Computational Approaches to Fallacy Detection in Natural Language Arguments

Foulis, Matt

Published in:
Online Handbook of Argumentation for AI

Publication date:
2021

Document Version
Peer reviewed version

Computational Approaches to Fallacy Detection in Natural Language Arguments

Matt Foulis

Centre for Argument Technology, University of Dundee

Abstract

One way to improve the quality of an argument and the reasoning employed is to look at the ways in which it has fallen short or erred. Fallacy theory provides us with a set of guidelines to appraise an argument and discover the most common circumstances in which these errors occur.

The goal need not be, primarily, to criticise an argument and call it a day, but rather, to develop a starting point from which a sound argument can be developed. Additionally, an understanding of fallacies provides the audience of an argument with a set of tools to prevent them from being misled and to critically appraise the arguments presented.

This paper presents the goal of our research, namely, to employ fallacy theory in a computational context. More specifically, the work aims to: develop artificial intelligence models to automatically identify fallacies; and develop a computationally focused fallacy compendium to promote further work and collaboration in this field.

1 Introduction

First developed by Aristotle [Aristotle and Barnes, 1984], fallacy theory was for a long time sidelined in logic textbooks, often being relegated to an appendix or considered unworthy of study. Prior to his seminal work in the field, Hamblin viewed the treatment of fallacies as: "as debased, worn-out and dogmatic a treatment as could be imagined—incredibly tradition-bound, yet lacking in logic and historical sense alike" [Hamblin, 1970]. Post-Hamblin, there has been a renewed interest in fallacy theory with extensive work on topics such as classification, identification, and structure of the various fallacies. However, fallacies themselves are often viewed by the public through their association with internet debaters' cries of 'Straw man!' or 'Ad Hominem!'. Despite this, fallacy theory can provide a valuable framework for computational based work with user-facing outputs.

One current area of research in which fallacy theory can prove useful is the identification of fallacies in fake news and misinformation. Fake news has rapidly become a major political and societal problem with extensive research being carried out into its forms, creation, identification, and prevention. (See [Sharma et al., 2019] and [Zhou and Zafarani, 2020] for a survey of current techniques and approaches). Within fake news, fallacies such as *cherry picking*, *red herring*, and *straw man* are used in an effort to present misleading and incorrect arguments to the audience.

Misleading arguments are not only found within fake news, however. With so much of our engagement with arguments taking place on platforms such as social media, there is an opportunity for computational tools to analyse the data produced and provide assistance both with argument creation and an understanding of misleading or incorrect arguments. Advances in Argument Mining have demonstrated the benefits that computational solutions can provide in not only locating where an argu-

ment is within the text, but providing further insights and critical evaluation of the argument itself [Lawrence and Reed, 2020]. Whilst current argument mining techniques are valuable, there is an opportunity for automatic fallacy identification to tackle a different problem and provide important benefits to argument evaluation, analysis, and construction. Being able to automatically identify fallacies within a piece of text and present this information to a user would allow them to gain an understanding of why the argument is invalid or misleading and about the common ways in which this can occur. This knowledge could then inform their future interactions with arguments.

This paper will discuss how a focus on fallacies within a computational context can help with such tasks and outline the research planned for this work.

2 Method

One approach to understanding and dealing with fallacies is to consider them within the context of a dialogue. [Walton and Krabbe, 1995] argued that by looking at the shifts between dialogue types, it is possible to gain a better understanding of certain fallacies . If an *ad baculum* or 'appeal to force' argument is used in what started as a persuasion dialogue, this would be inappropriate and may be an indication that the argument is fallacious and an illicit shift of dialogue type has occurred.

As well as a shift between dialogue types, fallacies can also be identified as the breaking of a rule of the dialogue. The pragma-dialectic theory [Grootendorst and Van Eemeren, 1992] lays out the rules for an 'ideal model for a critical discussion' which aims to be a dialogue type for the resolution of an argument. [Visser et al., 2018] develop this model in a computational dialogue game as a demonstration of how such games could be used to prevent the use of fallacious arguments.

This understanding of fallacies as illicit shifts between dialogue types or as the breaking of a dialogue rule can inform the work in fallacy detection and analysis. Although many arguments may not be fallacious in the correct context, the move between dialogue types shifts the argument from reasonable to fallacious.

This shift of dialogue type is often seen in fake news, in which fallacious arguments are used which shift the news report from a neutral statement of the facts, to a persuasion type dialogue. [Martino et al., 2020] demonstrate this with their work on the SEMEVAL shared task of identifying 18 different propaganda techniques in news articles. Many of these techniques such as *appeal to emotion* would be legitimate arguments in the right context, but when present in a news article they become fallacious and potentially misleading to a reader.

Whilst there is some overlap with work on fake news and propaganda detection, an approach focused on fallacies is worthwhile and could assist in both fake news detection and, more generally, in argumentative reasoning and evaluation. Making use of fallacy theory allows the wealth of theoretical work to be employed in a computational context.

Fallacies must first be identified so that they can be assessed and dealt with appropriately. Whilst fallacy identification has typically required human annotators with an understanding of fallacy types, recent advances in ML have demonstrated the possibility of achieving human-level results in complex tasks. For example, the tasks of language understanding and question answer identification using the BERT model [Devlin et al., 2019]. A goal of this research is, therefore, to develop artificial intelligence (AI) techniques that are capable of automatically identifying fallacies within natural language arguments.

To guide the approach taken in developing these AI techniques, consideration must be given to the type and structure of the fallacies to be identified. There exist many classifications of fallacies, however, these are typically formulated from a linguistic or logic-based perspective. Whilst there are several useful fallacy classifications, there is an opportunity for the development of a computer-based fallacy approach.

A second goal of the research, therefore, is the creation of a compendium of fallacies, based on existing fallacy theory, but with computational tasks such as automatic identification via Machine Learning methods in mind. The goal of this compendium is to create a 'living document' in which to describe fallacy

structure, provide examples and edge-cases, discuss linguistic and semantic indicators, and additional detection resources such as knowledge bases and context data and metadata.

A requirement of both the identification task and the compendium will be suitable datasets for the training and testing of AI techniques. One possibility is the extension of existing datasets such as US2016 from [Visser et al., 2020] and the Internet Argument Corpus v2 from [Walker et al., 2012]. These datasets could be further expanded to include labeling for both fallacies and dialogue types, including the shifts between them. Another approach involves the creation of a new dataset. Online crowd-sourcing facilities such as Amazon Mechanical Turk [1] can allow large datasets to be created by participants based on selected criteria.

Once a suitable dataset is created, investigation into the development and testing of suitable artificial intelligence techniques can begin. As new techniques are discovered and incorporated from related fields and tasks, these can be incorporated into the fallacy compendium in an effort to promote collaborative work and progress in this task.

3 Discussion

This section presents an example of an argument type that can be classified as fallacious or not depending on the dialogue type and context. Knowing the intended dialogue type will allow us to understand if an argument has broken the rules of that type, indicating that a shift has occurred and that the argument is likely fallacious.

An appeal to emotion argument is one in which an attempt to manipulate the audience's emotions is made in an effort to bring about or support the conclusion. An example of this is *argumentum ad misericordiam* or appeal to pity. In some cases, such as a charity appeal for donations, an appeal to the audience's pity may be seen as an acceptable argumentative tactic. However, in other cases, an appeal to the emotions of the audience may be inappropriate for the dialogue type and therefore fallacious.

[1] Available at https://www.mturk.com/

Let us consider an example of a political debate between two candidates for a local election. During the course of the debate, one of the participants argues:

(1) "We must increase our foreign aid spending. There are people starving, do you not care about them?".

Whilst the conclusion of increasing foreign aid may be valid and the most appropriate course of action, the argument raised makes use of a fallacious appeal to the audience's emotion. This argument closely resembles that of a charity appeal for donations, but the difference in dialogue type indicates the fallaciousness of the argument. The dialogue begins as a debate type, which can be considered as a mixture of persuasion and eristic types. However, when presented with this appeal to emotion argument, a shift in the dialogue type can be identified, moving the dialogue from a debate, to that of a negotiation type dialogue. The goal of the speaker has moved from persuasion of the other party to ensuring their position or aim is pursued amongst the options available.

It should be noted that not all shifts in dialogue type are inherently illegitimate. There are cases where both parties may benefit from and pursue a negotiation after having initially begun the dialogue as a persuasion type. The problem, however, as Walton notes, is when the the shift is illicit and one of the participants is not aware of the shift [Walton, 2010]. He argues, therefore, that the argument must at this point be evaluated based on the rules of the original dialogue type.

From our example, the importance of contextual information on our ability to perform this evaluation can be seen. This use of context has increased in argument mining tasks related to fallacy detection. [Lugini and Litman, 2020] demonstrated how including previous argumentative discourse units (ADU) from a conversation helped to improve a model's ability to correctly classify argument components. [Opitz and Frank, 2019] found that looking at context only whilst masking the content of argumentative units may in fact lead to improved prediction results. These examples demonstrate the use that dialogue context could have on the ability to detect and

classify fallacies.

In addition to providing AI models with this dialogical contextual information, features and indicators that describe the presence of the fallacious arguments can be developed. With this combination, the models will be able to recognise both the presence of the potentially fallacious argument, as well as a classification of fallaciousness based on the dialogue type. This understanding of context embedded into the model will allow more accurate classification to be made and improve the usefulness of the system.

4 Conclusion

In this paper, a plan of research which aims to make use of fallacy theory in a computational context has been discussed. By viewing fallacies as an illicit shift between dialogue types or as a breaking of the dialogue type's rule, this information can be used in combination with current and emerging artificial intelligence techniques to guide the work of automatic identification. Further discussed was a plan to develop a fallacy compendium to encourage collaboration and develop new approaches and methods in this field.

Acknowledgements

The author would like to thank Chris Reed and Elena Musi for their continued help and guidance. This work was funded in part thanks to ESRC grant no. ES/V003909/1.

References

[Aristotle and Barnes, 1984] Aristotle and Barnes, J. (1984). *Sophistical Refutations*. Princeton University Press.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Grootendorst and Van Eemeren, 1992] Grootendorst, F. v. E. and Van Eemeren, F. (1992). Argumentation, communication, and fallacies. a pragma-dialectical perspective.

[Hamblin, 1970] Hamblin, C. L. (1970). *Fallacies*. Methuen London.

[Lawrence and Reed, 2020] Lawrence, J. and Reed, C. (2020). Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

[Lugini and Litman, 2020] Lugini, L. and Litman, D. (2020). Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1475–1480.

[Martino et al., 2020] Martino, G. D. S., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., and Nakov, P. (2020). Semeval-2020 task 11: Detection of propaganda techniques in news articles. Barcelona (online). International Committee for Computational Linguistics.

[Opitz and Frank, 2019] Opitz, J. and Frank, A. (2019). Dissecting content and context in argumentative relation analysis.

[Sharma et al., 2019] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3):1–42.

[Visser et al., 2018] Visser, J., Budzynska, K., and Reed, C. (2018). A critical discussion game for prohibiting fallacies. *Logic and Logical Philosophy*, 27(4):491–515.

[Visser et al., 2020] Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., and Reed, C. (2020). Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

[Walker et al., 2012] Walker, M., Tree, J. F., Anand, P., Abbott, R., and King, J. (2012). A corpus

for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

[Walton, 2010] Walton, D. (2010). *The Place of Emotion in Argument*. Pennsylvania State University Press.

[Walton and Krabbe, 1995] Walton, D. and Krabbe, E. (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press.

[Zhou and Zafarani, 2020] Zhou, X. and Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5):1–40.