate text, given their complexity, we omit them from our current agreement calculation, and we focus our attention towards calculating the agreement between patterns alone. In our future work, we will also consider the agreement between the phrases. We use the following standards: i) *strict*, ii) *semi-lenient*, and iii) *lenient*. For each standard, we include sentiment and *promote* and *suppress* relations in the agreement calculation. For the strict standard, we consider a relation's explicitness. For the semi-lenient standard, we ignore this explicitness. For our lenient standard, we also ignore relation explicitness, and consider equivalent patterns. For the 8 patterns listed in Section 2.2, we consider each relation as equivalent.

Based on the above criteria, we report the kappa for 2 annotators as follows: i) .19 for the strict standard, ii) .29 for the semi-lenient standard, and iii) .43 for the lenient standard. Given the current complexities of this task and low reliability in previous work, we expected this to be an issue. In our future work, we will improve our guidelines and address other annotation problems.

5 Conclusion and Future Work

In this work, we composed a new task for deep argumentative structure analysis. We developed a small list of inference patterns for explaining argumentative relations. We examined the task feasibility by conducting an annotation study and reporting the coverage of argumentative explanations by the inference patterns. Our results indicate that argumentative relations can reasonably be explained.

In our future work, we will extend the amount of rhetorical patterns in order to increase our coverage. Furthermore, we plan to utilize other existing corpora in order to test our coverage in different domains. We will also continue to improve our guidelines and address any other problems. In the near future, we plan to make our annotated corpus and guidelines publicly available.

Acknowledgements This work was supported by JSPS KAKENHI Grant Numbers 15H01702, 16H06614, and CREST, JST.

References

[1] CABRIO, E., AND VILLATA, S. Generating abstract arguments: A natural language approach. In *Frontiers in Artificial Intelligence and Applications* (2012), vol. 245, pp. 454–461.

[2] CHEN, J., ZHANG, Q., LIU, P., QIU, X., AND HUANG, X. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2016), Association for Computational Linguistics, pp. 1726–1735.

[3] FENG, V. W., AND HIRST, G. Classifying Arguments by Scheme. In *Proc. of ACL* (2011), pp. 987–996.

[4] GREEN, N. L. Identifying argumentation schemes in genetics research articles. *NAACL HLT 2015* (2015), 12.

[5] HABERNAL, I., ECKLE-KOHLER, J., AND GUREVYCH, I. Argumentation Mining on the Web from Information Seeking Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing* (2014), pp. 26–39.

[6] HASAN, K. S., AND NG, V. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (2014), pp. 751–762.

[7] HASTINGS, A. C. A reformulation of the modes of reasoning in argumentation.

[8] HOUNGBO, H., AND MERCER, R. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. pp. 19–23.

[9] JI, Y., AND EISENSTEIN, J. One Vector is Not Enough : Entity-Augmented Distributed Semantics for Discourse Relations. *Transactions of the ACL 3* (2015), 329–344.

[10] LEVY, R., BILU, Y., HERSHCOVICH, D., AHARONI, E., AND SLONIM, N. Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland, August 23-29 2014* (2014), pp. 1489–1500.

[11] LIPPI, M., AND TORRONI, P. A Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology V* (2015), 1–25.

[12] MANN, W. C., AND THOMPSON, S. A. Rhetorical structure theory: Description and construction of text structures. In *Natural language generation*. Springer, 1987, pp. 85–95.

[13] PELDSZUS, A., AND STEDE, M. An annotated corpus of argumentative microtexts. In *Proceedings of the First Conference on Argumentation, Lisbon, Portugal, June. to appear* (2015).

[14] PELDSZUS, A., AND STEDE, M. Joint prediction in MST-style discourse parsing for argumentation mining. pp. 938–948.

[15] PERELMAN, C. The new rhetoric. In *Pragmatics of natural languages*. Springer, 1971, pp. 145–149.

[16] PERSING, I., AND NG, V. Modeling Stance in Student Essays. In *Proc. of ACL* (2016), pp. 2174–2184.

[17] PRASAD, R., DINESH, N., LEE, A., MILTSAKAKI, E., ROBALDO, L., JOSHI, A. K., AND WEBBER, B. L. The penn discourse treebank 2.0. In *LREC* (2008), Citeseer.

[18] RADEV, D. R. A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10* (2000), Association for Computational Linguistics, pp. 74–83.

[19] REED, C. Preliminary Results from an Argument Corpus. In *the IX Symposium on Social Communication* (2005), pp. 576–580.

[20] REED, C., GUREVYCH, I., STEIN, B., AND SLONIM, N. Nlp approaches to computational argumentation. In *ACL 2016 Tutorial* (2016), Association for Computational Linguistics.

[21] REED, C., AND PALAU, R. M. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation* (2008), pp. 91–100.

[22] RINOTT, R., DANKIN, L., ALZATE, C., KHAPRA, M. M., AHARONI, E., AND SLONIM, N. Show me your evidence–an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in NLP (EMNLP), Lisbon, Portugal* (2015), pp. 17–21.

[23] SONG, Y., HEILMAN, M., BEIGMAN KLEBANOV, B., AND DEANE, P. Applying Argumentation Schemes for Essay Scoring. *Proceedings of the First Workshop on Argumentation Mining*, 2011 (2014), 69–78.

[24] STAB, C., AND GUREVYCH, I. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (2014), pp. 1501–1510.

[25] STENETORP, P., PYYSALO, S., TOPIĆ, G., OHTA, T., ANANIADOU, S., AND TSUJII, J. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (2012), Association for Computational Linguistics, pp. 102–107.

[26] TEUFEL, S., SIDDHARTHAN, A., AND BATCHELOR, C. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3* (2009), Association for Computational Linguistics, pp. 1493–1502.

[27] WACHSMUTH, H., AL-KHATIB, K., AND STEIN, B. Using argument mining to assess the argumentation quality of essays.

[28] WALTON, D., REED, C., AND MACAGNO, F. *Argumentation schemes*. Cambridge University Press, 2008.