Domain modelling and NLP: Formal Ontologies? Lexica? Or a Bit of Both?

Massimo Poesio
University of Essex
Language and Computation Group

June 13, 2005

Abstract

There are a number of genuinely open questions concerning the use of domain models in NLP. It would be great if contributors to *Applied Ontology* could help addressing them rather than adding to an already long polemical literature ...

1 Empiricists vs. Formalists All Over Again?

In virtually every workshop on ontologies, terminology, or lexical acquisition I attended in the last years at NLP events I found myself watching (or getting involved in) fierce debates as to which approach to domain categorization is 'best': designing a clean, elegant ontology with a clear semantics and based on sound philosophical principles and / or scientific evidence; or relying on evidence from psychology and corpora, and on machine learning techniques, to acquire (automatically, as far as possible) a domain structure that in most cases will be rather messy. (The opposite sides of the argument are presented in (Wilks, 2002) and (Smith, 2004).) Are we back to the bad old days of the 'neat' vs 'fuzzy' debate? A fight between formalists and empiricists for the soul of domain modelling? In this article, I will summarize a computational linguist's view of the issues involved, and suggest ways in which the debate could become more constructive.

2 A Bit of History

Categorizing the world has been a favourite sport of philosophers ever since Aristotle, but the interest of the AI community in categories started with the work on semantic networks by Quillian (1968), and the subsequent proposals by Minsky on frame-based representations (1975).[1] These proposals introduced a

[1] It might be worth noting that the inspiration for this work was primarily psychological: to model associative memory in humans.

very intuitive way of representing knowledge that allowed an easy and efficient implementation of inferences such as inheritance and similarity by means of tree traversal algorithms like spreading activation. For the NLP community, domain ontologies encoded by semantic networks quickly became the solution to all problems in natural language interpretation that seemed to require an interaction with semantics, ranging from exploiting selectional restrictions to guide parsing (in examples such as *Bill saw a man with his telescope*–more on this below) to interpreting noun-noun compounds (e.g., to determine that the most likely interpretation of *airport long term car park* is the one indicated by the bracketing *[Airport [[long term] [car park]]]*). Semantic networks and frame based representations quickly became the dominant KR paradigm both in AI in general and in NLP, where they were used in classic early systems developed by Hendrix and colleagues at SRI, Woods, Brachman and colleagues at BBN (Woods *et al.*, 1980), and in the GUS system (Bobrow *et al.*, 1977). In one area of NLP research in which I am personally involved, anaphora resolution, domain ontologies encoded as semantic networks were used as key ingredients of well-known proposals by Charniak (1972), Sidner (1979), Alshawi (1987) and Carter (1987) to solve, e.g., **bridging references** like *the output* in (1), which refers to a part of *the A1A3A8*.

(1) *Tracing through the A1A3A8 from input pins 47 and 48 to the output*

Much of these early proposals, however, were very informal both with respect to what has been called the 'epistemological' side of knowledge representation (i.e., the semantics of the formalisms used to specify the domain ontology) and with respect to the ontological side proper (what types of objects exist). The calls for establishing domain modelling on more rigorous grounds thus led to a flurry of research: however, different communities identified different problems, leading to the establishment of at least two very different research traditions.

One school of thought advocated the establishment of more rigorous logical and philosophical foundations for domain modelling formalisms. This line of research aimed both to establish a 'Tarskian semantics' for the formalisms used to characterize domain ontologies, starting as early as (Schubert, 1976) and eventually leading to description logics (Baader *et al.*, 2003). Researchers also called for cleaner domain ontologies: this latter line of work started perhaps with (Sowa, 1984), but has been championed most consistently in the work by Guarino and colleagues eventually leading to DOLCE (Gangemi *et al.*, 2003). In NLP, the work on formal ontologies has been particularly influential in Natural Language Generation (NLG)–e.g., (Bateman *et al.*, 1995).[2]

A second, 'cognitive' school argued that the best way to identify epistemological primitives was to study concept formation, semantic priming and learning in humans, whereas the best approach to the construction of domain ontologies was to use (generally, unsupervised) machine learning techniques to automatically

[2] In this article I will not be concerned with another debate concerning semantic interpretation for NLP, that between the users of formalisms with reduced expressive power based on semantic networks, and the users of richer formalisms derived from formal semantics. For a quick discussion, see (Poesio, 2000).

extract such ontologies from language corpora and other data.[3] This approach found its philosophical foundations in the work of Wittgenstein, whereas from a psychological perspective, it was rooted in the work of Rosch, among others, who, many felt, had dealt a mortal blow to the 'classic' (i.e., Aristotelian) theory of concepts (for a useful discussion, see (Murphy, 2002). But this position also fit well with the move towards statistical approaches that took place at the beginning of the '90s. Among NLP researchers, the emphasis on machine learning was motivated by the feeling that automating the knowledge acquisition process was the only way to cope with the fact that knowledge changes continuously (new terms and new instances are introduced all the time). In addition, a machine learning approach was viewed as a solution to the context dependence of ontological information, indicated, e.g., by the fact that only certain parts of an ontology may be relevant for a given context (e.g., the fact that **chicken** is a subtype of the Phylum **Chordata** is relevant for biological applications, but probably not when interpreting kitchen recipes).

One of the recurring themes through this editorial is that both the 'formal ontology' and the 'cognitive' or 'concept-based' approaches to domain categorization have their role in NLP applications; I'll briefly discuss advantages and disadvantages of both in the following two sections. But it's also worth noticing that at least for the moment, and as far as I can tell for the foreseeable future, by far the most successful ontologies for NLP are ones that do not fit easily under any of these categories, and that many would not consider ontologies at all, of which WordNet (Fellbaum, 1998) is the most obvious example.[4] WordNet is an interesting compromise between the two positions sketched above. It is hand-coded, but its structure is dictated primarily by linguistic and psychological evidence–although the authors, though themselves psychologists, acknowledge that the 'hierarchical' approach they adopted may be at variance with recent psychological evidence (see, e.g., George Miller's chapter in the Fellbaum book). Because of this compromise, WordNet is caught between a rock and a hard place: on the one hand, it is attacked by formal ontologists for not being clean enough; on the other hand, it is criticized by NLP researchers for the limitations intrinsic to all hand-coded resources, such as its non-uniform coverage, or the sometimes arbitrary distinctions between senses (see also discussion of word-sense disambiguation below).

3 The Case for 'Clean' Domain Ontologies

Many NLP applications could benefit from ontologies organized according to formalist principles. Some obvious examples are any application allowing its users to extract information from a database containing scientific information

[3] I am conflating for convenience a number of sometimes quite distinct approaches under the 'cognitive' banner, although in a few cases the same researchers pursued both psychological and corpus-based approaches (e.g., in the work on HAL (Lund *et al.*, 1995) and on LSA (Landauer *et al.*, 1998)).

[4] Perhaps the most famous, and arguably most extensive, hand-coded ontology is CYC; however, CYC has had limited impact on the NLP community.

using natural language; or applications extracting information from, say, scientific texts to add to such a database (Gaizauskas *et al.*, 2003). Both types of applications have to map natural language terms into the terms of a specific domain, and WordNet cannot be of much help in such cases, as its coverage of specific domains is limited. The absence of a single, well-defined ontology for a given domain invariably leads to trouble for such applications. A simple example from our own work (Sanchez-Graillet and Poesio, 2004) on mining causal relations from the GENIA corpus[5] should suffice. The GENIA ontology with which the corpus is annotated has only limited coverage, so our system uses the UMLS Metathesaurus[6] to identify terms–but then we had to find a way of mapping the terms from the UMLS database into the terms of the GENIA ontology to evaluate the results. For any application dealing with larger amounts of data, not having to deal with a single unified ontology would mean an almost unsurmountable problem.

So why can't we simply use a formally-defined ontology as our domain model for NLP? One problem is that researchers soon discovered that such ontologies were of limited use for at least one of the tasks to which NLP researchers wanted to put them, using semantic types to determine attachment preferences (see next section). But even if we focus on the primary task for which we would like to use such ontologies, modelling the application domain, problems soon arise. The first is that particularly in the case of information extraction, and especially in bioinformatics, almost every text introduces new terms, so we cannot assume that all terms encountered in the texts we process will already be included in the ontology; in fact, making such an assumption would almost defeat the purpose of our systems, since it would limit their use to finding new properties about known terms. In other words, the ability to add new terms to an existing ontology is crucial even when using an ontology whose structure has been formally defined. The second problem is that researchers working in applications such as bioinformatics are not particularly rigorous either in their use of standard terms or in respecting ontological assumptions: e.g., in many biological papers terms referring to **protein**s are systematically used to refer to the **gene**s that carry the information related to the production of such protein. This suggests that even for such applications we may need what we'll call here a **lexicon**–an organization of the concepts in the domain based on language use as opposed to scientific knowledge. Last but not least, while developing a single domain ontology may be possible for domains such as medicine or biology, in other domains–particularly from the social sciences–things are not so clear. One example of particularly controversial domain is law, although formal ontologies of law do exist (Visser and Bench-Capon, 1998). Whether a classification system for information retrieval applications is best designed according formal criteria, arrived at automatically as done in thesaurus-based IR (Dumais and Chen, 2000), or using a combination of methods (Maedche and Staab, 2001), is a crucial question for work on the Semantic Web.

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/ genia/topics/Corpus/

[6] http://www.nlm.nih.gov/research/umls/

4 The Case for (Linguistically Organized) Lexica

It might be argued that although the empirical approach to NLP has been remarkably successful, most–though not all–of these successes have been in applications in which knowledge about the domain is not essential: speech recognition, part-of-speech tagging, or parsing. It is also true that current methods for the automatic acquisition of concept classification are not yet a threat to hand-coded resources such as WordNet. However, for purely 'language' tasks, automatically extracted classifications generally prove more useful sources of 'semantic' interpretation than domain ontologies in the sense discussed in the previous section. The best example of this is perhaps **selectional restrictions**. The selectional restrictions of verbs clearly affect parsing preferences, as shown by the contrast between *The children ate the cake with their hands*, in which the prepositional phrase expresses a modifier of the verbal phrase / action, and *The children ate the cake with blue icing*, where the prepositional phrase clearly expresses a modifier of the cake. In early work, computational linguists attempted to use domain ontologies to specify such preferences; but this approach was soon shown to be unworkable. Wilks (1975) convincingly argued that selectional restrictions express 'soft' preferences, that can be overridden in a number of ways–e.g, in metaphor, as in the famous example *This car drinks gasoline.* Accordingly, most recent work on determining selectional restrictions (Resnik, 1993) or on using selectional preferences to guide parsing (Hindle and Rooth, 1993) is based on probabilistic models of categorization. More recently, the usability of 'classic' ontologies for another language interpretation task, wordsense disambiguation, has also been called into question, as a result of competitions such as SENSEVAL. The assumptions that it is possible to clearly identify the sense of a word in context, or even to distinguish between its senses, have been challenged (Kilgarriff, 1997; Schütze, 1997). These two examples suggest that the conceptualizations appropriate for modeling domain knowledge may not be appropriate models of the type of information about concepts that is brought to bear in language interpretation (as said above, we will use the term 'lexicon' for this type of information).[7]

Computational linguistics research in language acquisition of the last ten years also suggests, however, that injecting some linguistic (and philosophical) insights into empirical models of language interpretation generally leads to improved results–and that this holds true for models of the lexicon, as well. In language modelling, the original word-based models of word prediction are being challenged by models taking syntactic information into account; whereas in parsing, the original structure-based probabilistic grammars have been replaced by grammars that take the role of heads into account. In the same way, older models of lexical acquisition such as LSA, in which the description of the mean-

[7] Plenty of psychological evidence suggests that effects such as semantic priming can be accounted for in terms of a categorization of concepts based on language use (Landauer *et al.*, 1998).

ing of a word depends solely on the words occurring in its vicinity, are being replaced by models in which the syntactic relations connecting these words play a role (Grefenstette, 1993). Our own work on lexical acquisition (Almuhareb and Poesio, 2004) suggests that better results at concept classification can be achieved by trying to identify the potential **attributes** or **roles** of a concept (in the sense of Description Logics) and by integrating this information in its definition.

5 The Way Forward

It would appear that the solution to the dilemma is pretty obvious–indeed, so obvious that it rarely goes unchallenged. You and I both know that whales are mammals, and that tomatoes are fruit; yet, we are also aware that in actual language use whales are most often treated as fish, and tomatoes as vegetables. In other words, we seem to perfectly capable to take advantage both of a source of domain knowledge based on language use (what we have called above a lexicon) and of one organized along scientific criteria (what we have called domain ontology, or we may also call **encyclopedia**). There clearly are plenty of questions to be answered concerning differences and similarities between these two types of knowledge, and how they are related; yet I feel that there is clearly something right about this distinction, and that trying to answer the questions originated by allowing for two sources of knowledge would be a much more profitable way of spending one's time than trying to develop a domain model that can be all things to all men, successful as WordNet and the Generalized Upper Model may have been (see also (Brewster *et al.*, 2005)). Applications providing routes through dialogue, such as SMARTWEB,[8] should be particularly appropriate to study this interaction, as they need to handle both types of knowledge. An example of language interpretation task at the frontier between lexical knowledge and encyclopedic knowledge is the resolution of **bridging references**. Given examples like (1), it would seem that resolving bridging references clearly requires access to domain knowledge. However, when one starts considering the range of bridging references found in natural language, one quickly finds that in many cases the precise relation between the anchor and the bridging reference is very difficult to identify (Poesio *et al.*, 2004).

6 Conclusions

The conclusions I draw from the observations above can be summarized as follows:

1. Domain ontologies are necessay for all sorts of NLP applications;
2. For many of these applications, a clean domain ontology would be very desirable, but

[8] http://www.smartweb-project.org/

3. Even in these applications, methods for automatically expanding the domain ontology from text will often be needed;

4. However, even for these applications we are likely to find lexicon more useful than a domain ontology;

5. Using the domain ontology instead of the lexicon is unlikely to give good results, but

6. The ontological distinctions identified in formal work may well provide useful guidance for work on lexical acquisition.

7. However, we still need to understand better the relation between lexica and domain ontologies; psychological evidence–e.g., on the resolution of bridging references–may well be useful for this purpose.

I hope that *Applied Ontology* will be the forum for this discussion to take place

References

Almuhareb, A. and Poesio, M. (2004). Attribute- and value-based clustering of concepts. In *Proc. of EMNLP*, Barcelona.

Alshawi, H. (1987). *Memory and Context for Language Interpretation*. Cambridge University Press, Cambridge.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., editors (2003). *The Description Logic Handbook*. Cambridge.

Bateman, J. A., Henschel, R., and Rinaldi, F. (1995). The generalized upper model 2.0. Technical report, GMD Darmstadt. Available online at http://www.darmstadt.gmd.de/publish/komet/genum /newUM.html.

Bobrow, D., Kaplan, R., Kay, M., Norman, D., Thompson, H., and Winograd, T. (1977). GUS, a frame-driven dialogue system. *Artificial Intelligence Journal*, **8**(2).

Brewster, C., Iria, J., Ciravegna, F., and Wilks, Y. (2005). The Ontology: Chimaera or Pegasus. In N. Kushmerick, F. Ciravegna, A. Doan, C. Knoblock, and S. Stabb, editors, *Proc. of the Dagstuhl Seminar on Machine Learning for the Semantic Web*. Available from http://www.smi.ucd.ie/Dagstuhl-MLSW/proceedings/.

Carter, D. M. (1987). *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester, UK.

Charniak, E. (1972). *Towards a Model of Children's Story Comprehension*. Ph.D. thesis, MIT. Available as MIT AI Lab TR-266.

Dumais, S. T. and Chen, H. (2000). Hierarchical classification of web content. In *Proceedings of SIGIR*, pages 256–263.

Fellbaum, C., editor (1998). *WordNet: An electronic lexical database*. The MIT Press.

Gaizauskas, R., Demetriou, G., Artymiuk, P., and Willett, P. (2003). R. gaizauskas and g. demetriou and p. artymiuk and p. willett. *Journal of Bioinformatics*, **19**(1), 135–143.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2003). Sweetening wordnet with DOLCE. *AI Magazine*, **24**(3), 13–24.

Grefenstette, G. (1993). SEXTANT: extracting semantics from raw text. *Heuristics*.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, **19**(1), 103–120. Special Issue on Using Large Corpora, I.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, **31**, 91–113.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, **25**, 259–284.

Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, **16**(2), 72–79.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, New York.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Poesio, M. (2000). Semantic analysis. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 6, pages 93–122. Marcel Dekker, New York.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to solve bridging references. In *Proc. of ACL*, Barcelona.

Quillian, M. R. (1968). Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 227–270. MIT Press, Cambridge, Massachusetts.

Resnik, P. (1993). *Selection and information: a class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Sanchez-Graillet, O. and Poesio, M. (2004). Extracting bayesian networks from text. In *Proc. of LREC*, Lisbon.

Schubert, L. K. (1976). Extending the expressive power of semantic networks. *Artificial Intelligence*, **7**(2), 163–198.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning*. CSLI, Stanford.

Sidner, C. L. (1979). *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

Smith, B. (2004). Beyond concepts: Ontology as reality representation. In A. Varzi and L. Vieu, editors, *Proc. of Third FOIS*, pages 73–84. IOS Press.

Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Addison Wesley, Reading, MA.

Visser, P. and Bench-Capon, T. (1998). A comparison of four legal ontologies for the design of legal knowledge systems. *Artificial Intelligence and Law*, pages 1–31.

Wilks, Y. A. (1975). An intelligent analyzer and understander of english. *Communications of the ACM*, **18**(5), 264–274. Reprinted in *Readings in Natural Language Processing*, Morgan Kaufmann.

Wilks, Y. A. (2002). Ontotherapy: or, how to stop worrying about what there is. In *Ontolex 2002 (Workshop held in conjunction with LREC 2002)*, Las Palmas, Canary Islands. Invited presentation.

Woods, W. A., Bates, M., Brachman, R., Bobrow, R., Cohen, P., Goodman, B., Israel, D., Schmolze, J., and Sidner, C. (1980). Research in knowledge representation for natural language understanding—annual report (9/1/79-8/31/80). BBN Report 4513, Belt Boranek and Newman, Cambridge, MA.