PhD Topic: Fallacy Detection in Online Argumentation

Oana Balalau and Fabian Suchanek
Inria, Télécom Paris, Institut Polytechnique de Paris

1 Context

Humans use argumentation daily to evaluate the validity of new ideas, convince an addressee, or solve a difference of opinion. An **argument** contains a statement to be validated (a proposition also called claim or conclusion), a set of backing propositions (called premises, which should be accepted ideas), and a logical connection between all the pieces of information presented that allows the inference of the conclusion. People might agree on the evidence (i.e., the facts) but might disagree on what decision or action to take. In this work, we will focus on **fallacies**: weak arguments that seem convincing, however, their evidence does not prove or disprove the argument's conclusion. Fallacies are generally divided into formal and informal fallacies. Formal fallacies are arguments that can be easily represented as invalid logical formulas, such as affirming the consequent, which is a wrong application of modus tollens. For example, we know that if someone robs a bank without being caught, they are rich. *"Steve Jobs was rich, hence he robed a bank"* is a fallacy. Although many informal fallacies can also be represented as invalid arguments, informal fallacies are easier to describe and understand without resorting to logical representations [Hansen, 2020]. For example, the argument *"Pharmaceutical companies are lying to us to increase their revenues, vaccine V is not safe to use."* is an informal fallacy, the ad hominem fallacy, in which we attack not the claim, *vaccine V is not safe to use*, but the person making the claim, in this case the pharmaceutic industry.

Fallacy detection is part of argumentation mining, the area of natural language processing dedicated to extracting, summarizing and reasoning over human arguments [Lawrence and Reed, 2019]. The task is closely related to propaganda detection, where propaganda consists of a set of manipulative techniques, such as fallacies, used in a political context to enforce an agenda [Da San Martino et al., 2019]. Fallacy detection has been approached so far only from a text classification perspective, using deep neural networks to pick up language patterns [Sahai et al., 2021]. However, when humans classify arguments as fallacies, they try to reason based on the evidence and claim presented and decide if there is an inference that can be made. Efforts have been to train neural classifiers on reasoning tasks, such as common sense reasoning, *"if someone robs a bank without being caught, they are rich"*, or inferring a conclusion from a set of premises [Helwe et al., 2021].

The **goal of this thesis is to improve fallacy detection** in natural language, by leveraging both language patterns but also additional information, such as common sense knowledge, encyclopedic knowledge and logical rules. To achieve this we will focus on how fallacies can be represented[1] and how we can classify reasoning patterns in argumentation.

The PhD candidate will be hired by Télécom Paris, after a short internship of 1-3 months.

2 Thesis Roadmap

Current state-of-the-art techniques for fallacy detection [Da San Martino et al., 2019, Sahai et al., 2021] use only language patterns that are picked up by architectures such as BERT [Devlin et al., 2019], and depending on type of fallacy that we are trying to classify, have an $F1$ score between 30% to 70% for token-level classification, showing a very unequal performance. A closer analysis [Sahai et al., 2021], showed that the classifiers were most likely finding simple patterns when they were achieving a high score, such as words related to a certain topic. Pretrained language models have acquired linguistic knowledge during training, however detecting a fallacious argument requires a better representation of the argument and of the context in which this argument is made. For this, we have to solve two tasks:

[1] https://nordf.telecom-paris.fr/en/

1. **Context representation.** In order to determine if an argumentation is correct, one might need to search for additional information not mentioned in the argument. Such information could be in the form of logical rules, common sense knowledge or encyclopedic knowledge. For this, we should **first represent the original argument as a structured information, and then augment it with information from other sources**. For example, the argument "Pharmaceutical companies are lying to us to increase their revenues, vaccine V is not safe to use.", can be represented as two triples (Pharmaceutical companies, lie, people) and (Vaccine V, is, not safe), plus representing that the second triple is supported by the first one. The triples do not contain any information on how the pharmaceutical companies and the vaccine V are connected, so we should add more information that could allow us to understand the rule between them. From a knowledge base, such as Yago or Wikidata, we could add the knowledge that "Vaccine V was produced by company A" and "Company A is a pharmaceutical company", thus finding the missing link.

2. **Argumentation Scheme.** Once we have represented the argument and its context, we move to understanding if the argument is correct. An argument is composed of the claim, the premises, and an inference rule. The inference rule describes the reasoning method of the argument. This can be either a method that is valid under any condition (such as a modus ponens or modus tollens), or a method that is defensable given certain conditions[2] (experts should state true statements in their domain of expertise). Such arguments are classified by their argumentation scheme [Macagno et al., 2017]. Most argumentation schemes can be attacked by critical questions [Walton, 2005]. The speaker has to be able to answer these critical questions to prove that the argument is not a fallacy. One example of an argumentation scheme is the ad hominem argument [Walton, 2010]: *premise: P is a person of bad character, conclusion: P's argument should not be accepted.* This argument is not always a fallacy and in order to investigate if it is correct we need to answer a few critical questions, such as: *i*) How correct is the attack on the character of P? *ii*) Is the issue of character relevant in the argument discussed? Another example of an argumentation scheme is the argument from expert opinion [Walton, 2005]. It has the premises *E is an expert in domain D, E asserts that A is known to be true, A is within D* and the conclusion *therefore, A may plausibly be taken to be true.* Some critical questions for this scheme are: *i*) Trustworthiness: Is E personally reliable as a source? *ii*) Backup Evidence: Is E's assertion based on evidence? **In conclusion, our goal is given a claim and a premise, predict the argumentation scheme.**

Once the argumentation scheme has been determined, we will have to see whether we can answer the critical questions. We might also need additional knowledge acquisition to check factual claims. This will then finally give us an estimate of how correct the argument is.

3 Related Work: Fallacies in Argumentation Mining

Ad hominem fallacies in conversations have been addressed in [Habernal et al., 2018]. The authors used the subreddit ChangeMyView, which is a forum for civilized discussions, to create a dataset of fallacies. The dataset contains comments that the moderators removed as they violated the rule of not being rude or hostile, hence committing an ad hominem fallacy. The authors investigated what could be some triggers for committing the fallacy and which neural networks are best for ad hominem classification.

Fallacious arguments are often made in the dissemination of propaganda. In [Da San Martino et al., 2019], the authors annotate journal articles with 18 propaganda techniques, out of which the majority are also fallacies. They also experiment with several BERT-based architectures for the task of classifying arguments as fallacies.

In [Sahai et al., 2021], the authors align informal fallacies mentioned on Reddit within the pragma-dialectic theory of argumentation, design a methodology for labeling easily fallacies in online discussions and evaluate several neural models on the task of predicting fallacious arguments.

[2] https://plato.stanford.edu/entries/reasoning-defeasible/

References

[Da San Martino et al., 2019] Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., and Nakov, P. (2019). Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

[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. In *NAACL-HLT (1)*.

[Habernal et al., 2018] Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2018). Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. 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 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

[Hansen, 2020] Hansen, H. (2020). Fallacies. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition.

[Helwe et al., 2021] Helwe, C., Clavel, C., and Suchanek, F. M. (2021). Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction*.

[Lawrence and Reed, 2019] Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

[Macagno et al., 2017] Macagno, F., Walton, D., and Reed, C. (2017). Argumentation schemes. history, classifications, and computational applications. *History, Classifications, and Computational Applications (December 23, 2017). Macagno, F., Walton, D. & Reed, C*, pages 2493–2556.

[Sahai et al., 2021] Sahai, S., Balalau, O., and Horincar, R. (2021). Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 644–657, Online. Association for Computational Linguistics.

[Walton, 2005] Walton, D. (2005). Justification of argumentation schemes. *The Australasian Journal of Logic*, 3.

[Walton, 2010] Walton, D. (2010). Formalization of the ad hominem argumentation scheme. *Journal of Applied Logic*, 8(1):1–21.