The Open University

Open Research Online

Automatic Grounding of Vague Geographic Ontology in Data

Allan Third, Brandon Bennett and David Mallenby⋆

School of Computing, University of Leeds, Leeds, LS2 9JT, UK
{thirda,brandon,davidm}@comp.leeds.ac.uk

1 Introduction

In constructing an ontological theory of a domain such as geography, it is important not only to take account of the vagueness and ambiguity which is inherent in many of the relevant concepts, but also to be able to relate the high-level definitions of the theory to actual sets of data of varying kinds. Any attempt to ignore or remove vagueness and ambiguity risks errors and conflict in the ontological theory with the knowledge of different domain experts, while an inability to ground the theory in real data limits its practical use. We present here a means of structuring such a theory to handle these issues in a principled manner, which lends itself to concrete implementation. We illustrate with reference to several examples from the domain of hydrography.

We say that a term is *vague* if there exist entities to which it is unclear whether that term applies or not. For example, in the hydrographic domain, the distinction between "stream" and "river" is vaguely defined — there are water features which it seems acceptable to classify as either large streams or small rivers. We could make these terms precise by stipulating a dividing line between their interpretations, but to do so is to move away from their actual use in the real world, in which vagueness is a real phenomenon.

Many terms in natural language are ambiguous — that is to say, they have two or more meanings. Ambiguity can often arise from different perspectives on a domain. From the point of view of the production of maps, the surface of the Earth can be seen as the surface of solid material with only air or water above it, whereas such a definition would omit regions such as the floors of caves, and so on, which would be considered part of the Earth's surface by an individual person walking. Terms such as "estuary" can be ambiguous between a feature of the two-dimensional geometry of water regions, and a region of water which has salinity lying within a certain range of values [Cameron & Pritchard, 1963]. A particular set of data will reflect the perspective of those who design it, and so the available data can affect the choice of definitions it is appropriate to use when relating a theory to real data. We believe that having an explicit model of this kind of ambiguity can make it possible to draw useful conclusions based on multiple sets of data embodying different perspectives on the domain in question.

Given a suitable set of data, and a high-level ontological theory, it is possible to provide an explicit mapping from individual data points and their attributes to the primitive symbols of the theory, in such a way that the data generate a model of that theory. We

refer to this process as "grounding" the theory on the data. For example, given a data-set consisting of points in space with the attributes of being either "land" or "water", a theory of hydrography defining terms such as "river" and "lake" can be grounded in those data by specifying which sets of points together constitute individual rivers and lakes. Grounding can simplify automated reasoning by providing a single explicit model of a theory which can be used to assist deduction.

We propose a way of structuring an ontological theory in such a way as to maintain the flexibility to handle vagueness and ambiguity. We also illustrate how such an approach can simplify the automation of the grounding process.

2 A Layered Theory

Let us divide an ontological theory of geography into three layers: a *general* layer, a *grounding* layer and a *data* layer. The general layer is intended, as the name suggests, to contain the definitions and axioms governing the use of high-level terms, close to the level of natural language. For example, it might define terms such as "river", "lake", "estuary", and so on. Where terms are ambiguous, their definitions in the general layer are intended to capture the meaning common to all contexts of use, where such a common meaning exists. For example, in any context, the term "river" denotes some kind of channel which can contain flowing water. The general layer might also define fundamentals, such as the relevant spatio-temporal logic and the overall structure of the globe and the various matter types which make it up.

The data layer, by contrast, is highly specific, and represents the contents of an individual data-set. It thus contains a set of ground formulae stating which attributes hold of which data point, and which relations hold between which data points. For example, it might contain formulae stating, for a set of regions in space, whether those regions represent, say, water or land on the Earth's surface. The data layer can also represent properties which can be computed straightforwardly from the data, such as spatial relations like "part-of" and "overlap" holding between regions of space, and computed geometric predicates such as whether a given region of water is considered "linear".

Between the general and data layers lies the grounding layer, which is intended to serve as a bridge between the two. The grounding layer takes the primitive predicates of the general layer, such as "water-channel", and provides definitions for them in terms of the predicates of the data layer, such as "water" and "linear". By taking all of these layers together, it is straightforward to see how the data provides a model for the high-level terms of the general layer, and it is possible to draw useful conclusions about the various concepts and facts represented in the theory.

Different grounding layers can be used to relate the general layer to different kinds of data-set, and can be chosen to represent the meanings of ambiguous terms in the general layer. A general term such as "estuary" can be grounded in the kind of example data layer described above by defining it in terms of the two-dimensional geometry of water regions. However, should a data-set be available which contains information on salinity, a grounding layer can be constructed which defines estuary along the lines of [Cameron & Pritchard, 1963]. It is thus possible to handle multiple perspectives on a

single term within the framework of one theory. As a consequence, it is possible to use the common terms of the general layer to draw conclusions which can be applied correctly to such different data sets simultaneously.

Another benefit of such a division is that analysis of definitions in the grounding layer can suggest ways of automating the grounding process. For example, given a grounding definition of a river in two-dimensional data as being a vaguely linear region of water, one can construct software which can identify such regions in data and augment the data layer accordingly.

3 Vagueness

We base our treatment of vagueness on *supervaluation semantics* [Fine, 1975]. It is argued in [Bennett, 2001a] and [Bennett, 2001b] that many vague geographic terms are such that, given a partial definition of a term — for example, the set of clear-cut cases of "river" — there remain many "acceptable" ways of making that term precise. That is to say, one interpretation may include certain borderline cases of "river" as genuine rivers, and another may not, without either interpretation contradicting our intuitive understanding of the term.

The applicability of the vague terms in which we are interested turns out to be dependent on certain precise properties which can take a range of values. For example, it seems clear that a river should be longer and wider than a stream, even if it is unclear at precisely which specific values of length and width the boundary between them lies. We can model this phenomenon in terms of *threshold* values on the relevant properties. Thus a specification of unique values for each threshold corresponds to one way of making all vague terms precise. Specifying a *range* of values for thresholds therefore fixes a *set* of ways in which those terms can be made precise. By building such thresholds into the formal definitions of vague terms, we acquire a straightforward means of controlling the set of admissible interpretations for those definitions, and, when grounding such definitions in real data, we can experiment with appropriate ranges for those thresholds and quantify over them to be able to carry out reasoning and draw conclusions respecting the vagueness of the relevant terms. Detailed discussion of this particular approach to the logic of vagueness can be found in [Bennett, 2006].

4 Conclusion

We have outlined a general approach to the structure of ontological theories which handles vagueness and ambiguity arising from different perspectives on the relevant domain, as well as lending itself to convenient automation of the grounding process. Our approach allows results regarding the high-level terms of an ontology to be applied to data arising from a range of different sources, and therefore allows the unification of such data sources at the conceptual level, while maintaining an understanding of the perspectives and goals implicit in the creation of each.

We have implemented our theory in software, and have successfully grounded an ontological theory of the hydrographic domain in real-world data sources, augmented with geometric analysis software to identify the interpretations of predicates such as

"linear". For details of this software, particularly the grounding level, see [Mallenby, forthcoming] and [Mallenby, submitted].

References

[Bennett, 2001a] Bennett, B. 2001a. Application of supervaluation semantics to vaguely defined spatial concepts. In Montello, D., ed., *Proceedings of the fifth international Conference on Spatial Information Theory (COSIT'01)*, number 2205 in LNCS, 108–123. Morro Bay: Springer.

[Bennett, 2001b] Bennett, B. 2001b. What is a forest? on the vagueness of certain geographic concepts. *Topoi* 20(2):189–201.

[Bennett, 2006] Bennett, B. 2006. A theory of vague adjectives grounded in relevant observables. In Doherty, P.; Mylopoulos, J.; and Welty, C. A., eds., *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning*, 36–45. AAAI Press.

[Cameron & Pritchard, 1963] Cameron, W. M., and Pritchard, D. W. 1963. Estuaries. In Hill, M. N., ed., *The Sea: The Composition of Sea-water*, volume 2. Harvard University Press.

[Fine, 1975] Fine, K. 1975. Vagueness, truth and logic. *Synthèse* 30:263–300.

[Mallenby, forthcoming] Mallenby, D. forthcoming. Grounding a geographic ontology on geographic data. In *8th International Symposium on Logical Formalizations of Commonsense Reasoning*.

[Mallenby, submitted] Mallenby, D. submitted. Preparing geographical data to facilitate geographical ontology grounding. *Computers and Geosciences*.