Computational Hermeneutics: Using Automated Reasoning for the Logical Analysis of Natural-Language Arguments

David Fuenmayor
(PhD Advisor: Prof. Christoph Benzmüller)

Dep. of Mathematics and Computer Science, Freie Universität Berlin, Germany
david.fuenmayor@fu-berlin.de

Abstract. While there have been major advances in automated theorem proving (ATP) during the last years, its main field of application has mostly remained bound to mathematics and hardware/software verification. We argue that the use of ATP in philosophy can also be very fruitful, not only because of the obvious quantitative advantages of automated reasoning tools (e.g. reducing by several orders of magnitude the time needed to test some argument's validity), but also because it enables a novel approach to the logical analysis of arguments. This approach, which we have called *computational hermeneutics*, draws its inspiration from work in the philosophy of language such as Donald Davidson's theory of *radical interpretation* and contemporary *inferentialist* theories of meaning, which do justice to the inherent circularity of linguistic understanding: the whole is understood (compositionally) on the basis of its parts, while each part is understood only in the (inferential) context of the whole. *Computational hermeneutics* is thus a holistic, iterative, trial-and-error enterprise, where we evaluate the adequacy of some candidate formalization of a sentence by computing the logical validity of the whole argument. We start with formalizations of some simple statements (taking them as tentative) and use them as stepping stones on the way to the formalization of other argument's sentences, repeating the procedure until arriving at a state of *reflective equilibrium*: A state where our beliefs have the highest degree of coherence and acceptability.

1 Background, Problem and Objectives

The traditional conception of logic as an *ars iudicandi* sees as its central role the classification of arguments into valid and invalid ones by identifying criteria that enable us to *judge* the correctness of (mostly deductive) inferences. However, logic can also be conceived as an *ars explicandi* aiming at rendering the rules implicit in our socio-linguistic argumentative praxis in a more orderly, more transparent, and less ambiguous way, thus making explicit the inferential structure of natural-language discourse.

During this research project, we want to explore the computer-supported application of formal logic (using automated theorem provers) to issues in philosophy concerned with: (i) the methodical evaluation (logic as *ars iudicandi*) and interpretation (logic as *ars explanandi*) of arguments and, building upon this, we want to tackle (ii) the problem of formalization: how to search *methodically* for the most appropriate logical form(s) of a given natural-language argument, by casting its individual statements into expressions of some sufficiently expressive logic (classical or non-classical).

There have been some few proposals regarding the ambitious project of defining an effective formalization procedure for natural language; being among the most well-known: Davidson's theory of meaning [13], Chomsky's generative grammar [11] and particularly Montague's universal grammar program [25] (and its intellectual heirs like Discourse Representation Theory [24] and Dynamic Predicate Logic [23] among others). Other proposals have, by contrast, focused on the task of providing definite adequacy criteria of formalization (e.g. [7,15,30,10,1,28]). However, there seems to be still no paradigmatic method for systematically arriving to an argument's formalization, nor definite criteria for rigorously judging its adequacy. The logical analysis of natural language continues to be essentially an artistic skill that cannot be standardized or taught methodically (aside from providing students with a handful of paradigmatic examples supplemented with commentaries).

This research aims at improving this situation by expanding on and implementing some of the most recent proposals found in the literature, with a special emphasis on the recent work of Peregrin and Svodoba [29] who, apart from providing *syntactic* and *pragmatic* (inferential) adequacy criteria, also tackle the difficult problem of finding a formalization procedure by proposing the method of *reflective equilibrium*, which is similar in spirit to the idealized scientific (i.e. *hypothetico-deductive*) method and, additionally, has the virtue of approaching this problem in a holistic way: the adequacy of the formalizations of all statements of a language (in our case, an argument) is assessed as a whole (in the spirit of Davidson's theory of meaning[1] [13,12] and Quine's (resp. Hempel's) *confirmation holism* [32]).

Following a holistic *hypothetico-deductive* approach for logical analysis was, until very recently, not feasible in practice; since it involves an iterative process of trial-and-error, where the adequacy of some candidate formalization for a sentence becomes tested by computing the logical validity of the *whole* argument. In

[1] As Davidson himself has put it: "[...] much of the interest in logical form comes from an interest in logical geography: to give the logical form of a sentence is to give its logical location in the totality of sentences, to describe it in a way that explicitly determines what sentences it entails and what sentences it is entailed by. The location must be given relative to a specific deductive theory; so logical form itself is relative to a theory." ([12] p. 140)

order to explore the vast combinatoric of candidate formalizations for even the simplest argument, we have to test its validity at least several hundreds of times (also to account for *logical pluralism*). It is here where the recent improvements and ongoing consolidation of modern automated theorem proving technology (for propositional logic, first-order logic and in particular also higher-order logic) become handy.

One of the aims of this research is to combine and apply theoretical insights from both inferential (proof-theoretical) and semantical (model-theoretical) approaches to the analysis and assessment of concrete arguments, particularly in ethics and politics.[2] An important deliverable of this research will consist in the development of an experimental software providing functionalities for automated logical analysis and evaluation of natural-language arguments.

2 Current and Previous Work

2.1 Computational Hermeneutics

In current work [20,18,21], we have introduced an approach named *computational hermeneutics* aimed at improving the tasks of logical analysis and interpretation of arguments in a semi-automated fashion by exploiting the computing power and usability of modern proof assistants (Isabelle/HOL [26] in particular). This method has been developed as a result of reflecting upon previous work on the application of Automated Theorem Proving (ATP) for the formalization and assessment of arguments in metaphysics (e.g. [4,5,17]) and is especially suited to the utilization of different kinds of classical and non-classical logics through Christoph Benzmüller's technique of *semantic embeddings* [3,2].

The ideas behind *computational hermeneutics* have been inspired by Donald Davidson's theory of *radical interpretation* [13,14], by drawing on his well-known *principle of charity* and a *holistic* account of meaning. We defend the view that the process of concept explication can take place in the very practice of argumentation through the explicitation of the inferential role that terms –both logical and non-logical– play in some theory or argument of our interest (see logical and semantic inferentialism [9,8,27]). In the case of formal arguments (in any arbitrary logic) this task of concept explication is carried out, for instance, by providing definitions (i.e. by directly correlating to a *definiens*) or by axiomatizing interrelations among terms. Concrete examples of the application of this approach to the logical analysis and assessment of ontological arguments can be found in [20,18,17] (using the *Isabelle* proof assistant).

There are ongoing efforts to formulate the *computational hermeneutics* approach as a combinatorial optimization problem in order to take advantage of existing

[2] In particular, Prof. Christoph Benzmüller's current research project: *Consistent Rational Argumentation in Politics (CRAP)* at the Free University Berlin.

algorithmic solutions (e.g. simulated annealing, evolutionary algorithms, etc.) and to integrate them with current state-of-the-art automated theorem provers and model finders in order to evaluate fitness and selection criteria. At this early stage of the project we can provide, as an illustration, a rough outline of the iterative structure of the *computational hermeneutics* approach:

1. Argument reconstruction (initially in natural language):

- **Add or remove sentences and choose their truth-values.** Premises and desired conclusions would need to become true, while some other 'unwanted' conclusions would have to become false. Deciding on these issues would expectedly involve a fair amount of human judgment.
- **Establish inferential relations,** i.e., determine the extension of the logical consequence relation: which sentences are to follow (logically) from which others. This task can be done manually or automatically by letting our automated tools find this out for themselves, provided the argument has already been formalized (even if only roughly after some few iterations). Automating this task frequently leads to the simplification of the argument, since current theorem provers are quite good at detecting idle premises/axioms (see, e.g., Isabelle's *Sledgehammer* tool [6]).

2. Choosing a logic for formalization, by determining the logical structure of the natural-language sentences occurring in the argument. The (partial) automation of this task can be supported by searching in a catalog of *semantic embeddings* of different logics in HOL (see [3,2] and next section) in order to select a candidate logic (modal, free, deontic, etc.) satisfying some given syntactic or semantic criteria (drawing, for instance, on the output of linguistic parsers).

3. Argument formalization (in the chosen logic), while getting continuous feedback from our automated reasoning tools about the argument's correctness (validity, consistency, non-circularity, etc.). This stage is itself iterative, since, for every sentence, we charitably (i.e. in the spirit of the *principle of charity*) try several different formalizations until getting a correct argument. Here is where we take most advantage of the real-time feedback offered by our automated tools. Some main tasks to be considered are:

- **Translate natural-language sentences into the target logic,** by relying either on our pre-understanding or on provided definitions of the argument's concepts. This step can be partially automated by using semantic parsers, but would also need some human-support.
- **Vary the logical form of already formalized sentences.** This can be done systematically and automatically by relying upon a catalog of (consistent) logical variations of formulas (see *semantic embeddings* in next section) and the output of automated tools (ATPs, model finders, etc).

- **Bring related terms together,** either by introducing definitions or by axiomatizing new interrelations among them. These newly introduced formulas can be translated back into natural language to be integrated into the argument in step 1.1, thus being disclosed as former *implicit* premises. The process of searching for additional premises with the aim of rendering an argument as formally correct can be seen as a kind of abductive reasoning ('inference to the best explanation'), which can be partially automated using current AI technology but still needs human support.

4. Are termination criteria satisfied? That is, have we achieved the *reflective equilibrium*? If not, we would come back to some early stage. Both termination and selection criteria are to be based on the adequacy criteria of formalization found in the literature.[3] Furthermore, the introduction of automated reasoning tools makes it feasible to apply these criteria while working with a database of correct and incorrect/fallacious arguments of a considerable size.

2.2 Semantic Embeddings of Non-Classical Logics in HOL

A parallel stream of work consists on improving and broadening Benzmüller's technique of semantic embeddings (see [3,2]), which allows us to take advantage of the expressive power of classical higher-order logic (as a metalanguage) in order to embed the syntax and semantics of another logic (object language). Using this technique we can, for instance, embed a modal logic by defining the modal *box* and *diamond* operators as meta-logical predicates and using quantification over sets of objects of a specially introduced type: 'possible world' (according to Kripke semantics). This approach allows us to reuse existing ATP technology for classical HOL and apply it for automated reasoning in non-classical logics (e.g. free, modal, temporal and deontic logics).

In previous work with Christoph Benzmüller [17,19], we presented a shallow semantic embedding for an intensional higher-order modal logic (IHOML) using the Isabelle/HOL proof assistant [26]. This logic has been originally introduced

[3] Consider, e.g., Peregrin and Svodoba's ([28,29]) proposed inferential criteria for evaluating the adequacy of formalization:
(i) The *principle of reliability*: "ϕ counts as an adequate formalization of the sentence S in the logical system L only if the following holds: If an argument form in which ϕ occurs as a premise or as the conclusion is valid in L, then all its perspicuous natural language instances in which S appears as a natural language instance of ϕ are intuitively correct arguments."
(ii) The *principle of ambitiousness*: "ϕ is the more adequate formalization of the sentence S in the logical system L the more natural language arguments in which S occurs as a premise or as the conclusion, which fall into the intended scope of L and which are intuitively perspicuous and correct, are instances of valid argument forms of S in which ϕ appears as the formalization of S." ([29] pp. 70-71).

by Melvin Fitting in his book *Types, Tableaus and Gödel's God* [16] and can be seen as an improved variant of the intensional logic originally developed by Montague [25] and later expanded by Gallin [22] by building upon Church's type theory and Kripke's possible-world semantics. In contrast to the earlier approaches, intensions and extensions are, in IHOML, both first-class objects. In this work we also presented an exemplary, non-trivial application of this reasoning infrastructure in the emergent field of *computational metaphysics*: the computer-supported formalization and critical assessment of Gödel's modern variant of the *ontological argument* and two of its proposed emendations (see [16,31] for further details).

Our approach to the semantic embedding of IHOML has built on previous work on the embedding of multimodal logics with quantification [3], which we have expanded to allow for restricted (aka. actualist) quantification, intensional terms, and their related operations. From an AI perspective we contributed a highly flexible framework for automated reasoning in intensional and modal logic. IHOML, which has not been automated before, has several applications, particularly, regarding the deep semantic analysis of natural language rational arguments, as envisioned in the *CRAP* research project mentioned above (see footnote 2).

3 Ongoing and Future Work

As illustrated above (and in greater detail in [20,18,17]) *computational hermeneutics* can be carried out in a semi-automatic fashion: We work iteratively on an argument by (i) fixing truth-values and inferential relations among its sentences; (ii) (tentatively) choosing a logic for formalization; and (iii) working back and forth on the formalization of its axioms and theorems (eventually adding new ones), making gradual adjustments while getting real-time feedback about the suitability of our changes (e.g. validating the argument, avoiding inconsistency or question-begging, etc.). This steps are to be repeated until arriving at a state of *reflective equilibrium*: A state where our beliefs have the highest degree of coherence and acceptability according to syntactic and, particularly, inferential adequacy criteria (such as the ones mentioned above).

Computational hermeneutics can thus be seen as an instance of the hypothetico-deductive (aka. scientific) method, since it features the sort of holistic mutual adjustment between theory and observation, which characterizes the idealized scientific inquiry. While modern ATP technology gives us the means to deductively draw inferences from our hypotheses (and to falsify them), the most challenging task remains how to *systematically* come up with the candidate hypotheses.[4]

[4] Together with Prof. Benzmüller's research team, we are currently exploring the potential of combining ATP with machine learning techniques, as applied particularly in *natural language processing* (NLP).

We are currently exploring the idea of approaching the problem of formalization as a combinatorial optimization problem, by using (among others) inferential criteria of adequacy to define the fitness/utility function of an appropriate optimization algorithm. The *principle of charity* would inspire our main selection criteria: an adequate formalization must validate the argument and guarantee some minimal qualitative features (consistency and independence of premises, invalidation of implausible conclusions, no 'question-begging', etc.). It is worth noting that, for the kind of non-trivial philosophical arguments we are interested in (e.g. ethics and metaphysics), such a selection criteria would aggressively prune our search tree. Furthermore, the evaluation of our fitness function is already, with today's technologies, not only completely automizable, but also seems to be highly parallelizable.

Significant resources will be devoted to the development of working software. The main touchstone for the validity of the gained insights will be the implementation of a software system, whose main functionality will be automated logical analysis: accepting an argument in natural-language as input and generating as output its most appropriate formalization (under consideration of different logics in view of some well-defined criteria). This software will interface and cooperate with other existing systems and technologies such as software for linguistic analysis and text mining/analytics, automated theorem provers (e.g. Leo-III, Satallax, Vampire) and interactive proof assistants (e.g. Isabelle, Coq). Further applications in areas like knowledge/ontology extraction, semantic web and legal informatics are currently being contemplated.

References

1. M. Baumgartner and T. Lampert. Adequate formalization. *Synthese*, 164(1):93–115, 2008.
2. C. Benzmüller. Recent successes with a meta-logical approach to universal logical reasoning (extended abstract). In S. A. da Costa Cavalheiro and J. L. Fiadeiro, editors, *Formal Methods: Foundations and Applications - 20th Brazilian Symposium, SBMF 2017, Recife, Brazil, November 29 - December 1, 2017, Proceedings*, volume 10623 of *Lecture Notes in Computer Science*, pages 7–11. Springer, 2017.
3. C. Benzmüller and L. Paulson. Quantified multimodal logics in simple type theory. *Logica Universalis (Special Issue on Multimodal Logics)*, 7(1):7–20, 2013.
4. C. Benzmüller, L. Weber, and B. Woltzenlogel Paleo. Computer-assisted analysis of the Anderson-Hájek controversy. *Logica Universalis*, 11(1):139–151, 2017.
5. C. Benzmüller and B. Woltzenlogel Paleo. The inconsistency in Gödel's ontological argument: A success story for AI in metaphysics. In *IJCAI 2016*, 2016.
6. J. Blanchette, S. Böhme, and L. Paulson. Extending Sledgehammer with SMT solvers. *Journal of Automated Reasoning*, 51(1):109–128, 2013.
7. U. Blau. *Die dreiwertige Logik der Sprache: ihre Syntax, Semantik und Anwendung in der Sprachanalyse.* Walter de Gruyter, 1978.
8. R. Brandom. *Articulating reasons.* Harvard University Press, 2009.
9. R. B. Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Harvard University Press, 1994.

10. G. Brun. *Die richtige Formel: Philosophische Probleme der logischen Formalisierung*, volume 2. Walter de Gruyter, 2003.
11. N. Chomsky. *The minimalist program.* MIT press, 2014.
12. D. Davidson. *Essays on actions and events: Philosophical essays*, volume 1. Oxford University Press on Demand, 2001.
13. D. Davidson. *Inquiries into Truth and Interpretation: Philosophical Essays*, volume 2. Oxford University Press, 2001.
14. D. Davidson. Radical interpretation. In *Inquiries into Truth and Interpretation.* Oxford University Press, 2001.
15. R. L. Epstein. The semantic foundations of logic. In *The Semantic Foundations of Logic Volume 2: Predicate Logic.* Oxford University Press, 1994.
16. M. Fitting. *Types, tableaus, and Gödel's god*, volume 12. Springer Science & Business Media, 2002.
17. D. Fuenmayor and C. Benzmüller. Automating emendations of the ontological argument in intensional higher-order modal logic. In *KI 2017: Advances in Artificial Intelligence 40th Annual German Conference on AI, Dortmund, Germany, September 25-29, 2017, Proceedings*, volume 10505 of *LNAI*, pages 114–127. Springer, 2017.
18. D. Fuenmayor and C. Benzmüller. Computer-assisted reconstruction and assessment of E. J. Lowe's modal ontological argument. *Archive of Formal Proofs*, Sept. 2017. http://isa-afp.org/entries/Lowe_Ontological_Argument.html.
19. D. Fuenmayor and C. Benzmüller. Types, Tableaus and Gödel's God in Isabelle/HOL. *Archive of Formal Proofs*, 2017. https://www.isa-afp.org/entries/Types_Tableaus_and_Goedels_God.html.
20. D. Fuenmayor and C. Benzmüller. A case study on computational hermeneutics: E. J. Lowe's modal ontological argument. *Journal of Applied Logic (special issue on Formal Approaches to the Ontological Argument)*, 2018. To appear.
21. D. Fuenmayor and C. Benzmüller. Computational hermeneutics: Using computers to interpret philosophical arguments (abstract). In *Logical Correctness, Workshop at UNILOG'2018, UNILOG'2018 Book of Abstracts*, 2018.
22. D. Gallin. *Intensional and Higher-Order Modal Logic.* N.-Holland, 1975.
23. J. Groenendijk and M. Stokhof. Dynamic predicate logic. *Linguistics and philosophy*, 14(1):39–100, 1991.
24. H. Kamp, J. Van Genabith, and U. Reyle. Discourse representation theory. In *Handbook of philosophical logic*, pages 125–394. Springer, 2011.
25. R. Montague. *Formal Philosophy: Selected Papers of Richard Montague. Ed. and with an Introd. by Richmond H. Thomason.* Yale University Press, 1974.
26. T. Nipkow, L. C. Paulson, and M. Wenzel. *Isabelle/HOL — A Proof Assistant for Higher-Order Logic.* Number 2283 in LNCS. Springer, 2002.
27. J. Peregrin. *Inferentialism: why rules matter.* Springer, 2014.
28. J. Peregrin and V. Svoboda. Criteria for logical formalization. *Synthese*, 190(14):2897–2924, 2013.
29. J. Peregrin and V. Svoboda. *Reflective Equilibrium and the Principles of Logical Analysis: Understanding the Laws of Logic.* Routledge Studies in Contemporary Philosophy. Taylor and Francis, 2017.
30. R. Sainsbury. *Logical Forms: An Introduction to Philosophical Logic.* Blackwell Publishers, 1991.
31. J. Sobel. *Logic and Theism: Arguments for and Against Beliefs in God.* Cambridge U. Press, 2004.
32. W. van Orman Quine. Two dogmas of empiricism. In *Can Theories be Refuted?*, pages 41–64. Springer, 1976.