Ontological Semantic Text Processing in the Biomedical Domain

Sergei Nirenburg, Marjorie McShane, Margalit Zabludowski,
Stephen Beale, Craig Pfeifer

Working Paper 03-05

March 8, 2005

Institute for Language and Information Technologies
University of Maryland Baltimore County

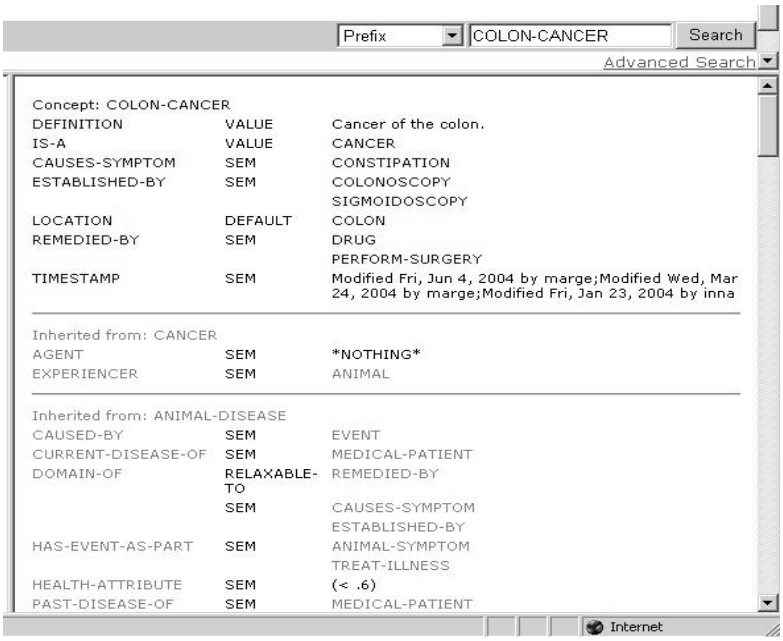
Abstract. This paper describes work in progress on expanding the capabilities of the Ontological Semantics (OntoSem) text processing environment, to the biomedical domain. OntoSem is an ontology-based approach to natural language processing that is geared toward high-end applications and offers solutions to problems that lie outside the range of current and projected knowledge-lean methods. The core system includes knowledge resources and processors for texts in general domains; however, recently we have branched into the biomedical domain, incorporating available biomedical knowledge resources into our architecture. This will permit our text-processing capabilities to be applied to biomedical applications, like question-answering using natural language as the input and output, and reasoning using ontologically-encoded event scripts.

1. An Overview of OntoSem

OntoSem is a text-processing environment that takes as input unrestricted text and carries out preprocessing, morphological analysis, syntactic analysis, and semantic analysis, with the results of semantic analysis represented as formal, fully disambiguated text-meaning representations (TMRs) that can then be used as the basis for many applications. Text analysis relies on a suite of processors and static resources that have been developed together with text processing as the goal.

The OntoSem language-independent **ontology** and **lexicon** are closely intertwined and are written using the same metalanguage of description as the TMRs. The ontology is represented as a directed acyclic graph structured on the IS-A relation but including extensive lateral property-based associations. It currently contains around 6,500 concepts, each of which is described by an average of 16 properties, some locally specified and others inherited. The number of concepts is currently relatively small because: a) the ontology is language independent, so fine-grained nuances of given *words* are not reflected there, they are described in the OntoSem lexicon using property-value specifications; b) we had, until recently, been describing lexical items at a relatively coarse grain-size (a new acquisition effort has refined the grain size and, accordingly, led to an increase in the number of concepts); c) we generally add a new concept only when we wish to attribute to it property values that distinguish it from its parent(s) and siblings (we can, in principle, distinguish any entities using property values, but this would require a tremendous amount of manual acquisition and would result in a grain-size of description that is finer than we are currently pursuing); d) there are many means, outside of ontological mapping, to describe word senses (see below); e) we have been developing the OntoSem knowledge resources in large part on a need-driven basis, for projects that have involved topics in economics (e.g., mergers and acquisitions, bankruptcy) and politics (e.g., travel and meetings); so while the resources have relatively broad coverage across domains, they are deep only in certain domains; f) until our recent work in the biomedical domain, we have not imported any external resources or targeted any specialized domains, which tend to have large inventories of terms that can be considered ontological concepts.

A screen shot of an excerpt from the ontological frame for COLON-CANCER is shown below (many inherited property values are not shown due to space constraints).



The top section of the description includes property-facet-value slots that are locally defined for the given concept. Facets permit further description of the nature of property values: the typical selectional restrictions listed under the SEM facet, default fillers are listed under DEFAULT, impossible fillers are listed under NOT, and extended selectional restrictions (e.g., a dog might eat paper, though it's not typical theme of eating) are listed under RELAXABLE-TO.

The OntoSem ontology also permits the specification of complex events, otherwise known as "scripts," implemented as contents of the HAS-EVENT-AS-PART properties of events. Complex events require an extension to the basic specification format because of the need to bind the variables (case roles and other properties) in component events to establish co-reference. Therefore, the format of the filler of HAS-EVENT-AS-PART allows a) Boolean operators *and*, *or* and *not*, b) conditional statements and c) loop statements. Component events in a script have a special status. They are not regular instances of concepts, as in the ontology no instantiation occurs, but their semantics is different from that of the general concepts to which they are related by name. Thus, the event of asking a question viewed as a component event of a tutoring script involves constraints different from those in the free-standing concept REQUEST-INFORMATION. Scripts in the general domain include bankruptcy, travel, meetings, and so on, with an indication of who participates, what they do, in what order the events occur, with what options and outcomes, etc.—all of which permits reasoning based on world knowledge.

Since the ontology is language independent, the OntoSem environment requires a **lexicon** for each language processed, which contains syntactic and semantic zones (linked using variables) as well as calls to "meaning procedures" (i.e., programs that carry out procedural semantics, like reference resolution; see [1]). The syntactic zone shows what syntactic configurations are possible for the word in the given sense, and

permits the encoding of phrasals and other complex entities with variables. The semantic zone most frequently refers to an ontological concept, either:

- directly: *dog* maps to DOG;
- as a combination of concepts: *weapon of mass destruction* is described as the union of CHEMICAL-WEAPON and BIOLOGICAL-WEAPON;
- as a value on an ontologically defined scale: all modifiers having to do with intelligence are described using values on the abstract scale of INTELLIGENCE that goes from 0-1; e.g., *brilliant (1), smart (.8), clever (.8), witless (.1), stupid (.1), of average intelligence (.5)*;
- using property-based modifications: e.g., *bistro* is a WAITERED-RESTAURANT that is small to medium-sized (SIZE < .5) and not formal (FORMALITY < .5); *Zionist* is a POLITICAL-ROLE that is the AGENT-OF a SUPPORT event whose THEME is Israel; the verb *asphalt* is a COVER event whose INSTRUMENT is ASPHALT; the verb *recall* (as in *they recalled the high chairs*) is a RETURN-OBJECT event that is CAUSED-BY a FOR-PROFIT-CORPORATION and whose THEME is ARTIFACT, INGESTIBLE or MATERIAL.

In addition to referring to ontological concepts, the meaning of words and phrases can be described extra-ontologically, for example, in terms of modality, aspect, time, etc. So, *might* in sentences like *This disease might be caused by environmental factors*, will be lexically described as adding epistemic modality (which concerns truth value) with a value of .5 (which conveys uncertainty) to the proposition.

The OntoSem lexicon currently contains around 20,000 highly specified entries, plus another 20,000 more coarsely specified entries from an earlier acquisition effort. Practically all closed-class entities and all senses of the most frequent verbs (based on corpus analysis) are covered.

When text is processed in OntoSem, the result is a text-meaning representation, or TMR, which is an unambiguous representation of its meaning that draws upon ontological and lexical specification and is written using the TMR knowledge-representation metalanguage. A sample TMR for the input *He asked the UN to authorize the war*, is as follows:

REQUEST-ACTION-69

AGENT	HUMAN-72
THEME	ACCEPT-70
BENEFICIARY	ORGANIZATION-71
SOURCE-ROOT-WORD	ask
TIME	(< (FIND-ANCHOR-TIME))

ACCEPT-70

THEME	WAR-73
THEME-OF	REQUEST-ACTION-69
SOURCE-ROOT-WORD	authorize

ORGANIZATION-71

HAS-NAME	UNITED-NATIONS
BENEFICIARY-OF	REQUEST-ACTION-69
SOURCE-ROOT-WORD	UN

HUMAN-72

HAS-NAME	COLIN POWELL

AGENT-OF REQUEST-ACTION-69
SOURCE-ROOT-WORD he ; *reference resolution has been carried out*

WAR-73
THEME-OF ACCEPT-70
SOURCE-ROOT-WORD war

This says that the word *ask* instantiates the 69th instance of the concept REQUEST-ACTION, whose agent is HUMAN-72 (the instantiation of *he*, which was resolved as 'Colin Powell' using reference-resolution procedures), whose beneficiary is ORGANIZATION-71 (the instantiation of UN, which was resolved to 'United Nations' using reference-resolution procedures), and whose theme is ACCEPT-70 (the instantiation of *authorize*, whose theme is WAR-73 – the semantic representation of the word *war*).

Automatic production of TMRs relies on a suite of processors which cover preprocessing, syntactic analysis and semantic analysis, including lexical disambiguation, reference resolution and other context-dependent aspects of language-based reasoning (see [2] for a description of the OntoSem analyzers and [3] for the resolution of semantic ellipsis).

In addition to the ontology and lexicon, OntoSem uses two other knowledge bases. The first is an **onomasticon**, or lexicon of proper names, which contains approximately 350,000 entries and is growing daily using automated extraction techniques. Onomasticon entries are associated with concept types at a relatively high level of the ontological tree (e.g., HUMAN, LOCATION), to permit automated acquisition of entity-concept mappings. The second is the **fact repository**, which contains real-world facts represented as numbered "remembered instances" of ontological concepts (e.g., REQUEST-ACTION-69 above). See [4] and the publications and tutorials as http://ilit.umbc.edu for further details.

2. Benefits of Applying OntoSem to the Biomedical Domain

In this section we place our current work in the biomedical domain in the context of our broader approaches to resource building and application development.

2.1. *High-End Text Processing*

In contrast to environments that use "light semantics" or no semantics at all (e.g., using what essentially amounts to advanced fuzzy textual string matching as the basis of information extraction) OntoSem can:

- formally represent, in TMR, the distinction between such statements as *Drug route affects colon cancer survival* and *Drug route might affect colon cancer survival*, the latter supplied with epistemic modality with a value of .5;
- represent paraphrases of a locution by the same TMR, as in *Drug route might <could possibly> affect colon cancer survival*;
- resolve reference so that *Drug route might affect it* will be represented in TMR the same as *Drug route might affect colon cancer survival*, making both statements equally available as data for question-answering, summarization, knowledge extraction, etc.;
- facilitate a natural language user interface for knowledge base querying;
- attributing partial semantic analyses to sentences that fail to be fully parsed syntactically.

These basic OntoSem text processing capabilities can be applied to any domain, as can the general domain knowledge resources, which account for a large amount of the text in any specialized field.

2.2. *Building a Biomedical Fact Repository from TMRs*

One direction of recent work on OntoSem has been to develop the capability of automatically populating a fact repository from TMRs, such that the fact repository can be used as the search space for responding to queries. The fact repository is written in the TMR language and, as such, is language independent. Moreover, because reference resolution is carried out during the generation of TMRs, facts about a given instance of a concept are stored in the same fact repository entry. This permits broad queries like *Tell me about X* to be answered in a direct way, by converting into a natural language response all the information stored about the given concept instance. In addition, such a fact repository acts as a structured, language independent, semantic rendering of potentially vast quantities of text.

2.3. *Portability to Other Languages*

OntoSem is a multilingual text-processing environment that has been used in the past for languages such as Spanish, Russian, Chinese and Turkish. Whereas it is typical to assume that lexicons are language-specific whereas ontologies are language-independent, most aspects of OntoSem semantic representations are language-independent. This means that if we consider lexically specified semantic representations – no matter what lexicon they originate from – to be building blocks for *word meaning* (as opposed to concept meaning, as is done in the ontology), then the job of writing a lexicon for a new language based on the lexicon for an existing language consists mainly of providing translation of the head words as well as idiosyncratic syntactic or syntax-semantics linking information. The theoretical basis for this practical (and successfully tested) approach to new lexicon development derives from the Principle of Practical Effability [4], which states that what can be expressed in one language can *somehow* be expressed in all other languages, be it by a word, a phrase, etc. Porting OntoSem to other languages in the biomedical domain would mean, for example, that a Chinese speaking researcher could query the same database as an English speaking colleague, with each using his or her preferred language.

2.4. *Importing Resources*

Developing the OntoSem lexicon and ontology is labor intensive and requires highly trained acquirers. Experience has taught us that there are no shortcuts, and we have become only more convinced over the years that most experiments aimed at saving time (e.g., automatically merging ontologies that cover the same domains, or applying resources to NLP that are not built for it, like machine-readable dictionaries and psycholinguistic-oriented word nets) are of limited practical value [5]. Having said that, we neither can nor should reinvent wheels: the key is to incorporate available resources into our own in closely monitored ways, not begrudging the necessary, non-trivial amount of hand massaging that will bring those resources up to the standards of OntoSem. That is, the gap between available resources and ours is not merely in terms of formalism, it is in terms of depth of semantic content.

2.5. *MeSH and The Metathesaurus*

MeSH is the National Library of Medicine's (NLM's) tree of medical subject headings, arranged hierarchically. The Metathesaurus is NLM's ontology of hundreds of thousands of medical terms along with their synonyms and morphological variants. These resources overlap in part (MeSH being much smaller) and use the same concept identifiers (CUIs). Since we are processing them together, we will refer to them hereafter as M/M.[1]

At the outset, we were hoping that: a) M/M would cover only the biomedical domain, thus not overlapping extensively with our basic, general-domain ontology (merging ontologies that cover the same domain is a well-known hurdle); b) we could attach high-level nodes of M/M to our ontology without needing to "open up" and manually review or edit the trees; c) the available resources would be rich in property-value descriptors.

Our initial review of the resources made it clear that point (c) would not obtain because 61% of the concepts have no property-value specifications at all, and of those that have properties, the properties are not – for our purposes – very useful, as they convey either very general relationships, relationships that would fall out of the IS-A hierarchy anyway (like "narrower than" and "broader than"), or relationships that we could not adequately interpret (like "allowable quantifier"; possibly, further work with the resource could yield better interpretation). Therefore, we settled on knowledge extraction geared toward creating an IS-A hierarchy based on the PAR (parent) and CHD (child) relationships between CUIs (concept identifiers). We used the preferred term for each CUI as the name of the concept, changing its form as necessary for our knowledge base (e.g., multiple words were joined by a hyphen). We then generated lexicon entries for each SUI (string identifier: i.e., lexical or morphological variant of a CUI) and mapped it to the relevant CUI.

Although the IS-A hierarchy is only a fraction of what we ultimately want in our ontology (values of many other properties being crucial for advanced NLP and reasoning), we continued to pursue this experiment for two reasons. First, the inventory of concepts and their linking to different synonymous terms will immediately increase the coverage of OntoSem analysis in the biomedical domain. Second, because of the existing subsumption relations in our ontology, everything that is known about the OntoSem concept to which an imported subtree is anchored is inherited by all the concepts in that subtree; so even if a given imported concept does not have any local properties specified for it, it still inherits properties that can support text processing. Suppose we import a subtree of surgical procedures, which will be linked to the OntoSem node PERFORM-SURGERY, an excerpt of which is shown below:

PERFORM-SURGERY		
DEF	VALUE	A type of medical service where an animal is cut open to treat disease or injury.
IS-A	VALUE	TREAT-ILLNESS
SUBCLASSES	VALUE	COLECTOMY, MESENTERIC-LYMPHADENECTOMY, ...
AGENT	SEM	SURGEON
	RELAXABLE-TO	DOCTOR
INSTRUMENT	INV	SCALPEL
BENEFICIARY	SEM	MEDICAL-PATIENT
DOMAIN-OF	INV	REMEDY-FOR
HAS-EVENT-AS-PART	SEM	CUT
LOCATION	SEM	OPERATING-ROOM
REMEDY-FOR	SEM	ANIMAL-DISEASE

All of the properties of PERFORM-SURGERY, including its typical agent, instrument, etc., will be propagated to the new descendants. For example, knowing that an appendectomy is a surgical procedure will help to resolve the reference for *he* in *He gave her an appendectomy*, should one candidate referent for *he* be a surgeon and the other, a child.

Of course, ideally we would prefer (and intend in the future) to develop an ontological script for APPENDECTOMY, including the typical amount of time spent operating, the actual procedures involved, possible complications, etc.; however, even before that goal can be attained, there are many more levels of concept description that we can seek to achieve that would be more contentful than importing an iconic concept. For example, we can acquire the following:

APPENDECTOMY

HAS-EVENT-AS-PART REMOVE

THEME APPENDIX

To reiterate, although the paucity of property-value specifications in M/M leaves much knowledge acquisition to be done in OntoSem in the long term, it does not negate the potential immediate usefulness of the resource. What does, however, pose immediate problems is a set of issues deriving from the actual content of the M/M hierarchy – issues that make it impossible for us to directly import high-level subtrees without manual inspection and correction.

Problems with directly importing M/M into OntoSem

1. In OntoSem, there is a strict division between concepts, which are stored in the ontology (e.g., CITY, DICTIONARY), and instances of concepts, which are stored in the fact repository (e.g., Paris, France; Medical Entities Dictionary). In M/M, no such distinction exists, making it necessary for us to manually extract all concept instances from M/M before importing any subtrees into our ontology.

2. M/M contains some concepts outside of the medical domain, and OntoSem contains concepts within the medical domain, leading to ontology merging issues. Whereas we will readily exclude M/M concepts from the general domain (since they are loosely if at all related to medical concepts via properties anyway), we certainly do not want to overwrite OntoSem medical concepts – like COLON-CANCER – with M/M ones (in order to facilitate the wholesale importation of M/M subtrees) because the OntoSem concepts are described more extensively.

3. In OntoSem, the IS-A relation is interpreted strictly, in terms of subsumption, whereas in M/M the parent/child relationship (that we used to build our IS-A tree) reflects a more loose association (see [8] for a discussion of the semantics of IS-A). For example, in M/M the following are all siblings: Gait; Lower extremity pain walking; Lower limb length difference; Barefoot walking; and Extensor thrust. If these concepts were retained as siblings in OntoSem, they would need to inherit the same properties, potentially with different value sets. However, in our terms they are different types of entities which need to inherit different types of properties from different subtrees: whereas Barefoot walking is a type of walking (descendant of WALK), Lower Extremity Pain Walking is a symptom (descendant of SYMPTOM), and Lower limb length difference should most likely be a binary property that we have not yet defined but need to, as by LIMB-LENGTH with values SAME/DIFFERENT.

4. In OntoSem, many phrases are lexicalized with a descriptive semantic representation rather than being rendered as ontological concepts. For example, if we were acquiring a biomedical resource from scratch, we would choose to represent Limb Pain (a concept in M/M) as PAIN (LOCATION: LIMB). This preference for description over concept proliferation is not a requirement, but in order for any reasoning about 'Limb pain' to be carried out (e.g., that it refers to pain in a limb, which is described ontologically as an ARM or LEG), we will need to specify *somewhere* what the collocation means – otherwise it is an opaque linguistic atom.

5. In M/M, many concepts have a very large number of parents whereas in OntoSem it is rare to have more than two parents since multiple inheritance requires checking that all inheritance from both parents is either valid, blocked or merged. Of the concepts in M/M, 651,000 have one or two parents but another 30,000 have 3 or more parents (represented as number of concepts: number of parents): 17075:**3**, 6787:**4**, 3434:**5**, 1907:**6**, 1203:**7**, 715:**8**, 432:**9**, 1,000:**>=10**. As mentioned earlier, many of these "parents" are not parents in the narrow sense of the term used in OntoSem but, instead, concepts related in some unspecified way: e.g., of the 38 root nodes of the hierarchy, a number are sources of information, like SNOMED Intl. 1998 and Medical Entities Dictionary. Therefore, before importing the resource we must induce the true line of descendancy and retain that in the IS-A hierarchy, then use manual analysis to flesh out and record the other relationships involved.

6. M/M contains errors, as must be expected for a resource of its size: e.g., over 14,000 concepts are listed as parents of themselves. However, due to the high quality of the OntoSem resources, such errors must be filtered out in order not to diminish the reliability of our resources in ways that will be difficult to detect and correct later. Unfortunately, most errors are not as easy to detect as entities being parents of themselves.

3. Development plans, Final Thoughts

Importing resources always comes at a cost – in fact, many costs. The first necessary task, which we have already accomplished with M/M, concerns understanding the offerings and manipulating formalisms. The next is selecting which aspects of the resource to use, which is much more complex. As described above, we cannot import M/M in full, nor can we manually reconfigure the entire resource to ideally fit the needs of OntoSem. As a compromise, we are manually seeking useful and unproblematic subtrees, in effect starting from the bottom of the M/M hierarchy and working our way up. For example, the subtrees headed by M/M concepts like Bacteria and Arteries can be directly mapped to the OntoSem concepts BACETERIM and ARTERY because (a) these subtrees contain fully comparable types of entities as defined in OntoSem (i.e., there is no mixing of physical objects, symptoms, events, etc., as in the examples in point 3 above), and (b) BACTERIUM and ARTERY are leaf nodes in OntoSem, meaning that there will be no merging conflicts. All subtrees thus imported will still be subject to manual inspection. Following this relatively cheap, lower-tree concept importation – which will be accompanied by importing the associated lexicon entries into our lexicon – we will turn to manual acquisition using M/M as a resource, in much the same way as we have been using WordNet as a resource for manual lexicon acquisition. The resources needed for such manual incorporation are significant, but not nearly those needed to create the original resource from scratch.

Whereas early on we thought envisioned the M/M ontology as a "plug in" to our base ontology, it has become clear that this will not be possible both due to the many-point integration with the base ontology and due to the necessity of adding at least minimal properties to each concept, like the body part affected by a given disease or surgery (we expect to be able to do this semi-automatically with follow-up manual checking).

We have recently been pursuing (in cooperation with Tim Oates of UMBC) the machine learning of ontological property values, which will be useful for supplementing the imported biomedical subtrees. We create TMRs for multiple texts about a given entity (e.g., elephants) and from the cited property values (e.g., height, weight, food consumed), infer generalized ontological constraints (e.g., an elephant is between x and y pounds and eats foods a, b, and c). This "learning by reading" technique is ideally suited to a knowledge-rich environment like OntoSem and promises to ease the knowledge bottleneck with ever greater efficacy with the continual improvement of our knowledge resources and processors.

A recent evaluation of the OntoSem environment is reported in [9]. We needed to build a new evaluation scheme because in semantics-rich text processing the typical recall and precision statistics are of minimal interest; the real issues involve why certain mistakes in TMR were made and how they can be avoided by better using available knowledge or acquiring more knowledge. Although lexical and syntactic coverage remains an issue for our system, as for all knowledge-rich systems, OntoSem has been used and continues to be used in applications like question-answering and knowledge extraction, setting it squarely among the available options to serve current, real-world needs. At the conference, we intend to report the results of OntoSem processing of medical texts using the newly imported and enhanced medical knowledge.

References

1. M. McShane, S. Beale and S. Nirenburg. Some Meaning Procedures of Ontological Semantics, *Proceedings of LREC*, 2004.
2. S. Beale, S. Nirenburg and M. McShane. Just-in-time Grammar. *Proceedings of 2003 International Multiconference in Computer Science and Computer Engineering.* Vegas, Nevada.
3. M. McShane S. Beale, and S. Nirenburg. OntoSem Methods for Processing Semantic Ellipsis. *Proceedings of HLT/NAACL 2004 Workshop on Computational Lexical Semantics*, Boston, Mass.
4. S. Nirenburg and V. Raskin. *Ontological Semantics*. MIT Press, 2004.
5. S. Nirenburg, M. McShane, and S. Beale. The Rationale for Building Resources Expressly for NLP. *Proceedings of LREC*, 2004.
6. O. Bodenreider, J. A. Mitchell, Alexa T. McCray. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. *Proceedings of the AMIA 2002 Annual Symposium.*
7. A. T. McCray, O. Bodenreider, J. Malley and A. C. Browne. Evaluating UMLS Strings for Natural Language Processing. *Proceedings of AMIA Annual Symposium* 2001: 448-452.
8. W. Woods. What's in a Link: Foundations for Semantic Networks, in D.G. Bobrow & A. Collins (eds.), *Representation and Understanding, Academic Press*, 1975; reprinted in, Collins & Smith (eds.), *Readings in Cognitive Science*, section 2.2.
9. S. Nirenburg, S. Beale and M. McShane. Evaluating the Performance of the OntoSem Semantic Analyzer. *Proceedings of the ACL Workshop on Text Meaning Representation*, 2004.