The Rationale for Building Resources Expressly for NLP

Sergei Nirenburg, Marjorie McShane and Stephen Beale

Institute of Language and Information Technologies
University of Maryland Baltimore County
{sergei, marge, sbeale}@umbc.edu

Abstract

In this paper we argue for the need of NLP-specific resources to support truly high level, semantically oriented applications. We describe what, in our experience, constitutes useful knowledge for such applications and why most extant resources are not sufficient for this purpose, leading our Ontological Semantics group to build its own. We suggest that extensive time and energy *are* being spent on resources for NLP, though not on developing ones of higher utility but, rather, on trying to discover ways of using less than ideal ones. We believe that a more useful long-term approach to the problem of knowledge acquisition for NLP would be to acquire what is needed from the outset, since it is likely that in the end such work will prove necessary anyway.

Introduction. A frequent question asked of our Ontological Semantics (OntoSem) group is, what available knowledge resources do you use? WordNet? FrameNet? XTAG?, etc. The question is valid: a number of research groups are building resources that are claimed to have if not primary then secondary applicability to natural language processing (NLP). So, if one were to assume that knowledge is knowledge – with the implication that any and all knowledge is valuable – then one would expect the developers of a knowledge-based system like OntoSem to voraciously incorporate everything available. We, however, do not do this because past attempts to incorporate resources that were not built explicitly to support semantic-rich text processing were less time efficient than starting from scratch; and, in a practical, application-oriented environment like OntoSem, the potential theoretical insights from experiments in resource merging become secondary to the practical necessity of building systems. Thus, we have been developing a suite of interconnected static resources and processors that are specifically targeted at high-end applications. In this paper we present a brief overview of OntoSem, describe why a number of the most widely reported resources are less applicable to NLP than is widely believed and hoped, and present the opinion that, as a field, we should develop resources that are truly sufficient for high-end NLP rather than spend the same significant amount of time and effort attempting to utilize resources borrowed from other fields or developed for other purposes, with inevitably inferior results.

A Snapshot of Ontological Semantics. OntoSem is a text processing environment that takes as input unrestricted raw text and carries out its tokenization, morphological analysis, syntactic analysis, and semantic analysis to yield formal text-meaning representations (TMRs). Text analysis relies on:

- the OntoSem language-independent **ontology**, which is represented using its own metalanguage and currently contains around 5,500 concepts, each described by an average of 16 properties ("features"), selected from the hundreds of properties defined in the ontology; the number of concepts is intentionally restricted, so that mappings from lexicons are many-to-one;
- an OntoSem **lexicon** for each language processed, whose entries contain (among other information) syntactic and semantic zones (linked through special variables) as well as procedural-semantic attachments that we call "meaning procedures;" the semantic zone most frequently invokes ontological concepts, either directly or with modifications, but can also describe word meaning extra-ontologically, for example, in terms of parameterized values of modality, aspect, time, etc., or combinations thereof;
- a **fact repository**, which contains real-world facts represented as numbered "remembered instances" of ontological concepts (e.g., SPEECH-ACT-3186 is the 3186th instantiation of the concept SPEECH-ACT in the world model constructed during text processing as the embodiment of text meaning);
- the OntoSem text **analyzers**, covering everything from tokenization to TMR creation;
- the TMR language, which is the **metalanguage** for representing text meaning, compatible with the metalanguage of the ontology and the fact repository.

Details of this approach to text processing can be found, e.g., in Nirenburg and Raskin *forthcoming* and Nirenburg et al. 2003. The ontology itself, a brief ontology tutorial, and an extensive lexicon tutorial can be viewed at http://ilit.umbc.edu.

TMRs represent, to our knowledge, the most semantically rich, automatically generated expressions of text meaning of any extant system. They require detailed lexical and world knowledge, most of which must be manually acquired. Many believe that manual knowledge acquisition is too expensive to be feasible, so they work on circumventing this problem: some groups attempt to maximize the use of noisy knowledge in NLP applications, e.g., Krymolowski and Roth (1998); and numerous groups attempt to adapt WordNet for use in NLP (especially with respect to problems of ambiguity): e.g., Mihalcea and Moldovan (2001) automatically generate a more coarse-

grained WordNet, Agirre et al. (2001) add topic signatures to synsets, and Gawronska and Erlendsson (2001) introduce "pointers" between noun and verb synsets.

Within the OntoSem group, the general approach to knowledge acquisition and the development of NLP is quite different and is based on the following tenets:

a) semantically rich NLP is an ambitious but achievable goal and the big payoff – a program that can reason, share knowledge and communicate in human language – is well worth the effort;
b) in order to reach such a goal, we cannot bind ourselves to *a priori* insufficient resources;
c) the practical severity of the knowledge bottleneck is exaggerated: e.g., 12K OntoSem lexicon entries – including the entire closed class and most of the hardest, polysemous verbs – were built by one person in one year, in conjunction with ontology development; achieving a 100K lexicon, which is a reasonable size for broad-coverage text processing, should take significantly less than 8 more person years – very little, in the big scheme of things;
d) acquisition of high-quality knowledge already partially is and should be more automated (our group is just one of those that are currently pursuing various avenues for automation level enhancement), though an analyst must be kept in the loop to handle difficult issues and maintain quality;
e) as mentioned above, it is not the case that little time and money is being spent on resources: every researcher trying to find ways to use or improve non-optimal resources is spending time on resources, just not necessarily on building new high-quality ones;
f) we are interested in the last 10-20% of precision that current stochastic methods fail to achieve.

In the sections below we discuss a number of resource-related issues for NLP, specifically discussing FrameNet, XTAG and WordNet (as representative resources) in relation to OntoSem.

Resources for Syntax. A prevalent area of study in NLP is syntactic parsing. A number of available resources can support parsing English, most notably, FrameNet and XTAG.

FrameNet is primarily a lexicography-oriented resource that includes an inventory of subcategorization frames, lexical items that evoke each frame, corpus examples of each frame with arguments/adjuncts indicated, and the possible ways in which arguments/adjuncts can be syntactically realized (Baker and Sato 2003). We have not utilized FrameNet as a resource because it is currently too small, does not use a standard inventory of case roles (e.g., it has MOVERS and MEANS of transportation as opposed to roles like AGENT, THEME and GOAL; see Fillmore and Lowe 1998), highlights textual collocation information and does not provide a sufficient semantic representation of the meaning of the verb in an ontologically-based metalanguage. However, we foresee its corpus examples as being potentially useful for targeted testing and evaluation of OntoSem TMRs.

XTAG does not address semantics at all but has excellent broad syntactic coverage of English to support effective parsing. The XTAG lexicon associates each lexical item with a class, and that class membership indicates which syntactic transformations the verb permits. We have been working to incorporate XTAG's parsing into the OntoSem environment not only because of its broad coverage (far broader than that of the syntactic component of our current OntoSem lexicon), but also because the OntoSem lexicon indicates only the active voice, with other transformations understood in general terms by the syntactic analyzer, but with no verb-specific parameterization.

Our first experiment with XTAG involved automatically generating syntactic zones of OntoSem lexicon entries using a format converter. For example, all transitive verbs were mapped to our basic transitive template, with the default linking pattern of variables and case roles. While we had hoped that this would save acquirer time – and would leave us the trace of the XTAG category to which the word belonged – it actually did not do so for four practical reasons: a) the most time-consuming part of acquisition is representing the semantics through ontological concepts and extra-ontological means; choosing a basic template – of which we have an on-line inventory – is very fast; b) very often the syntactic template and semantic one do not follow the default pattern since modality, reification of property fillers, and other methods of description are necessary; c) it took a long time to sort through the very large XTAG lexicon for the common verbs that our small acquisition team must currently concentrate on; d) without semantic information about what word sense is intended, it can be difficult to orient oneself when filling in the XTAG templates with OntoSem semantic information.

Having set aside (at least temporarily) that experiment, we turned to another experiment that is still in progress. It involves using XTAG's syntactic information separately from the OntoSem lexicon, in a sort of 2-stage parse. Specifically, we are working on a program that will take the information in the OntoSem lexicon, automatically identify which XTAG verb type it is, and apply the transformations to it so that we will have rules that cover all possible syntactic uses. This program is not simple because, unlike XTAG, OntoSem has semantic mappings that must be maintained, and some of the transformations add additional semantics (e.g., the diathesis *It is John who slept* carries discourse information that the active diathesis does not).

Semantic resources. No available resources that claim to provide semantic support for NLP have proved directly applicable to OntoSem, though some have been indirectly useful: e.g., WordNet is among the many on-line and paper sources of synonyms that our acquirers can use during manual acquisition. A comparison between the representation of verbs expressing change in WordNet and OntoSem will serve as illustration of the difference in semantic richness of these two resources.

In describing the presentation of verbs of change in WordNet Fellbaum (1999b: 252) writes: "...Verb phrases like *change magnitude, change shape,* and *change surface* were entered [as nodes in WordNet] on the basis of purely semantic considerations. These concepts were needed to distinguish three groups of verbs that were otherwise all daughters of one node containing the verb *change*. To have represented verbs like *increase, dwindle,* and *wax* as sisters of verbs like *flatten, bend* and *twist* as well as of verbs like *buckle, fold,* and *smoothen* just did not seem

felicitous and seemed to result in a semantically non-homogenous class."

OntoSem takes the semantic specification of verbs denoting change a large step further, representing these notions beyond iconic listing in a hierarchy. All verbs of change in OntoSem are lexically mapped to the ontological concept CHANGE-EVENT but their respective lexicon entries specify their meaning in terms of preconditions and effects. Take, for example, the verb *increase*, whose meaning depends on the theme of the increase. E.g., if the THEME of the increase is mapped to a SCALAR-ATTRIBUTE – like *price* (mapped to COST) or *height* (mapped to HEIGHT) – then the PRECONDITION has a lower value on the given abstract scale (0-1) than the EFFECT does. A call to a meaning procedure that incorporates the correct scalar into the representation of the change event is listed in the lexical entry for all change events. So, a TMR for *the price increased* (in presentation format) will be:

```
CHANGE-EVENT
  THEME   COST
  PRECONDITION.COST.VALUE < EFFECT.COST.VALUE
  TIME < SPEEECH-ACT.TIME
```

For lack of space, we will mention only three of the many limitations of WordNet. **First**, it does not handle the semantics of adjectives well, as reported in Fellbaum (1999a); compare this with OntoSem's fundamental treatment of even the most polysemous of adjectives, as described in Raskin and Nirenburg 1999. **Second**, different diatheses of a given verb are presented in different parts of the hierarchy: e.g., active *sell* has a superordinate of *exchange* while middle *sell* has a superordinate of *be* – which Fellbaum describes as a result of the design of WordNet (Fellbaum 1999b: 256-257). She further notes that "Researchers who have tried [to] find the semantic properties that are both necessary and sufficient to characterize the class of verbs that can undergo middle formation have not been completely successful..." (259) The search for such semantic overlap is, in our opinion, an invented problem: there need not be any such properties, and an environment for representing semantics should best start from the needs presented by the language rather than the restrictions of a given formalism. **Third**, complex expressions and complex notions (even if expressed succinctly) cannot be integrated, as reported by Fellbaum (1998) (who focuses on idioms, but the same issues arise with semantically compositional expressions). Among the types of excluded entities are: a) idioms that do not fit into any of WordNet's categories N(P), V(P), Adj(P) or Adverbial(P): e.g., *the more the merrier*; b) structures that require negation like *not give a hoot*; c) full sentences; d) idioms that contain variables, like *blow one's stack*; e) idioms that express concepts that can't be paraphrased by a single notion, like *drown ones' sorrows*; f) idioms meaning *become smth..*, as in *hit the roof*. OntoSem, by contrast, permits all of these types of entities, with their corresponding semantic representations, to be expressed in lexical entries that can include variables, optional elements, and expressions of any length or complexity. In short, OntoSem imposes no limits on the granularity of semantic (not to mention syntactic) expressiveness: semantics can be expressed by any combination of ontological mappings, preconditions and effects, property values, values of mood or aspect, etc.; and if a means of representation does not exist, we create it to fill a practical need.

Still and all, the main issue we have with WordNet is its weakness in supporting ambiguity resolution.

Resolving Ambiguity. Ambiguity is the killer challenge for NLP. It is the reason why MT is not simply a code-breaking problem, as was hypothesized by Weaver in the 1940s. It is, therefore, reasonable to say that if a lexicon and an ontology used for NLP do not support disambiguation, it cannot be sufficient for truly high-level applications.

WordNet's inability to support ambiguity resolution is understandable because ambiguity poses virtually no problem for humans, and WordNet seeks to depict how humans organize lexical knowledge. In other words, if WordNet accurately depicts how humans organize lexical knowledge, then use of the resource should presuppose all of the world knowledge, pragmatics, goals and general analytical skills possessed by humans. Machines, however, do not have these advantages. A relevant comparison is the utility of a thesaurus to a native speaker versus its relative opaqueness to a language learner.

WordNet is used by many as a source of knowledge in NLP simply because it is there. Its actual efficacy varies among applications: e.g., Vieira and Poesio (1998) found it of little help in reference resolution, and its utility in query expansion for information retrieval has been mixed (see below). The widespread use of WordNet for NLP has spurred efforts to make it a better NLP resource, with version 2.0 including more noun-verb links and a topical organization for certain domains. However, the nature of this resource as a hierarchy of semantically undefined lexical items remains, we believe, a insurmountable disadvantage for machine processing.

Take, for example, the illustration of WordNet use cited by Fellbaum (1999b: 250-251): "...If users query the verb *brush*, they will find the different senses of this verb each with a different superordinate: one, *brush* as a subordinate, or troponym, of *create* (as in the sentence "He brushed a hole in the coat"), another sense whose superordinate is *clean* ("She brushed the suit"), and a third sense that is a subordinate of *remove* ("He brushed away the crumbs")." This example underscores why WordNet is not sufficient for word-sense disambiguation. In order to disambiguate, a lexicon must contain the information that: the *create* sense – if even listed at all due to its very rare usage – requires an object indicating some sort of hole or opening; the *clean* sense requires an object that indicates a piece of clothing or furniture; and the third sense requires a PP complement headed by *away* or *off*. (Incidentally, the most prevalent sense, of brushing one's hair or teeth, is not mentioned.) Without such information, automated disambiguation cannot be achieved.

Another phenomenon presenting similar hurdles for automatic disambiguation is the lexicalization of metaphors: e.g., Fellbaum (1999a) reports that *heart* as "affection" (*win the hearts of people*) and *heart* as a bodily organ have the same status in WordNet. In OntoSem, by contrast, we deal with non-literal language using a combination of phrasal lexical entries (e.g., *win someone's heart*) and productive processing of non-literal language.

Gonzalo et al. (1998) report that although WordNet can be potentially useful for query expansion, it has yielded few successful experiments because badly targeted

expansion (i.e., for a wrong word sense) degrades performance more than no expansion at all. His group set up an experiment in which they manually disambiguated then tested WordNet's potential to improve text retrieval. But the fact is, by the time one has static resources and programs that are capable of disambiguating, it is unlikely that they will need WordNet's query expansion (unless, of course, one wants to include a user in the loop, as is done by Bagga et al. 1997). Gonzalo et al. conclude that "...the queries have to be disambiguated to take advantage of the approach; otherwise, the best possible results with synset indexing does not improve the performance of standard word indexing." Thus, the utility of WordNet – despite widespread attempts to incorporate it into NLP systems – remains under question.

Closing Thoughts. It has become common practice to consider as self-evident that the extensive citing of WordNet in the NLP literature is proof of its utility. That conclusion is, actually, unfounded: people are certainly trying their best to find good use for it since it is available, but that does not imply that their attempts have shown great promise or that success will improve with better machine learning techniques. A common result of machine-learning efforts with and without WordNet is a small increase in results using WordNet and no indication of where the given work can proceed. Take as examples two experiments from the realm of word sense disambiguation (WSD): Stetina et al. (1998) achieve 75.2% accuracy by choosing the first lexical word sense, and 80.3% using WordNet, and Mihalcea and Moldovan (1998) reach 58% precision in WSD using semantic density in WordNet. However, here the experiments stop: the ML methods have been used, they do the best they can with the available resources but are still far from 100%. These relatively low ceilings of results are expected if the difficult problems of NLP are set upon using resources that do not target the difficult problems, and using procedures that – because they do not use sufficient amounts of deep knowledge – have to be satisfied with results that may be state-of-the-art but are unimpressive in absolute terms.

Naturally, the argument from the other side is that the field – not to mention society – needs results right away, and there is no time to build large knowledge resources. Our response is that time will be spent either way, and if time is spent on developing the resources the community really needs for higher-end applications, in the long run it will be well worth the effort.

Another line of criticism, while conceding attainability of knowledge acquisition, questions the utility of the knowledge of the kind OntoSem uses and generates. A very brief response is that this knowledge is absolutely essential for the success of any of the automatic reasoning systems and has already been successfully used in this capacity in a question-answering system AQUA where it supplied knowledge to enable the operation of the JTP (Fikes et al., 2003) reasoning module. Of course, OntoSem semantic analysis itself involves reasoning, and indeed it uses its own results to attain improved analysis.

One final word concerns EuroWordNet and the various language-specific word nets developed on the pattern of WordNet (see, e.g., *Computers and the Humanities, vol. 32*, 1999). The impetus to follow a well-understood research paradigm is clear and understandable, as is the desire to provide machine processing with at least *some* sort of language knowledge. We believe, however, that it would be useful to step back and ask whether we need such word nets at all. Our opinion is that, for most applications, we need something better and that building a single ontology with high-quality, language-specific lexicons mapped to it is the best hope for real progress in the field.

References

Agirre, E., O. Ansa, D. Martínez & E.H. Hovy (2001). Enriching WordNet concepts with topic signatures. In [N-WordNet].

Bagga, A., J.Y. Chai and A.W. Bierman (1997). The role of WordNet in the creation of a trainable message understanding system. In Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference.

Baker, C.F., Fillmore, C.J., and Lowe, J.B. (1998). The Berkeley FrameNet project. In Proceedings of the COLING-ACL, Montreal, Canada.

Baker, C.F. and H. Sato (2003). The FrameNet data and software. Poster and Demonstration at the ACL, Sapporo, Japan.

[Col-WordNet] Proceedings of the COLING-ACL Workshop on the Usage of WordNet in Natural Language Processing Systems, Montreal, 1998.

Fellbaum, C. (1998). Towards a Representation of Idioms in WordNet. In [Col-WordNet] (pp. 52-57).

Fellbaum, C. (1999a). A semantic network of English: The mother of all wordnets. Computers and the Humanities 32: 209-220.

Fellbaum, C. (1999b). Verb semantics via conceptual and lexical relations. In E. Viegas (Ed.), Breadth and Depth of the Lexicon (pp. 247-262). Dordrecht, Holland: Kluwer Academic Publishers.

Fikes, R., J. Jenkins , and G. Frank (2003). JTP: A System Architecture and Component Library for Hybrid Reasoning. Proceedings of the Seventh World Multiconference on Systemics, Cybernetics, and Informatics. Orlando, Florida, USA.

Gawronska, B. & B. Erlendsson. (2001). Reducing ambiguity by cross-category connections. WordNet and Other Lexical Resources Workshop, NAACL.

Gonzalo, Julio, Felisa Verdejo, Irina Chugur and Juan Cigarran. (1998). Indexing with WordNet synsets can improve text retrieval. In [Col-WordNet].

Krymolowski, Y. & D. Roth. (1998). Incorporating knowledge in natural language learning. In [Col-WordNet].

Mihalcea, R. & D. Moldovan. (1998). Word sense disambiguation based on semantic density. In [Col-WordNet].

Mihalcea, R. & D. Moldovan. (2001). Automatic generation of a coarse grained WordNet. In [N-WordNet] (pp. 35-41).

Nirenburg, S. M. McShane & S. Beale. (2003). Operative strategies in Ontological Semantics. In Proceedings of the HLT-NAACL-03 Workshop on Text Meaning, Edmonton, Alberta, Canada, June.

Nirenburg, S. and V. Raskin (forthcoming). Ontological Semantics. Cambridge Mass.: The MIT Press.

[N-WordNet] Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources, Pitt., PA, June.

Raskin, V. & S. Nirenburg. (1999). An applied ontological semantic microtheory of adjectival meaning for natural language processing. *Machine Translation*.

Stetina, Jiri, Sadao Kurohashi and Makoto Nagao. (1998). General word sense disambiguation method based on a full sentential context. In [Col-WordNet].

Vieira, R. & M. Poesio. (1998). Processing definite descriptions in corpora. In S. Botley and T. McEnery (Eds.), *Corpus-based and Computational Approaches to Anaphora.*