Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Artificial Argumentation for Humans

Serena Villata
Université Côte d'Azur, CNRS, Inria, I3S, France
villata@i3s.unice.fr

Abstract

The latest years have seen an increasing interest in the topic of Artificial Intelligence (AI), the challenges it is facing, and the recent advances it has achieved, e.g., intelligent personal assistants. Differently from the past, where research on AI was mainly confined in research labs, the topic is now attracting interest from a wider audience, including policy-makers, information technology companies, and philosophers. Alas, these advances have also raised a number of concerns on AI's social, economic, and legal impact. Hence, the definition of design principles and automated methods to support transparent intelligent machine deliberation is highly desirable. Argumentation is important for handling conflicting beliefs, assumptions, opinions, goals, and many other mental attitudes. Argumentation pervades human intelligent behavior, and I believe that it is a mandatory element to conceive autonomous artificial machines that can exploit argumentation models and tools in the cognitive tasks they are required to carry out. Results in this area will allow reducing the gap between humans and machines towards a good AI hybrid society.

1 Introduction

Since the early years of the field, Artificial Intelligence has the goal to understand the principles governing intelligent behavior and to encode such principles in so-called *intelligent machines*. In the latest years, progress in AI seems to be accelerating, e.g., given the recent results in Machine Learning, Natural Language Processing (NLP) and Knowledge Engineering, leading to important investments in AI also from main information technology companies. Alas, all that glitters is not gold, and together with the increasing popularity of AI and the expectations on it, new concerns are also rising around the development of intelligent machines. We are at a crossroads: on the one hand, AI can be enormously beneficial for human flourishing, but on the other hand, we need to take care about the design of AI machines in order to reach a so-called *good AI hybrid society* [Russell, 2017; Cath *et al.*, 2018]. In this society, intelligent machines have the capability to form teams with humans.

Argumentation is the process by which arguments are constructed, compared, evaluated in some respect and judged in order to establish whether any of them are warranted. The idea of *argumentation* as the process of creating arguments for and against competing claims, was a subject of interest to philosophers and lawyers. In recent years, however, there has been a growth of interest in the subject from formal and technical perspectives in Computer Science, and a wide use of argumentation technologies in practical applications [Atkinson *et al.*, 2017]. The field of artificial argumentation plays an important role in Artificial Intelligence research. The reason for this is based on the recognition that if we are to develop robust intelligent machines able to act in mixed human-machine teams, then it is imperative that they can handle incomplete and inconsistent information in a way that somehow emulates the way humans tackle such a complex task (Figure 1).

During this first of part my career, I focused on different problems that I believe stand in the way of reaching this ambitious goal. I started from the observation that, in their deliberation process, humans use argumentation either internally, by evaluating arguments and counterarguments, or externally, by entering into a debate where arguments are exchanged. The three pillars of the development of argumentation-enhanced intelligent machines are, from my point of view: *(i)* modeling and reasoning on socio-cognitive components like trust using computational models of argument which are able to deal with incomplete and conflicting information, *(ii)* mining argument structures in natural language text to detect, e.g., potential fallacies, recurrent patters, and inner strength, and *(iii)* analyzing and understating the role of emotions in real world argumentative situations (e.g., debates) to inject such information in the computational models of argument to better cast incomplete and inconsistent information when emotions play a role.

On a broader scale, my view of argumentation-enhanced intelligent machines passes through the use of argumentation technologies to support the transparency of the deliberation process (*why* the machine deliberated in a certain way), and to support the extraction and reasoning on argumentation structures from different settings (e.g., clinical trials, social media posts, political debates) which, being generated by humans, require a high capability to deal with incompleteness and inconsistency. Humans argue. Machines should be able to argue too if we aim to achieve mixed teams in a hybrid society.

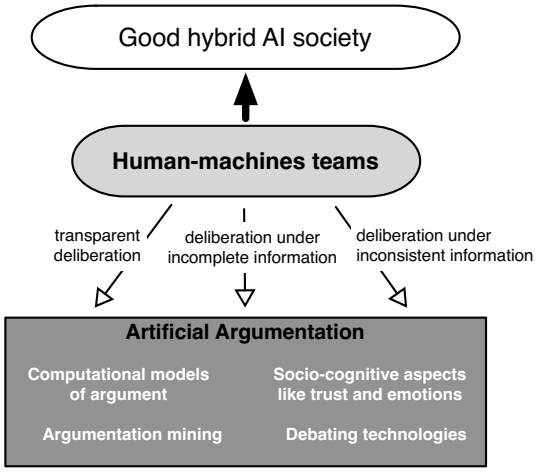

Figure 1: Towards argumentation-enhanced intelligent machines.

2 Reasoning on Trustful Arguments

Computational models of argument (COMMA) involve different ways for analyzing arguments and their relations. There are, at the higher level, two ways to formalize a set of arguments and their relationships, abstract argumentation and structured argumentation. Abstract argumentation has been introduced by Dung [Dung, 1995]. It sees each argument as an abstract entity and arguments are related to each other by means of attack relations. Structured argumentation [Besnard and Hunter, 2008; Prakken, 2010] is a framework in which more details about the arguments are considered. In particular, each argument is seen as composed by the premises, the claim and the inference rules used to achieve the claim from the premises.

Arguments are accepted following a set of formal criteria called *argumentation semantics* [Baroni *et al.*, 2011], which produce zero, one, or several sets of accepted arguments. Being interested since my PhD in providing computational models of argument that take into account socio-cognitive components in modeling the arguments and assessing their acceptability, I started from the observation that the acceptability of the arguments in human dialogues does not rely only on objective considerations but it is also highly influenced by the trustworthiness of the information source providing them. Hence, I focused on how to determine the acceptability of the arguments given the trustworthiness of the intelligent agents proposing them. In addition, the reasoning process addressed by the agents concerning the extent to which they trust the other information sources leads to the emergence not only of conflicts among the arguments but also of the conflicts among the sources. Since argumentation is a mechanism to reason about conflicting information, it seemed the suitable methodology for reason about trust.

In [da Costa Pereira *et al.*, 2011], we presented a *fuzzy labeling algorithm* to assign an acceptability *degree* to the arguments, depending on the trustworthiness of the agents who propose them, and the relations with the other arguments in the framework. We adopted possibility theory to represent uncertainty about the information, and to model the fact that information sources can be only partially trusted. We proved that the algorithm converges, and the convergence speed was estimated to be linear, as it is generally the case with iterative methods. In [da Costa Pereira *et al.*, 2016], we provided an experimental validation of this algorithm with the aim of carrying out an empirical evaluation of its performance on a benchmark of argumentation graphs. Results showed the satisfactory performance of our algorithm, even on complex graph structures as those present in the benchmark of the First International Competition on Computational Models of Argumentation (ICCMA) [Thimm and Villata, 2017]. While the information provided by a source should be assessed by an agent on the basis of several criteria: most notably, its content and the trust one has in its source, in turn, the observed quality of information should feed back on the assessment of its source, and such feedback should intelligently distribute among different features of the source, e.g., competence and sincerity. In [Paglieri *et al.*, 2014], we extended our fuzzy labeling algorithm into a formal framework in which trust is treated as a multi-dimensional concept relativized to the sincerity of the source and its competence with respect to specific domains. Both these aspects influence the assessment of the information, and also determine a feedback on the trustworthiness degree of its source.

Lately, the fuzzy labeling algorithm has been extended in [Cabrio *et al.*, 2017] to deal with bipolar abstract argumentation [Cayrol and Lagasquie-Schiex, 2013], an abstract argumentation framework where two kinds of relations hold between arguments, i.e., attack and support. In particular, the usefulness of this algorithm has been tested on a real use case, i.e., the explanation of the decisions of a question answering system. The bipolar fuzzy labeling algorithm is exploited to reconcile the information returned by the QAKiS question answering system[1], which queries over different and possible inconsistent data sources (i.e., the language-specific DBpedia chapters), and to explain the reasons underlying the proposed ranking.

In multi-agent systems, trust is also used to minimize the uncertainty in the interactions of the agents especially in case of conflicting information from different sources. Besides conflicts among information there can also be conflicts about the trust attributed to the information sources. In [Villata *et al.*, 2013], we explored how to express the possibly conflicting motivations about trust and distrust using argumentation. The methodology of meta-argumentation [Boella *et al.*, 2009] allowed us to model both information (i.e., arguments) and information sources (i.e., intelligent agents) as (meta-)arguments and to argue about them. We defined a focused representation of trust such that trust concerns not only the sources but also the arguments and the relations between them. When two pieces of information coming from different sources are conflicting, they can be seen as two arguments attacking each other. When an information source explicitly expresses a negative evaluation about the trustworthiness of another source, this negative evaluation is seen as an "attack" against the trustworthiness of the second source which is modelled as an argument as well.

[1] http://qakis.org/qakis2/

3 Mining Arguments in Natural Language

The modeling of arguments and their sources together in the same framework and the fact of considering also socio-cognitive aspects such as *trust* has been a fundamental step to move from computational models of argument closer to artificial argumentation for humans. However, a huge gap was still to overcome: humans do not express their arguments under the form of a logical formula nor through a node in a graph. Humans express arguments in natural language, i.e., through (possibly long) textual sentences. Overcoming this gap means to be able to move from arguments expressed in natural language to a formal representation of them (e.g., through *argumentation schemes* or structured argumentation), meaning at the final stage that the formal reasoning frameworks proposed thus far in the COMMA community may be exploited to reason over real arguments, which are often incomplete and conflicting. This is a very challenging goal as it involves several research areas from the AI panorama: NLP to provide the methods to process natural language text, to identify the arguments and their components (i.e., premises and claims) in texts and to predict the relations among such arguments; Knowledge Representation and Reasoning (KRR) to contribute with the reasoning capabilities upon the retrieved arguments and relations so that, for instance, fallacies and inconsistencies can be automatically identified in such texts, and Human-Computer Interaction (HCI) to guide the design of good human-computer digital argument-based supportive tools. The joint efforts of a set of researchers in these areas including myself allowed for the emergence of a new research field called *Argument(tation) Mining* (AM). AM has been defined as "the general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand" [Habernal and Gurevych, 2017].

The argument mining pipeline [Lippi and Torroni, 2016] is composed of three main steps: first, the argument components are identified in the text; second, the boundaries of such components are defined; third, the intra-argument relations (relations among the evidences and the claim composing an argument) and the inter-argument relations (relations among different arguments, e.g., support and attack) are predicted. Usually supervised learning methods (e.g., Support Vector Machines, Naïve Bayes classifiers, Logistic Regression, and Recurrent Neural Networks) are used to face these tasks, leading to the need of defining beforehand annotated datasets for the specific task and application scenario.

The first application scenario we considered were online debate platforms like Debatepedia (now iDebate). With the growing use of the Social Web, an increasing number of applications for exchanging opinions with other people are becoming available online. These applications are widely adopted with the consequence that the number of opinions about the debated issues increases. In order to cut in on a debate, the participants need first to evaluate the opinions of the other users to detect whether they are in favour or against the debated issue. In [Cabrio and Villata, 2013], we proposed and evaluated the use of natural language techniques to identify the arguments and their relations in online debate posts. In particular, we adopted the Textual Entailment (TE) approach, a generic framework for applied semantics, where linguistic objects are mapped by means of semantic inferences at a textual level. TE is then coupled together with an abstract bipolar argumentation system which allows to identify the arguments that are accepted in the considered online debate. In addition, we applied a data-driven approach to the analysis of the different proposals put forward for modelling the support relation in the COMMA field. Results confirmed that there is not a unique interpretation of the support relation, as the different combinations of additional attacks among the arguments involved in a support relation are significantly represented in the analyzed corpus. Using the same methodology, in [Cabrio *et al.*, 2013], we proposed an automatic framework to support the management of argumentative discussions in wiki-like platforms. More precisely, for our experiments, we used the history of Wikipedia for the top five more edited pages.

Also the problem of understanding the stream of messages exchanged on social media such as Facebook and Twitter is becoming a major challenge for automated systems. The tremendous amount of data exchanged on these platforms as well as the specific form of language adopted by social media users constitute a new challenging context for existing argument mining techniques. In particular, this is due to the peculiarities of the language used to write textual messages on social media. In [Bosc *et al.*, 2016] we constructed a resource of natural language arguments called DART (Dataset of Arguments and their Relations on Twitter) where the complete argument mining pipeline over Twitter messages is considered: *(i)* we identify which tweets can be considered as arguments and which cannot, and *(ii)* we identify what is the relation, i.e., support or attack, linking such tweets to each other. We also worked on the creation of a complete argument mining pipeline over Twitter messages, whose goal is to compute the set of tweets which are widely recognized as accepted, and the different (possibly conflicting) viewpoints that emerge on a topic, given a stream of messages. New issues emerge when dealing with arguments posted on such platforms, such as the need to make a distinction between personal opinions and actual facts, and to detect the source disseminating information about such facts to allow for provenance verification. In [Dusmanu *et al.*, 2017], we applied supervised classification to identify arguments on Twitter, and we presented two new tasks for argument mining, namely facts recognition and source identification. We studied the feasibility of the approaches proposed to address these tasks on a set of tweets related to the Grexit and Brexit news topics.

Another challenging scenario to apply argument mining is the political one. Politicians deliver speeches and participate into TV debates, therefore being able to identify fallacies and inconsistencies in their argumentation in an automated way would be a valuable contribution to the society. In [Menini *et al.*, 2018], we applied argument mining techniques, in particular relation prediction, to study political speeches in monological form, where there is no direct interaction between opponents. We argued that this kind of technique can effectively support researchers in history, social and political sciences, which must deal with an increasing amount of data in digital form and need ways to automatically extract and analyze ar-

gumentation patterns. We tested and discussed our approach based on the analysis of documents issued by R. Nixon and J. F. Kennedy during 1960 presidential campaign. We relied on a supervised classifier to predict argument relations (i.e., support and attack). The application of argument mining to such data allows not only to highlight the main points of agreement and disagreement between the candidates over the campaign issues such as Cuba, but also an in-depth argumentative analysis of the respective viewpoints on these topics.

4 Correlating Emotions with Arguments

Human argumentation involves also a final component that cannot be overlooked, i.e., the *emotional* one. The assessment of the arguments, their persuasive power truly passes through the objective evaluation of their coherence and the trustworthiness of their sources, but it also involves emotional aspects that play an important role. In particular, argumentation is seen in the COMMA field as a mechanism to support different forms of reasoning such as decision making and persuasion and it always casts under the light of critical thinking. In the latest years, several computational approaches to argumentation have been proposed to detect conflicting information, take the best decision with respect to the available knowledge, and update our own beliefs when new information arrives. The common point of all these approaches is that they assume a purely rational behavior of the involved actors, be them humans or intelligent agents. However, this is not the case as humans are proved to behave differently, mixing rational and emotional attitudes to guide their actions. Some works have claimed that there exists a strong connection between the argumentation process and the emotions felt by people involved in such process. In [Benlamine *et al.*, 2015], we advocated a complementary, descriptive and experimental method, based on the collection of emotional data about the way human reasoners handle emotions during debate interactions. Across different debates, people's argumentation in plain English is correlated with the emotions automatically detected from the participants through emotion recognition tools, their engagement in the debate, and the mental workload required to debate captured through Electroencephalography (EEG) headsets. Results showed several correlations among emotions, engagement and mental workload with respect to the argumentation elements. For instance, high engagement is correlated with negative emotions showing that participants are mentally involved in producing arguments to rebut those which are not in line with their viewpoint, and neuroticism and conscientiousness have both a negative impact on the debaters' brain indexes ending up into a reduced mental engagement index and an increased cognitive load. Beside their theoretical value for validating and inspiring computational models of argument, these results have applied value for developing artificial agents meant to argue with human users or to assist users in the management of debates.

As a further natural step, we decided to move to a specific kind of argumentation, namely *argumentative persuasion*. It implies two parties where the former tries to get the latter to do (or not do) some action or to believe (or not believe) something. It usually employs one of the three persuasion strategies, i.e., Ethos (relying on the authority of the persuader), Pathos (soliciting the emotions) or Logos (grounded on logical arguments), depending on the topic of the debate and the persuader. Several approaches have been proposed to model argumentative agents following one or more of these strategies to persuade the other agents and change their beliefs. However, none of them explored how the choice of a strategy impacts the mental states of the debaters and the argumentation process. In [Villata *et al.*, 2018], we address this issue by setting a field experiment with real debaters to assess the impact of persuasion strategies on the mental engagement and emotions of the participants, as well as on the persuasiveness power of the arguments exchanged during the debate. Also in this case, participants were equipped with EEG headsets and emotion recognition tools. Our results showed that the Pathos strategy is the most effective in terms of mental engagement.

5 Future Directions

The main focus of my career has been so far on making artificial argumentation closer to human argumentation. The general aim of my research is to define argumentation-enhanced intelligent machines which need to deal with incomplete, conflicting and uncertain information. More precisely, my contributions fall within the general areas of computational models of argument (focusing on the definition of formal models of argument supporting deliberation and explanation by taking into account external components like trust, emotions and norms), and argument mining (focusing on the definition of empirical methods for detecting argumentative structures from text, considering application scenarios like social media posts, medical trials, and political debates).

The problems described here cannot at all be considered as closed or completely solved. It is important to note that this line of research has a high degree of multi-disciplinarity, meaning that a lot of interactions are required with experts in linguistics, cognitive science, and psychology. These interactions are not always successful, and a shared vocabulary needs to be built to properly interact. Yet, when fruitful collaborations are established, then the results are far better than expected. Moreover, it should also be considered that encoding socio-cognitive components into a formal argumentation framework often leads to an incomplete representation of these fuzzy, multi-faceted elements. The risk is always to come up with a framework modeling part of the component in a fine-grained way, but that is hardly extendable to include another aspect. From the point of view of argument mining, it appears evident that the argumentative sentences "in the wild", i.e., in natural language text as the ones reported in the examples, are pretty far from the prototypical argumentation patterns usually investigated in COMMA, increasing the complexity of the task. Despite the good results obtained in some application scenarios, for other kinds of documents (e.g., legal cases) system performances should improve. It is important to note that also human agreement (generally viewed as the upper bound on automatic performance in annotation tasks) is affected by the complexity of the AM tasks.

Acknowledgments

This work has been partially supported by EU H2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 for the project MIREL: MIning and REasoning with Legal texts.

References

[Atkinson *et al.*, 2017] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo R. Simari, Matthias Thimm, and Serena Villata. Towards Artificial Argumentation. *AI magazine*, 38(3), 2017.

[Baroni *et al.*, 2011] P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410, 2011.

[Benlamine *et al.*, 2015] Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. Emotions in argumentation: an empirical evaluation. In *Proc. of the 24th Int. Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 156–163, 2015.

[Besnard and Hunter, 2008] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008.

[Boella *et al.*, 2009] Guido Boella, Dov M. Gabbay, Leendert W. N. van der Torre, and Serena Villata. Meta-argumentation modelling I: methodology and techniques. *Studia Logica*, 93(2-3):297–355, 2009.

[Bosc *et al.*, 2016] Tom Bosc, Elena Cabrio, and Serena Villata. Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. In *Proc. of the 6th Int. Conference on Computational Models of Argument, COMMA 2016*, volume 287 of *Frontiers in AI and Applications*, pages 21–32. IOS Press, 2016.

[Cabrio and Villata, 2013] Elena Cabrio and Serena Villata. A Natural Language Bipolar Argumentation Approach to Support Users in Online Debate Interactions. *Argument and Computation*, 4(3):209–230, 2013.

[Cabrio *et al.*, 2013] Elena Cabrio, Serena Villata, and Fabien Gandon. A Support Framework for Argumentative Discussions Management in the Web. In *ESWC - 10th Int. Conference on The Semantic Web*, volume 7882 of *LNCS*, pages 412–426. Springer, 2013.

[Cabrio *et al.*, 2017] Elena Cabrio, Serena Villata, and Alessio Palmero Aprosio. A RADAR for information reconciliation in question answering systems over linked data. *Semantic Web*, 8(4):601–617, 2017.

[Cath *et al.*, 2018] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics*, 24(2):505–528, 2018.

[Cayrol and Lagasquie-Schiex, 2013] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. *Int. J. Approx. Reasoning*, 54(7):876–899, 2013.

[da Costa Pereira *et al.*, 2011] Célia da Costa Pereira, Andrea Tettamanzi, and Serena Villata. Changing ones mind: Erase or rewind? In *Proc. of the 22nd Int. Joint Conf. on Artificial Intelligence, IJCAI*, pages 164–171, 2011.

[da Costa Pereira *et al.*, 2016] Célia da Costa Pereira, Mauro Dragoni, Andrea G. B. Tettamanzi, and Serena Villata. Fuzzy labeling for abstract argumentation: An empirical evaluation. In *Proc. of the 10th Int. Conference on Scalable Uncertainty Management, SUM 2016*, volume 9858 of *LNCS*, pages 126–139. Springer, 2016.

[Dung, 1995] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.

[Dusmanu *et al.*, 2017] Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument Mining on Twitter: Arguments, Facts and Sources. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 2317–2322. ACL, 2017.

[Habernal and Gurevych, 2017] I. Habernal and I. Gurevych. Argumentation mining in user-generated web discourse. *Comput. Linguist.*, 43(1):125–179, 2017.

[Lippi and Torroni, 2016] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10, 2016.

[Menini *et al.*, 2018] Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. Never retreat, never retract: Argumentation analysis for political speeches. In *Proc. of the 32nd AAAI Conference on Artificial Intelligence*, pages 4889–4896, 2018.

[Paglieri *et al.*, 2014] Fabio Paglieri, Cristiano Castelfranchi, Célia da Costa Pereira, Rino Falcone, Andrea Tettamanzi, and Serena Villata. Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. *Comp. and Math. Organization Theory*, 20(2):176–194, 2014.

[Prakken, 2010] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.

[Russell, 2017] Stuart J. Russell. Provably beneficial artificial intelligence. *Exponential Life, The Next Step*, 2017.

[Thimm and Villata, 2017] Matthias Thimm and Serena Villata. The first international competition on computational models of argumentation: Results and analysis. *Artif. Intell.*, 252:267–294, 2017.

[Villata *et al.*, 2013] Serena Villata, Guido Boella, Dov M. Gabbay, and Leendert Van Der Torre. A socio-cognitive model of trust using argumentation theory. *Int. Journal of Approximate Reasoning*, 54(4):541–559, 2013.

[Villata *et al.*, 2018] Serena Villata, Sahbi Benlamine, Elena Cabrio, Claude Frasson, and Fabien Gandon. Assessing persuasion in argumentation through emotions and mental states. In *Proc. of the 31st Int. Florida Artificial Intelligence Research Society Conference, FLAIRS*, pages 134–139. AAAI, 2018.