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

David Fuenmayor[1] and Christoph Benzmüller[2,1]

[1] Freie Universität Berlin, Germany
[2] University of Luxembourg, Luxembourg

Abstract. We present a method for finding a most adequate logical formalization of a natural language argument. Our approach, which we call *computational hermeneutics*, is grounded on recent progress in the area of automated theorem proving. It is inspired by Donald Davidson's work on *radical interpretation*; a systematic approach to interpretation that does justice to the inherent circularity of understanding: the whole is understood (compositionally) on the basis of its parts, while each part is understood only in the context of the whole (*hermeneutic circle*). *Computational hermeneutics* is 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 The Motivation

While there have been major advances in automated theorem proving (ATP) during the last years, its main field of application has mostly remained bounded to mathematics and hardware/software verification. We argue that the use of ATP in philosophy can also be very fruitful (see for example the results reported in [5]), 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 argument's validity), but also because it enables a novel approach to the logical analysis of arguments, which we call *computational hermeneutics*.

As a result of reflecting upon previous work on the application of ATP for the computer-supported evaluation of arguments in metaphysics [5,6,12,1,4], we have become interested in developing a systematic formalization procedure for natural-language arguments with regard to their assessment using automated tools. Unsurprisingly, the problem of finding the adequate formalization of natural-language discourse is far from trivial and has been tackled in the past

without much practical success (see e.g. the research derived from Davidson's theory of meaning and interpretation [10], Chomsky's generative grammar [7] and particularly Montague's universal grammar [14]). The logical analysis of natural language continues to be considered as a kind of artistic skill that cannot be standardized or taught methodically (aside from providing students with a handful of paradigmatic examples supplemented with commentaries). Our research aims at improving this situation by expanding on (and implementing) Davidson's work on *radical interpretation* [10] and the recent work of Peregrin and Svodoba [15,16], who tackle the difficult problem of formalization in a holistic way, by introducing the iterative method of *reflective equilibrium*. As an illustration, let us examine one of their proposed criteria for evaluating the adequacy of formalization, the *principle of reliability* ([16] p. 70):

> "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."

According to this criteria, the adequacy of each individual formalization of an argument's sentence is assessed by computing the argument's validity *as a whole* (which depends itself on the way we formalize its constituent sentences). As we see it, this circle is a virtuous one: it does justice to holistic accounts of meaning relying on the inferential role of sentences. As Davidson [9] 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." (p. 140)

2 Radical Interpretation and the Principle of Charity

What is the use of *radical interpretation* in argumentation? The answer is trivially stated by Davidson himself, by arguing that "all understanding of the speech of another involves radical interpretation" ([8], p. 125). Furthermore, the impoverished evidential position we are faced with when interpreting some arguments (particularly philosophical ones) corresponds very closely to the starting situation Davidson contemplates in his thought experiments on *radical interpretation*, where he shows how an interpreter could come to understand someone's words and actions without relying on any prior understanding of them. Davidson's program builds on the idea of taking the concept of truth as basic and extracting from it an account of interpretation satisfying two general requirements: (i) it must reveal the compositional structure of language, and (ii) it can be assessed by evidence available to the interpreter [8,10].

The first requirement (i) is addressed by noting that a theory of truth in Tarski's style (modified to apply to natural language) can be used as a theory of interpretation. This implies that, for every sentence s of an object language L, a sentence of the form: ' "s" is true in L iff p' (aka. *T-schema*) can be derived, where p is a translation of s into the metalanguage used for interpretation (note that the sentence p is being *used*, while s is only being *mentioned*). Thus, by virtue of the recursive nature of Tarski's theory of truth [18], the structure of the object language becomes revealed. From the point of view of *computational hermeneutics*, the object language L corresponds to the idiolect of the speaker (natural language), and the metalanguage is constituted by the formulas of our logic of formalization (classical HOL as well as other non-classical logics through the technique of *semantic embeddings* [2,3]) plus the expression "is valid in the logic XY".

As an illustration, we present an instance of the *T-schema* taken from our preliminary case studies in *computational hermeneutics* (e.g. [13,12]): ' "There is only one God" is true iff "$\exists x.\ God\ x \wedge \forall y.\ God\ y \rightarrow y = x$" is valid in HOL'. Notice that the *used* metalanguage sentence p has the form: ' "q" is valid in HOL', where the *mentioned* sentence q corresponds to the formalization of the object sentence s. Taking q as an interpretation of s certainly helps us to clarify our understanding of s and to shed light –by virtue of compositionality– on the meanings (and/or inferential roles) of its individual words.

The second general requirement (ii) states that the interpreter can have access to *objective* evidence in order to judge the appropriateness of her interpretations (i.e. access to the events and objects in the 'external world' which make sentences true). This particularly implies access to the speaker's attitudes regarding the truth or falsity of sample sentences, under specified circumstances observable by speaker and interpreter alike. In *computational hermeneutics*, this kind of objective evidence is provided by the output of automated reasoning tools. Thus, the computer acts as an (arguably unbiased) arbiter deciding on the truth of a sentence in the context of an argument.

A central concept in Davidson's account of *radical interpretation* is the *principle of charity*, which he holds as a condition for the possibility of engaging in any kind of interpretive endeavor. The *principle of charity* has been summarized by Davidson by stating that "we make maximum sense of the words and thoughts of others when we interpret in a way that optimizes agreement" ([10] p. 197). Hence the principle builds on the possibility of intersubjective agreement about external facts among speaker and interpreter. The *principle of charity* can be invoked to make sense of a speaker's ambiguous utterances and, in our case, to presume (and foster) the validity of an argument. Consequently, in *computational hermeneutics* we assume from the outset that an argument's conclusions indeed follow from its premises and disregard formalizations that do not do justice to this postulate.

3 Holistic Approach: Why Feasible Now?

Following a holistic 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 order to explore the vast combinatoric of possible 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.

To get an idea of this, let us imagine the following scenario: A philosopher working on a formal argument wants to test a variation on one of its premises or definitions and find out if the argument still holds. Since our philosopher is working with pen and paper, she will have to follow some kind of proof procedure (e.g. tableaus or natural-deduction calculus), which, depending on her calculation skills, may take some minutes to be carried out. It seems clear that she cannot allow herself many of such experiments on such conditions.

Now compare the above scenario to another one in which our working philosopher can carry out such an experiment in just a few seconds and with no effort, by employing an automated theorem prover. In a best-case scenario, the proof assistant would automatically generate a proof (or the sketch of a countermodel), so she just needs to interpret the results and use them to inform her new conjectures. In any case, she would at least know if her speculations had the intended consequences, or not. After some minutes of work, she will have tried plenty of different variations of the argument while getting real-time feedback regarding their suitability.[3]

We aim at showing how this radical *quantitative* increase in productivity does indeed entail a *qualitative* change in the way we approach formal argumentation, since it allows us to take things to a whole new level (note that we are talking here of many hundreds of such trial-and-error 'experiments' that would take months or even years if using pen and paper). Most importantly, this qualitative leap opens the door for the possibility of fully automating the process of argument formalization.

[3] The situation is obviously idealized, since, as is well known, most of theorem-proving problems are computationally complex and even undecidable, so in many cases a solution will take several minutes or just never be found. Nevertheless, as work in the emerging field of *computational metaphysics* [11,17,5,6,4,12] suggests, the lucky situation depicted above is not rare.

4 Ongoing and Future Work

Computational hermeneutics is well suited for the utilization of different kinds of classical and non-classical logics through the technique of shallow *semantic embeddings* [2,3], 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 the technique of semantic embeddings we can, for instance, embed a modal logic by defining the modal operators as meta-logical predicates in HOL and using quantification over sets of objects of a definite type w (the type of possible worlds or situations). This gives us two important benefits: (i) we can reuse extant ATP technology for classical logic and apply it for automated reasoning in non-classical logics (e.g. free, modal, temporal or deontic logics); and (ii) the logic of formalization becomes another degree of freedom and can be fine-tuned dynamically by adding/removing axioms in our metalanguage (HOL).

In previous work (e.g. [12,13]) we have illustrated how the *computational hermeneutics* approach can be carried out in a semi-automatic fashion: We work iteratively on an argument by (i) tentatively choosing a logic for formalization; (ii) fixing truth-values and inferential relations among its sentences; and (iii) working back and forth on the formalization of its axioms and theorems, by 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 criteria of adequacy, such as the one shown above (see Peregrin and Svodoba [15,16]).

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 is characteristic of 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. Here we see great potential in the combination of ATP with machine learning techniques.

We are currently exploring the way to fully automate this process. The idea is to tackle 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 provide our main selection criteria: an adequate formalization must validate the argument. 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, with today's technologies, not only completely automizable, but also seems to be highly parallelizable.

References

1. M. Bentert, C. Benzmüller, D. Streit, and B. Woltzenlogel Paleo. Analysis of an ontological proof proposed by Leibniz. In C. Tandy, editor, *Death and Anti-Death, Volume 14: Four Decades after Michael Polanyi, Three Centuries after G.W. Leibniz*. Ria University Press, 2016.
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. Automating Gödel's ontological proof of God's existence with higher-order automated theorem provers. In T. Schaub, G. Friedrich, and B. O'Sullivan, editors, *ECAI 2014*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 93 – 98. IOS Press, 2014.
6. C. Benzmüller and B. Woltzenlogel Paleo. The inconsistency in Gödel's ontological argument: A success story for AI in metaphysics. In S. Kambhampati, editor, *IJCAI 2016*, volume 1-3, pages 936–942. AAAI Press, 2016.
7. N. Chomsky. *The minimalist program*. MIT press, 2014.
8. D. Davidson. Radical interpretation interpreted. *Philosophical Perspectives*, 8:121–128, January 1994.
9. D. Davidson. *Essays on actions and events: Philosophical essays*, volume 1. Oxford University Press on Demand, 2001.
10. D. Davidson. *Inquiries into Truth and Interpretation: Philosophical Essays*, volume 2. Oxford University Press, 2001.
11. B. Fitelson and E. N. Zalta. Steps toward a computational metaphysics. *Journal of Philosophical Logic*, 36(2):227–247, 2007.
12. D. Fuenmayor and C. Benzmüller. Automating emendations of the ontological argument in intensional higher-order modal logic. In G. Kern-Isberner, J. Fürnkranz, and M. Thimm, editors, *KI 2017: Advances in Artificial Intelligence*, volume 10505, pages 114–127. Springer, 2017.
13. D. Fuenmayor and C. Benzmüller. A case study on computational hermeneutics: E. J. Lowe's modal ontological argument. Preprint available on PhilPapers, submitted, 2017.
14. R. Montague. *Formal Philosophy: Selected Papers of Richard Montague. Ed. and with an Introd. by Richmond H. Thomason*. Yale University Press, 1974.
15. J. Peregrin and V. Svoboda. Criteria for logical formalization. *Synthese*, 190(14):2897–2924, 2013.
16. 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.
17. J. Rushby. The ontological argument in PVS. In *Proc. of CAV Workshop "Fun With Formal Methods"*, St. Petersburg, Russia, 2013.
18. A. Tarski. The concept of truth in formalized languages. *Logic, semantics, metamathematics*, 2:152–278, 1956.