THE WHYS AND HOWS OF ONTOLOGICAL SEMANTICS

Victor Raskin
Purdue University

1. Introduction

OS work has been reported to this Colloquium since it started back at IU in 2004. It has dealt primarily with the currently funded projects, most of them in applying NLP to IAS. This work, along with other developments and applications, is continuing and being submitted for presentation at MCLS-06. In this paper, I would like, however, to address a puzzling question: Why are excellent groups and scholars, including those present and represented here, continuing efforts in achieving increasingly explicitly semantic objectives with non-semantic methods with predominantly non-semantic methods (syntax-cum-statistics), without using such available resources as ontologies and ontological lexicons? Since I have never missed an opportunity to express my sincere appreciation of, and admiration for, these groups, scholars, and their work. my question is based on pure intellectual curiosity, not on thinly veiled superiority. I want to know whether my distinguished colleagues detect a major flaw in computational/ontological semantics that I am prevented from seeing due to certain prejudices and predilections they do not have.

2. Prehistory

The NLP Lab at Purdue University in cooperation with CS/Colgate (1982-87), CMT/CMU (1987-1994), and CRL/NMSU (1994-2000) has established and tested in a number of applications a knowledge- and meaning-based approach to NLP called ontological semantics (OS). Since 1999, NLP Lab has cooperated with CERIAS in applying the approach to information assurance and security (IAS) tasks. The term 'ontological semantics' is also used by Sergei Nirenburg's ILIT/UMBC for a somewhat different approach, sharing some theoretical premises reported in Nirenburg and Raskin (2004) but differing in more attention to surface syntax, somewhat shallower semantics of closed-class items, and aversion to incremental implementations. OS is a mature phase of computational semantics approach initiated by Raskin at the Computational Linguistics Lab at Moscow State University in the early 1960s. In the West, computational semantics was founded by Yorick Wilks, Eugene Charniak, and Roger Schank in Lugano, Switzerland, supported by a small private European foundation some three decades ago. Wilks is the only founder who has survived the ruthless anti-semantical onslaught by the largely American syntactico-statistical approach (SSA), dominating NLP even now, when, since the late 1990s, most major US Government BAAs and RFPs have posited explicitly semantic objectives.

3. OS vs SSA: Any Strawmen Here?

The difference between the two schools is critical. Infected by "fear of semantics," SSA tries to use the standard surface parsing and elaborate statistical methods to devine the meaning of text from the observable syntactic and statistical behavior of its units without any attempt to represent meaning directly, and it prides itself on not developing knowledge resources, using none or grabbing opportunistically whatever is easily available on the Web, such as WordNet, even if the resource is not particularly suitable or designed for NLP.

OS, committed to the cause of weak AI, has created large knowledge resources modeling the ones used by humans processing language and information: the constructed ontology representing the human knowledge of the world; the large lexicons representing the lexical entries in terms of pertinent ontological concepts; the ontological parser which represents the meaning of text, clause by clause and sentence by sentence, to closely approximate the way humans understand them. Supporting the resources is the elaborate acquisition platform ensuring smooth homogeneous acquisition of lexical entries and supporting concepts by various acquirers with varying degrees of training, seamless integration of fully automated methods with a constrained and tightly limited human intelligence, and the incremental progress towards full automation in acquisition.

4. The Case of the Semantic Web

The most publicized use of the term 'semantic' is probably The Semantic Web (SW), whose visionary founder was recently knighted by Queen Elizabeth II, a major NLP expert. But even this initiative remains largely non-semantic.

An enormous effort by many SW visionaries and talented scholars has been spent on developing and perfecting the formalism for letting the semantic content of various Web sites to interact with each other. So why don't they already? For the same reason they never will— unless the SW developers realize that they have done nothing to ensure that there is a methodology for translating the Web content into their formalisms.

For various sociological and academic reasons, SW—just like early NLP—has been developed by computer scientists, logicians, and mathematicians. As the first half-century of NLP development shows, these groups tend to possess an unlimited amount of naivete about natural language, regularly confusing their competence in their native tongue with the linguistic savvy about it. As a result, they lack descriptive techniques, are commonly unable to determine the meaning of the sentence, and dismiss meaning anyway because they feel comfortable only with the "objective" methods of counting observable words and word combinations and—not so well—analyzing the surface syntax. Their inability to assess meaning results in the unavailability of accurate working systems for processing natural language, and

the people who need such systems dismiss them for poor quality: the US Government, the major funder of NLP research, has a pathetic record of deployment of the delivered systems. The few linguists who were allowed to join these efforts were selected for their formalistic inclinations and tried hard to support the others' efforts in avoiding semantics; many of them had no semantic training.

Even since the late 1990s, when the US Government started funding primarily semantic projects in NLP, often using the very language of ontological semantics in the calls for proposals, the anti-semantic forces have revived their doomed attempts to approximate meaning without investing in the ontologies and lexicons but rather by recycling syntactic and statistical methodologies in different packaging. The SW developers are akin to these forces, and if they ever worry about the migration of Web content at all—and there is preciously little evidence of any awareness of this problem in print—they probably assume that it will be done somehow by somebody like the statisticians/syntacticians in NLP.

SW has generated an enormous white elephant called Web Ontology Language, for which the acronym is, strangely, OWL. The large number of *a priori* and explicitly undefined rules have been developed for expressing content in the recommended formalism. Rarely, are these illustrated with a few convenient examples. OWL comes without a methodology for training in acquisition or for acquisition itself. It is tacitly assumed that the Web content owners will somehow learn OWL by themselves and will voluntarily spend a considerable effort in translating their content into OWL and they will do it uniformly, bravely solving the problems of homonymy and ambiguity in unexplored ways. The closest precedent to this wishful thinking was Chairman Mao's idea of increasing steel production in PRC by obligating regular citizens to manufacture steel in their backyard vats. The Chinese did. The Web content owners have and will not—but if they attempt it, the result will be the same as with the Chinese—unacceptably low quality of the product and its consequent non-usability.

5. OS can save the semantic web

Contrary to this approach, OS offers a large developed ontology, with multiple properties interrelating concepts to each other in accordance with human intuition and with easy and tried extensibility to new domains, a battery of lexicons, one for each language, with each sense of a lexical entry clearly and systematically defined in ontological terms, and the ontological parser, producing text meaning representations and increasingly approximating human understanding. Incredibly importantly, OS also comes with an explicit acquisition toolbox, combining human and computerized limited training with an increasingly automated hybrid human-computer system for uniform acquisition of ontological concepts and lexical entries. The lack of such system has rendered CYC, a lovely idea, unusable. Failure to incorporate OS will also leave SW on the drawing tables.

6. Fear of Semantics

There are several reasons for the rejection of computational/ontological semantics.

First, the sociological reason, mentioned above in connection with the Semantic Web: the prevalence of non-linguists in the NLP effort. This is compounded by the educational reason: neither the non-linguists nor the linguists they hire are liberated of the fear of semantics by their education: most linguists think that semantics is "awfully hard," intuitive and unformalizable—and that after a decade of the moribund formal semantics in the 1990s, which formalized everything that could be formalized and ignore most of natural language semantics.

Occasionally, one hears grumbles about the subjectivity of the engineered ontology and ontological lexicons. There are no grumbles about the subjectivity and idiosyncracy of natural language, which actually only exists as idiolects. All of these fears were discussed and dismissed as early as 1995 (see Nirenburg et *al.* 1995). The full paper will deal with these in more updated detail.

There is yet another group of people who are moving into NLP and filling the positions that would more successfully be manned by linguists: psycholinguists, like psychologists a group whose approach is that of statisticians with underspecified theories that often operates under the moniker 'cognitive science.' As statisticians, they bolster the statistical emphasis of SSA and, simultaneously, as bad theoreticians, they either confirm the fears of computer scientists about useless theories muddling their clean formalistic approaches or enforce the anti-theoretical stance by letting them fall by the wayside as well. The use of a new jingoism: vector-space models with singular value decomposition, inappropriately termed 'latent semantic analysis' (LSA), to replace semantics, is a prominent example for this development.

7. References

Nirenburg, S., and V. Raskin 2004. Ontological Semantics. Cambridge, MA: MIT Press.

Nirenburg, S., V. Raskin, and B. Onyshkevich 1995. *Apologiae ontologiae.* In: J. Klavans, B. Boguraev, L. Levin, and J. Pustejovsky (eds.), Symposium: Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity. Working Notes. AAAI Spring Symposium Series, Stanford, CA: Stanford University, 1995, 95-107. Reprinted in a shortened version in: Proceedings of TMI-95, Centre for Computational Linguistics, Catholic Universities Leuven Belgium, 1995, 106-114.