

Ontological Semantic Support for a Specific Domain

Victor Raskin
NLP Lab/CERIAS
Purdue University
West Lafayette, IN

vraskin@purdue.edu

Katrina Triezenberg
NLP Lab/CERIAS
Purdue University
West Lafayette, IN

kattriez@purdue.edu

Evguenia Malaia
NLP Lab/CERIAS
Purdue University
West Lafayette, IN

emalaya@purdue.edu

Olga Krachina

NLP Lab/CERIAS
Purdue University
West Lafayette, IN

okrachin@purdue.edu

Abstract

This paper attempts, out of necessity, to reach two different audiences—never such a great idea! For those familiar with the ontological semantic approach, it focuses, in Section 3, on the crucial issue of expanding ontological semantic resources to new specific domains, which is a decisive factor for spreading the use of the ontological semantic legacy resources and their enrichment and improvement by the fast-growing ontological semantic community. For the majority of the MCLC participants, it also quickly introduces the basic notions of the approach in Section 1 and attempts, in Section 2, to explain why it should be preferred to other approaches if one is in the business of developing the real applications and wants to do it on a principled theoretical basis. The major statement on the approach is, of course, Nirenburg and Raskin (2004), but this paper tries to bring some aspects to a sharper focus as well as updating and upgrading that text written in 1998-2001 and indicating the divergence within the approach. The paper is complemented, within this volume, by Spartz et al. (2005), to which most methodological details are delegated.

1 Ontological Semantics: Basics

The massive 1994-2000 research effort that led to the establishment of the body of the theory of ontological semantics and the 1998-2002 effort at formulating its purview and premises (Nirenburg and Raskin 2004) have brought about the following legacy resources:

- the 5,500 language-independent **ontology**, a tangled conceptual hierarchy, with each conceptual node containing a set of properties, each consisting of a slot, a facet, and a filler, where the first and the last are also concepts (see examples below);
- the 40,000-lexical-entry English **lexicon**, with the meaning zone of each entry anchored in an ontological concept with the appropriate constraints on the property fillers (see examples below);
- several smaller lexicons for a couple of dozen other languages, the biggest of those for Spanish;
- several language-dependent onomastica (pl. of **onomasticon**), i.e., lexica of proper names (see examples below);
- The **Text Meaning Representation (TMR) language** for a compositional and supracompositional representation of the meaning of the sentence and text in ontological terms (see examples below);
- The prototype for the
 - **Preprocessor** taking care of:
 - tokenization,

- ecological processing,
 - morphological processing, and
 - syntactic parsing;
- **Analyzer** that assigns a TMR to sentences and texts;
- **Generator** that transforms a TMR into a text in a natural language or into a set of data;
- The **fact repository** containing every processed TMR and the procedures for searching, comparing, and modifying the property slots;
- The **acquisition toolbox** (see Spartz et al. 2005 for details) for 3-tiered semi-automatic acquisition of ontological concepts and lexical entries, with the massive acquisition implemented by the minimally-trained lowest-tier (Tier 1) acquirers contributing only the intuitive native knowledge of word meaning elicited in a tightly controlled environment under the close automatically triggered supervision by a small number of higher-tiered acquirers (Tier 2, the acquisition managers, determining the class membership for the templates, and Tier 3, the master acquirers, developing the templates for new types of meanings).

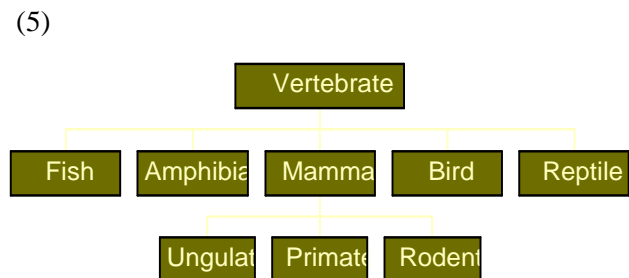
“Sliding down” one level from the ontology top is represented in (1), another level down one of the 3 major branches in (2), two levels down another major branch in (3), three levels down the third major branch in (4).

- (1) ALL
 - Objects
 - Events
 - Properties
- (2) Events
 - Mental events
 - Social events
 - Physical events
- (3) Objects
 - Intangible object
 - force
 - energy
 - Physical object

- animate
- inanimate
- computer data
- physical systems
- Social object
 - geopolitical entity
 - organization
- Mental object
 - abstract object
 - representational

- (4) Properties
 - Case roles
 - agent
 - beneficiary
 - destination
 - experiencer
 - instrument
 - location
 - path
 - purpose
 - source
 - theme
 - Attributes
 - Literal attribute (La)
 - Object-La (O-La)
 - Physical-O-La
 - Social-O-La
 - Event-La
 - Scalar attribute
 - Object-scalar-attribute
 - Event-scalar-attribute

It is obvious, of course, that an ontology will reflect an established classification in (5) very well, and it is mentioned here only for the sake of an easy introduction of the first concept example in (6).



(6)			
MAMMAL			
Definition	value	A warm blooded	
		animal with fur	
Is-a	value	vertebrate	
Subclasses	value	ungulate	
		primate	
		rodent	
English-1	map-lex	mammal	
Italian-1	map-lex	mammifero	

agent ^\$var1
 theme ^\$var2
 instrument NL

But the ontology establishes such a hierarchy among all of its concepts, thus placing all of its concepts in parent-child and sibling relationships. By allowing multiple parenting, the ontology turns itself from a simple tree to a tangled hierarchy, or a lattice.

Most verbs and nouns have lexical entries in the lexicon that are “anchored” in an ontological concept: in the MAMMAL concept above, the English and Italian slots contain the lexical entries in these two languages which are anchored in the concept. Similarly, the English verb say (8) is anchored in the (simplified) ontological concept INFORM (7).

(7) inform			
definition		“the event of asserting something to provide information to another person or set of persons”	
is-a		assertive-act	
agent		human	
theme		event	
instrument		communication-device	
beneficiary		human	

(8) say-v1			
syn-struct			
1	root	say	
	cat	v	
	subj	root	\$var1
		cat	n
	obj	root	\$var2
		cat	n
2	root	say	
	cat	v	
	subj	root	\$var1
		cat	n
	comp	root	\$var2
sem-struct			
1 2		inform	

The syn-struct zone of the lexical entry contains the two possible syntactic constructions (in loose LFG notation—don’t ask!) for its object, an NP, as in *Spencer said a word*, or a complement, as in *Spencer said that he would be late*. The sem-struct presents the meaning for both syntactic usages as the ontological concept INFORM, with its agent slot filled with the meaning of the subject of both constructions, its theme constrained to the meaning of the object in the first construction and of the complement, in the second, and its instrument constrained to NATURAL-LANGUAGE (NL).

The onomasticon has lexical entries for items like *Turkey*, whose sem-struct contains the ontological concept NATION, with its area, population, form of government, GNP, military forces, etc., filled with the actual numbers and concepts for this particular nation.

The simplified TMR for a simple sentence, such as *Mary drove from New York to Boston on Wednesday* combines an event with its property fillers (plus some parametrization—see Nirenburg and Raskin 2004: 284ff.) to look like (9). A “real-life” TMR for a sentence, such as (10), would take longer than a whole column here (ibid: 172-173) and, interestingly, represent it as a 7-event structure because there is no necessary correlation between surface verbs and ontological events nor between surface nouns and ontological objects.

(9) go			
agent	human		
		name Mary	
source		city	
		name New York	
destination		city	
		name Boston	
instrument		automobile	
time		day-of-the-week	
		name Wednesday	

(10) Dresser Industries said it expects that major capital expenditure for expansion of U.S. manufacturing capacity will reduce imports from Japan.

2 Why Ontological Semantics?

This approach is diametrically opposed to the multiple ingenious attempts to avoid ‘brute-force’ semantics and to replace direct representation of meaning through the “simpler” syntactic and statistical processing. The arguments against semantics, much muted recently because of the increased government RFP’s discrimination against the non-semantic and even non-ontological-semantic proposals, have included the cost of the ontological semantic enterprise and the lack of the trained personnel. Both arguments are factually correct and inconsequential. The cost of the ontological semantic enterprise ran into the millions at the peak of its development but, once expended, it does not have to be considered again, and as this paper claims, the cost of adapting and expanding the ontological semantic resources is low. It is also true that few NLP groups include qualified semanticists because the community, which had ruthlessly discriminated against computational semantics for decades, has not prepared enough descriptive semanticists and because, co-incidentally, theoretical linguistics, another potential contributor of the trained personnel, has drifted away from comprehensive linguistic description. The answer to this second anti-semantic argument is that, after the completion of the deployment of the legacy resources at the turn of the century, the ontological semantic approach requires very few highly trained semanticists to maintain and expand the resources. Besides, the success of ontological semantic applications has led to the increase of the demand for appropriate training. In other words, the rationale for the non-semantic approaches to computing text meaning have been reduced to the inertia and perpetuation of one’s favorite methodologies, aka. as the old dogs vs. new tricks dilemma.

As the faith in the practical utility of the non-semantic methods wanes, it is indeed notable that most of the NLP groups nevertheless remain in this mold and pursue their favorite methods inventively, ingeniously, and interestingly. These methods cannot, however, be used to successfully implement real-life, non-toy applications—certainly not within the range of customer acceptance. So, increasingly, the non-semantic NLP groups disconnect themselves from the business of practical NLP. When they respond to the semantic RFPs they position their work as a methodology to

emulate human semantic competence by studying the result of their efforts, such as pre-tagged word sense disambiguation and then studying and mimicking their non-semantic behavior, for instance, with the help of ever increasingly complex and inventive statistical methods. They can claim a non-representational approach to meaning in natural language, originally a Wittgensteinian approach (Wilks 1971). The dichotomy between representation and non-representation can even be presented as the principled basis of the distinction between analytical and phenomenological (“Continental”) philosophy. The syntactico-statistical meaning avoiders in the 2000s’ NLP will be mighty surprised to hear themselves assigned to the phenomenological camp because they consider meaning unknowable (except for the small part covered by grammaticalized semantics and thus accounted for, to an extent, by formal semantics of the 1990s—see, for instance, Heim and Kratzer 1998) and the phenomena they deal with, observable surface syntax and other statistically countable features of texts to be the terra firma required by science.

The response of ontological semantics is, of course, a rigorous, formal theory of meaning which does not limit itself to a part susceptible to one favorite method and which is successfully implemented in a growing array of practical applications, often in new areas of enterprise, such as search engines (check out www.hakia.com/testing) and information assurance and security (Raskin et al. 2002). But the main claim of ontological semantics is along the lines of, “We have already done it—all you have to do is use it.” And this is what the last section and Spartz et al. (2005) are about.

3 Expanding Ontological Semantics to a New Domain

With the increased interest in formal and engineering ontology, much interesting work was done in the late 1980s and 1990s on the formal apparatus for importing, exporting, and meshing ontologies (see Nirenburg and Raskin 2004: 136ff. and references there), including a pretty well developed Knowledge Interchange Format (KIF). The crucial argument for ontological semantics is whether it can be adapted, modified, and/or expanded to fit the needs of any NLP application, independently of

its theoretical and practical orientation and at a reasonable cost, and while the formal means for doing that may be available, it is the content expansion that is of essence here. The customary confusion of formalism with content afflicts NLP along with the philosophy of language, formal semantics and other arenas of romantic formalization, most recently the recently receding infatuation with the Semantic Web (Berners-Lee et al. 2001), where much impressive work has been done, at much cost, on the formal interchanges again but the task of filling the formalism with knowledge content has been dismissively and fatally left to the unaided and, in all likelihood, unwilling website owner.

In this connection, one problem that is irrelevant to the problem of seamless, feasible, and cost-efficient expansion of ontological semantics to new domains is the formalism—even though this is the most frequently asked question. Anybody well-versed in mathematical logic knows that the effability (mutual translatability) principle pertains to formalisms much more directly and simply than to natural languages, so the choice of formalism is a matter of taste and habit. These factors are not irrelevant in terms of the choice personnel and their prior training but it is not a substantive issue. Ontological semantics uses its own set of formalisms, and they are both interchangeable with each other and with any other formalism, whether already compromised by government approval or not.

The legacy ontology and lexicons are a mixture of the top-level general concepts and dependent lexicons developed on the principles of coverage and logic, primarily by the method of rapid propagation (Nirenburg and Raskin 2004: 323-326) and convenience, on the one hand, and of the opportunistic corpus-driven acquisition (ibid: 326-331). The latter is a result of expanding the legacy resources into one specific domain after another, and the choice of those domains depends pretty definitively on available funding. As a result, at any point of expansion, there are some domains which are developed to a considerable depth, many levels down, and others which are barely present. The financial subdomain of mergers and acquisitions, for instance, was represented in great detail due to substantial funding through the massive Mikro-Kosmos project, yet to CRL/NMSU in the mid-1990s. The Purdue effort has moved into the do-

main and subdomains of information assurance and security (Raskin et al. 2002).

Expanding ontological semantics includes the following major stages (cf. Spartz et al. 2005):

- determining the purview of the domain;
- establishing and collecting the representative corpus of texts in the domain;
- establishing the new lexical entries and new senses of the existing lexical entries;
- mapping out the ontological events and objects in the domain and checking on their coverage in the legacy ontology;
- acquiring new ontological concepts and lexical entries/senses.

Our previous experiences in ontological semantic expansion has shown a pretty stable statistics of no more than 350 new ontological concepts and under 1,500 lexical entries per domain and less per subdomain, achievable in about 6 person/months at a cost of under \$20,000 for Tier 2 acquirers.

Since its separation from the sister groups around 2000, the Purdue NLP effort has expanded the legacy resources to the domain of information security and implemented a number of applications in them. It has developed and improved a comprehensive universal methodology for the optimal, parsimonious, and economical ontological semantic expansion into a specific domain, focusing on the higher levels of acquisition, such as systematic and homogeneous solutions for discovering the pertinent branch of an ontology for a concept candidate, for solving the dilemma of adding a child to an ontology vs. modifying the parent's property slots and/or fillers, for limiting the number of senses in the lexicon and deciding whether to add a new sense, for mapping out the ontology of the domain, etc. Two subdomains of the domain of information assurance and security, that of digital identity management and of website privacy policy statements, are being treated ontologically-semantically within the medium-size 3-year NSF ITR and CyberTrust research grants by two of the co-authors (Krachina 2005, Malaia 2005).

References

- Berners-Lee, T., J. Hendler, and O. Lassila 2001. The Semantic Web. *Scientific American*, May 17.

- Heim, I., and A. Kratzer 1998. **Semantics in Generative Grammar**. Oxford: Blackwell.
- Krachina O. 2005. Role of Ontological Semantics in Handling Privacy Policies. Poster presented at the 6th CERIAS Annual Research Symposium, Purdue University, West Lafayette, IN, March 23
- Malaia, E. 2005. Digital Identity Management in Ontological Semantics: Methodology and Practice of Domain Representation. Poster presented at the 6th CERIAS Annual Research Symposium, Purdue University, West Lafayette, IN, March 23.
- Nirenburg, S., and V. Raskin 2004. *Ontological Semantics*. Cambridge, MA: MIT Press.
- Raskin, V., S. Nirenburg, M. J. Atallah, C. F. Hempelmann, and K. E. Triesenberg 2002. Why NLP should move into IAS? (14 pp.). In: Steven Krauer (ed.), **Proceedings of the Workshop on a Roadmap for Computational Linguistics at COLING 2002: 19th International Conference on Computational Linguistics**, Taipei University, Taipei, Taiwan, August 2002.
- Spartz, J., E. Malaia, and C. Falk 2005. Methodology and Tools for Ontological Semantic Acquisition. In this volume.
- Wilks, Y. A. 1971. Decidability and Natural Language. *Mind* 80.