The Gene Ontology as a Source of Lexical Semantic Knowledge for a Biological Natural Language Processing Application

Cornelia M. Verspoor
Los Alamos National Lab
PO Box 1663, MS B256
Los Alamos, NM 87545
verspoor@lanl.gov

Cliff Joslyn
Los Alamos National Lab
PO Box 1663, MS B265
Los Alamos, NM 87545
joslyn@lanl.gov

George J. Papcun
Los Alamos National Lab
PO Box 1663, MS B265
Los Alamos, NM 87545
gjp@lanl.gov

ABSTRACT

Mappings between the Gene Ontology (GO) and terms found in a corpus of selected Medline abstracts are studied as a means to augment the lexicon for a text processing application, and as a source of lexical semantic knowledge. We perform an analysis of term overlaps between GO node terms, gene products in the GO, and a domain corpus in order to evaluate the relevance of the GO to our biological natural language processing application. We find that the GO covers a significant portion of the middle- and high-frequency terms in the corpus. We furthermore apply rules based on text parallelism, text insertion and modification relations to hierarchical relations in the GO to infer lexical semantic relations, and discuss the results of this inference. The results demonstrate the potential for using mappings from ontologies to augment lexica for text processing.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Language Parsing and Understanding—*knowledge representation formalisms and methods, text processing*

General Terms

Algorithms, Design, Standardization

Keywords

Information extraction, lexical semantics, ontological representations

1. INTRODUCTION

In any natural language processing (NLP) application, there is a critical need to manage lexical resources in a manner which supports representation of syntactic and semantic constraints on lexical use. In domains which contain much highly specific terminology, such as the biological domain, it is often a daunting task to construct such lexical resources. We turn, therefore, to existing terminological and ontological resources for the domain. In particular, the semantic grounding provided by an ontology can be extremely important for enabling precise analysis of the meaning conveyed in relevant text sources.

In this paper, we discuss explorations we have made into using the Gene Ontology (GO, http://www.geneontology.org) [1] as a source of lexical semantic knowledge for a text processing application in the biological domain. The target use for the resulting lexicon is a prototype system, currently under development, that aims to extract regulatory relationships from biological text [5], and which depends on the existence of domain-specific lexical resources. While our customer has supplied some lists of terms that are associated with particular semantic types, these lists are invariably incomplete and exist independently of any domain ontology. We therefore look to the GO as a source of richer semantic data for lexical resources, specifically investigating its relevance to our domain corpus and its potential as a datasource enabling the incorporation of semantic generalizations into our NLP system. We will discover that by attempting to mine lexical semantic relations from the GO for use in our NLP application, we also open the possibility of automatic ontology extension by feeding the relations we mine back into the ontology. The work is an elaboration of ideas presented in [7].

2. ANTECEDENT ISSUES

In investigating the utility of the GO as a lexical semantic data source, there are several preliminary issues we must address. We begin by establishing the formal properties of the lexicon used in the NLP system and of the ontology.

The mathematical structure of the lexicon is $\mathcal{G} = \langle T_{text}, \mathcal{T}, G \rangle$, where:

- T_{text} is a term list
- $\mathcal{T}$ is a tree on T_{text}
- $G\colon T_{text} \mapsto \mathcal{T}$ assigns terms to positions in the tree.

The tree $\mathcal{T}$ defines hierarchical relations between concepts, and the mapping G allows for a single term to be mapped

to multiple concepts in the tree (to capture polysemy), or for multiple terms to be mapped to a single concept in the tree (to capture synonymy).

The mathematical structure of the Gene Ontology is $\mathcal{O} = \langle R_{nodes}, T_{labels}, \mathcal{P}, F \rangle$, where:

- R_{nodes} are phrases, node terms
- T_{gps} are gene products annotated to the nodes
- $\mathcal{P} = \langle R_{nodes}, \leq \rangle$ is a partial order on R_{nodes}
- F: $T_{gps} \mapsto 2^{\mathcal{P}}$ is the labeling function, mapping each gene product to a set of ontology nodes.

The goal of our work is to investigate whether it is possible to bootstrap from the structure of the ontology to the structure of $\mathcal{T}$ and to automatically define the mapping G, in whole or in part. To reach that goal, we had to answer some preliminary questions on the terminological overlap between the distinct data sources, the GO and the domain corpus:

1. Are the nodes and gene products really distinct?
$$T_{nodes} \cap T_{gps} \sim \emptyset$$
2. How many ontological terms are also in our domain corpus?
$$T_{common} := T_{nodes} \cap T_{text}; |T_{common}| = ?$$
3. If $|T_{common}|$ is small, then why? Is it still sufficiently large to warrant utilizing the ontology as a data source for the domain?
4. What about $|T_{nodes} - T_{text}|$ vs. $|T_{text} - T_{nodes}|$?
5. Should the lexicon be supplemented by $T_{nodes} - T_{text}$?

3. COLLECTION OF LEXICAL DATA

We began our investigations by creating two independent lexical data sets: one derived from a domain corpus, and the second derived from the GO itself.

Our domain corpus is comprised of 9,336 Medline abstracts downloaded from the National Library of Medicine's PubMed website [6]. The abstracts were selected as relevant to the goals of the information extraction system we are building. The resulting corpus is 2.3 million words in size.

The objective of GO is to provide controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products. We worked with the March 2003 version of the GO. The GO contains several kinds of data: (a) terms, (b) gene products, (c) associations between gene products and terms, a term associated to a gene product indicating a biological process, molecular function, or cellular component of the gene product, (d) isa hierarchical relations between terms, and (e) part of relations between terms. We worked with the terms and gene products, as separate datasets, in our lexical analysis.

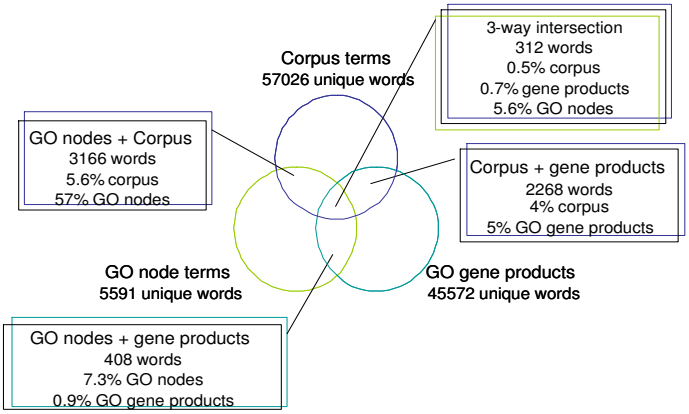

Figure 1: Term overlaps of unstemmed terms in the three term sets

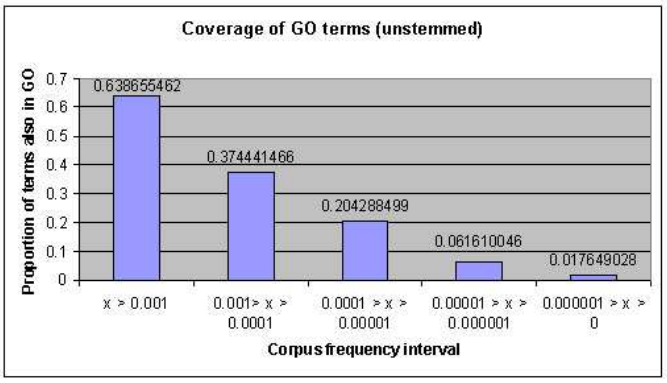

Figure 2: Coverage of unstemmed corpus terms by unstemmed GO terms, by corpus term frequency

After collecting the core datasets, we extracted the individual terms in each set as tokens delimited by whitespace and/or punctuation. That is, we isolated individual unique words in each dataset. This gave us the three sets of terms, T_{gps}, T_{nodes}, and T_{text}. We also calculated the frequency of each of the extracted terms T_{text} in the domain corpus. Finally, we set about to answer the questions of overlap between the three term sets, to determine whether it was worth proceeding. The results of this analysis can be seen in Figures 1-3.

Figure 1 shows the results for unstemmed terms, that is, terms taken directly from the text and not modified or normalized in any way, such that plural and singular forms of nouns or different tenses of the same verb are considered distinct terms. The figure shows that in our domain corpus, a substantial quantity (57%) of the 5591 unique terms occurring in the GO nodes also occurred in the corpus. Relative to the 57062 unique terms extracted from the corpus, however, the overlap was quite small at 5.6%. On the surface, this suggests that the GO on its own is not a sufficiently broad-coverage vocabulary to serve as a rich lexical data source for text processing. However, when we analyzed where the overlapping terms occurred in the frequency-ranked list of corpus terms, shown in Figure 2, we discovered that the cov-

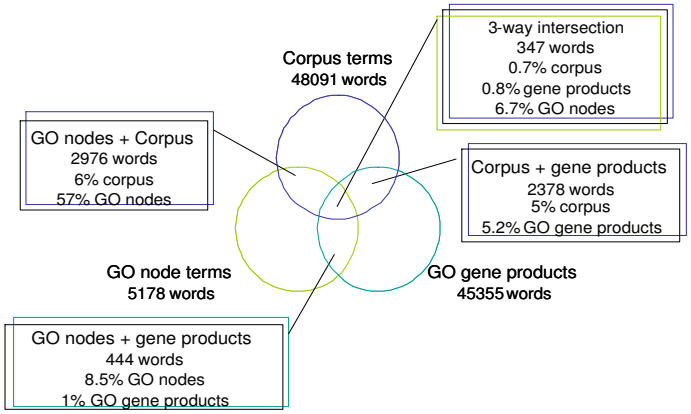

Figure 3: Term overlaps of stemmed terms in the three term sets

Table 1: Recurrence of multi-word phrases from GO nodes in the corpus

433	signal transduction	409	kinase activity
322	cell cycle	261	insulin secretion
203	cell proliferation	197	cell growth
185	growth hormone	177	cell death
170	binding activity	156	tumor necrosis factor
155	cell surface	148	insulin-like growth factor
146	epidermal growth factor	120	plasma membrane
113	cytochrome c	87	glucose metabolism
79	adrenergic receptor	76	activation of MAPK
74	extracellular matrix	73	lipid metabolism
69	tumor suppressor	69	cell adhesion
64	glucose transport	64	cyclase activity
59	cell differentiation	55	enzyme activity

erage of terms in the high- and middle-frequency ranges was quite good, indicating that in fact for many of the terms we are likely to encounter regularly as we process domain texts, the GO could provide a semantic grounding. This was reinforced by superficial analysis of the GO terms not found in the corpus that suggested that the terms in the set $T_{nodes} - T_{text}$ are primarily specific molecule and protein or protein family names rather than more general covering terms. Thus the set of GO node terms can provide a highly relevant lexicon for the domain. Furthermore, to the extent that that these terms can be anchored semantically by exploiting the ontology, there is no reason not to incorporate them into a lexicon for the domain, even when they do not occur in our domain sample.

We explored the overlap of gene product annotations with both the corpus and the GO node terms themselves. The fact that the overlap with the GO node terms is non-null (albeit a small proportion, 5%) suggests that there are some highly specific nodes in the GO that perhaps cannot be expected to have much utility generically in text processing. Any GO node that is tied to a particular gene product is unlikely to occur often in a corpus. However, one might expect greater overlap of gene products with the corpus than we found (only 4% of the corpus terms), given the size of set of gene products in the GO (45,472 terms). This perhaps indicates some bias in the corpus sample we have selected.

The results for stemmed terms are similar, as shown in Figure 3. Here we see that the figures for overlap between each pair of term sets, as well as for the three-way intersection, are very slighly higher than for the parallel sets of unstemmed terms, however on inspection of the term lists it is clear that the stemming is introducing spurious and undesired matches. For instance, *regular* and *regulate* both are reduced to *regul*; *was* and *this* reduced to *wa* and *thi*, respectively, which may inadvertently conflate with completely unrelated terms. On the other hand *studies* reduces to *studi* rather than *study* so the appropriate conflation is not possible in this case. What is needed in order for the incorporation of stemming to be truly effective is a more sophisticated morphologically-based stemmer.

We next looked at how often the full phrasal node labels occur verbatim in the domain corpus. We found that very few occur in the corpus, even fewer if you eliminate single-word labels from the set due to its redundancy with the previous dataset. Overall, only 986 out of 16,475 phrases (6% of the GO node labels) occurred directly in the corpus, of which only 564 were multi-word phrases (3.4%). However, many of the multi-word phrases occurred multiple times in the corpus. Table 1 gives the top-ranking phrases and their corresponding counts in the domain corpus. It is worthwhile to represent these phrases as indivisible lexical units in the lexicon for our NLP system; their recurrence is significant enough to indicate that the phrase as a whole has substantial semantic import for the domain, and the GO provides direct semantic grounding for them (e.g. *kinase activity* is a kind of *enzyme activity* which in turn is a kind of *molecular function*). Thus, we treat these lexical phrases as constructions [3, 5], defined as any learned relationship between form and meaning, which should be explicitly represented independent of any compositional analysis available.

Through the collection of the lexical data introduced in this section, we discovered the answers to the five main questions we were interested in exploring. What we learned is that there is an appropriate overlap between the GO and our domain corpus to warrant utilizing the GO as a lexical data source for our NLP application. However, what we are ultimately interested in using the GO for is *lexical semantic* information; so far we have only explored its utility for *lexical* information. It is the semantic aspect to which we next turn.

4. INFERRING LEXICAL RELATIONS

There are two basic strategies we are utilizing to exploit the world knowledge present in GO for our NLP application. They might be termed the *direct* and the *indirect* strategies, respectively.

The direct strategy refers to directly utilizing the hierarchical relations in the GO for subsumption checking. This is applicable in the case where lexical items correspond directly to GO node phrases, as in the *kinase activity* example above in which the generalization to *molecular function* is allowed simply by following the structure of the GO.

The indirect strategy refers to reasoning upon the ontological relations represented in the GO in order to establish

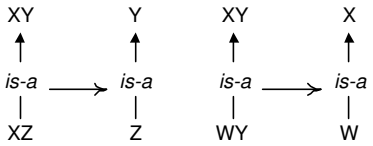

Figure 4: Text Parallelism Rule

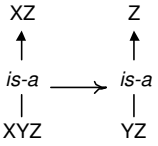

Figure 5: Insertion Rule

ontological relations among individual terms. Specifically, relations between heads of phrases are inferred from the relation between the phrases as a whole. We are exploring the extent to which relations in the GO can be exploited in establishing relations between individual terms in the lexicon.

Currently, our reasoning strategy for inducing lexical semantic relations from the GO utilizes three simple rules. These are not intended to capture the full range of lexical semantic relations which might be induced from the GO, but rather are a first attempt in exploring whether there are meaningful relations that can be induced at all.

1. **Text Parallelism.** This rule attempts to infer an individual lexical relation from a recognized parallelism between phrases where there is some textual overlap between words. See Figure 4. For instance, from the GO relation "lipoprotein metabolism *isa* protein metabolism" we deduce "lipoprotein *isa* protein"; from "lipoprotein biosynthesis *isa* lipoprotein metabolism" we deduce "biosynthesis *isa* metabolism".

2. **Insertion.** This rule handles the case in which a word (or words) are inserted in the middle of a term, creating a child term as a specialization of a parent term. See Figure 5. We have implemented the rule to allow grouping to the right, based on the right-branching structure of English. While this grouping will not always reflect the most intuitive structure of a phrase, in the context of the GO this seems to be more common than a left-branching structure and without implementing full parsing we need to make a (somewhat arbitrary) choice. When this rule is applied, the GO relation "adult feeding behavior *isa* adult behavior" results in the inference "feeding behavior *isa* behavior"; from "chemosensory jump behavior *isa* chemosensory behavior" we deduce "jump behavior *isa* behavior"

3. **Modifier.** This rule handles the case in which one term is a specialization of the other through the introduction of a pre- or post-modifier. See Figure 6. In this case, the rule disallows an inference, following from the recognition that the the modifiers generally modify the entire phrase, and any relation at the level of individual lexical item doesn't make sense. For instance, there is no clear lexical relation to be inferred from "positive gravitactic behavior *isa* gravitactic behavior" or "larval feeding behavior (sensu insecta) *isa* larval feeding behavior".

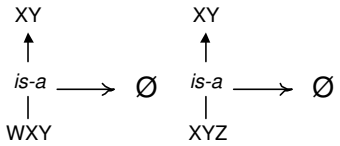

Figure 6: Modifier Rule

Table 2: Lexical semantic relations induced from GO

581	biosynthesis isa metabolism
577	catabolism isa metabolism
44	receptor isa binding
38	deoxyribonucleoside isa nucleoside
35	ribonucleoside isa nucleoside
33	permease isa transporter
27	Saccharomyces isa Fungi
22	porter isa transporter
15	oxidation isa metabolism
14	tRNA isa RNA
14	inhibitor isa regulator
13	ribonucleotide isa nucleotide
11	proliferation isa activation
11	differentiation isa activation
11	deoxyribonucleotide isa nucleotide
10	rRNA isa RNA
10	mRNA isa RNA
9	snRNA isa RNA
8	modification isa metabolism
8	methylation isa modification

These rules can be applied to each parent-child pair in the GO, giving us a set of additional parent-child pairs that can be integrated to form a lexical semantic network. Figure 7 illustrates how these lexical semantic inferences can be linked via the terms they involve to the GO.

In applying these rules, we generated additional parent-child pairs for 5,638 out of 16,849 parent-child pairs in the GO (33%). This corresponded to 2,865 unique parent-child relations. The top-ranking relations are shown in Table 2. We believe that these reflect some fundamental relations; these can form the starting point for a domain ontology at the lexical level as well as the phrasal level. Some of these relations do correspond to existing parent-child pairs in the GO (such as the first two in the table, which correspond to generic physiological processes), but others do not, such as the relationship between RNA, and tRNA, mRNA, rRNA, and snRNA. Overall, only 21 of the 2,865 generated relations already existed in the GO; in Table 2 all but the first two are new parent-child relations not found in the GO.

The rules also result in some problematic inferences. For instance, the right-branching preference in the Insertion rule when applied to "adult male behavior *isa* adult behavior" results in the inference "male behavior *isa* behavior". This inference is not incorrect, but intuitively one would prefer the inference of "adult male *isa* male" from this source relation. This could perhaps be modeled through the incorporation of statistical parsing or, more straightforwardly, reference

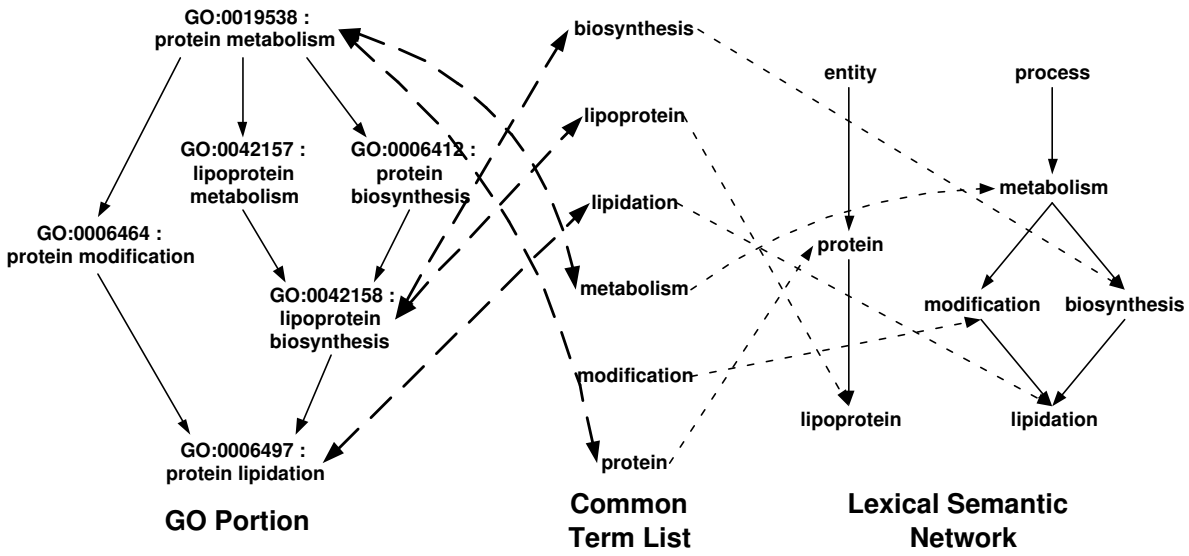

Figure 7: Mappings from the GO to a lexical semantic network

to the relative mutual information of the alternative phrasal analyses. We have not yet tried this.

The Parallel rule sometimes leads to inferences that, independently, seem quite odd. For instance, application of the rule to "maternal behavior *isa* reproductive behavior" and "mating behavior *isa* reproductive behavior" results in "maternal *isa* reproductive" and "mating *isa* reproductive". The inferred relations are rather forced and difficult to interpret. What seems to be going on in this case is (a) there is a context-dependent interpretation of the relationship between the adjective and the noun in these two phrases which is lost when the nominal context is removed (where the parent/child relation expresses something like "maternal behavior" *isa* "behavior in support of successful reproduction") and (b) the *isa* relation does not adequately capture the relation between the parent and the child – in what sense is a maternal behavior *really* a reproductive behavior?

The lexical semantic network which is generated via these rules from the GO can be used to augment the GO itself, in order to extend the GO from a collection of phrasal relations to a more detailed ontology. Along the way, this approach will help to validate the information in the GO by highlighting instances where the *isa* relation may be insufficient, or even by identifying cases where there might be inconsistencies in the GO through recognition of a cycle in the lexical semantic network.

5. APPLICATION

Ultimately, our goal is to incorporate these lexical relations into a NLP system which aims to extract regulatory relationships from Medline abstracts [5]. Our system is built on top of the General Architecture for Text Engineering (GATE) framework (http://gate.ac.uk) [2]. This framework provides the software glue supporting a pipeline NLP architecture in which modules add (linguistic) annotations to documents, default modules for doing NLP tasks such as tokenization and part-of-speech tagging, and a grammar definition language JAPE (Java Annotation Pattern Engine) supporting definition of patterns for information extraction.

The lexicon in our system is represented in terms of gazetteers (term lists) in GATE. GATE itself only supports the assignment of major and minor types to a given list of lexical items. This alone does not provide sufficient semantic granularity to enable precise relation extraction, and furthermore does not allow us to take advantage of the semantic structure provided by the grounding of the terms in the GO. We therefore incorporate extensions to GATE provided by OntoText Lab (http://www.ontotext.com) which allow us to define mappings of ontological categories from GO to lexical features in the GATE lexicon. With this in place, lexical items can be considered by the NLP system in the far richer semantic context provided by the GO. This is achieved by incorporating subsumption checking into the patterns which drive the information extraction.

The hierarchical structure of the GO can be exploited to represent semantic constraints and generalizations in linguistic patterns, since each term derived from the GO is associated with a node in the ontology. For instance, a rule may require that a particular argument be some type of protein metabolism. With reference to the GO, and the additional lexical semantic relations we have induced or perhaps added manually, we can verify that this holds for a given word or phrase identified in the text. These types of constraints allow us to more accurately identify particular relationships.

As an example, in our NLP system we may wish to discriminate a protein reference from a gene reference in text. There are a set of verbs which select for a protein as subject and a gene as object and can be reliably used to make this discrimination. Examples of such verbs are *transactivate*, *upregulate* and *downregulate*. These could be described as PROTEIN-INTERACTSWITH-GENE-RELATIONS, as they are all verbs exhibiting selectional preferences in their subcategorization of [PROTEIN verb GENE]. In our GATE lexicon, we

will treat these verbs as distinct, because they are not synonymous and because each verb is associated with its own set of surface forms. Without any semantic generalization, a separate pattern reflecting the selectional preferences of each verb would need to be defined. However, the existence of an ontology can facilitate the creation of a single pattern [PROTEIN PROTEIN-INTERACTSWITH-GENE-RELATION GENE] which covers all three cases (and any others that pattern similarly). Each verbal base form (the GATE type associated with the set of surface forms) can be mapped to a node in the ontology which is lexically equivalent via term equivalence – see again Figure 7 – and then the subsumption of this term under the PROTEIN-INTERACTSWITH-GENE-RELATION can be validated in the ontology itself.

The previous example cannot be addressed with the information currently available in the GO, but suggests the importance of ontological knowledge representation in information extraction. Another example which can be addressed with the GO is the problem of protein function inference. For instance, we may have an application in which we are required to identify all sentences in which a protein is acting metabolically. Rather than having to spell out all the different kinds of metabolic function, we can draw on the structure of the ontology. For instance, we might define a pattern [PROTEIN serves a METABOLISM function], where we verify that the word in the pre-*function* position maps to a node in the ontology subsumed by *metabolism*. The term *biosynthetic*, for example, maps to the lexical type *biosynthesis*, that maps to a corresponding node in the ontology, that is in turn subsumed by *metabolism*. So the sentence "The lipoprotein serves a biosynthetic function" could be identified as satisfying the more general pattern.

6. CONCLUSION AND FUTURE WORK

In this work we have investigated the potential for exploiting the Gene Ontology, an ontology in the biology domain, as a source of the kind of lexical semantic knowledge that is needed for a natural language processing system in the same domain. We have seen that quantitatively, the overlap between the data in the GO and in our domain corpus is sufficient to warrant utilizing the GO as a lexical data source; taking gene products and node terms together we cover approximately 10% of the corpus terms, and those which are covered are the most frequently occurring terms. But lexical overlap is not sufficient to enable the use of the ontology for our NLP application; we must also show that the semantics of the ontology can be exploited. We have shown that the application of some simple inference rules to the parent/child pairs in the GO can result in the creation of a semantic network that captures core lexical relations for the domain, and can be used to enable generalization in our information extraction system. The GO itself could be augmented, and in turn validated, with these lexical relations.

In future work, we would like to explore using even more of the data in the GO. We might investigate whether it is possible to draw on the definitions of terms in the GO to establish additional lexical relations; words which are used to define a given word can be assumed to have a contextual relationship with that word. This in turn can be used in the NLP system to support word sense disambiguation in the face of words with multiple meanings or in the case of overlapping multi-word units. This is in the spirit of word sense disambiguation work based on machine readable dictionaries [4]. We might also try to make use of the *synonym* relations in the GO. Finally, we would like to do some analysis of the phrasal characteristics of the domain corpus itself, in order to explore further the utility of the GO node terms taken as a unit in the domain.

We must also investigate the properties of the lexical semantic network which is generated from the GO. What are the problems that come up in taking a generated set of individual parent/child pairs and attempting to combine them into a single network? What are the implications of such problems for the GO itself?

More importantly, in this work we have focused on the potential for using the GO to make semantic generalizations. The next obvious step is to draw on those semantic generalizations in the definition of our information extraction system. We have not yet attempted to define patterns that draw on either the original or the inferred relations in the GO. Only when we do this will be evaluate the utility of the world knowledge represented in the GO, either explicitly or implicitly, for our applications. The examples we have given hint at this utility, but the proof will be in the execution of a large-scale information extraction system.

7. ACKNOWLEDGMENTS

This work was funded in part through a Los Alamos National Laboratory collaboration with Procter & Gamble Corporation. A big thank you to our summer student Cheng Lee who wrote many of the scripts that helped us collect the data in this paper.

8. REFERENCES

[1] M. Ashburner, C. Ball, and J. B. et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[2] H. Cunningham. GATE, a General Architecture for Text Engineering . *Computers and the Humanities*, 36:223–254, 2002.

[3] C. Fillmore. Syntactic intrusion and the notion of grammatical construction. In *Berkeley Linguistics Society*, volume 11, pages 73–86, 1985.

[4] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the Fifth International Conference on Systems Documentation*, pages 24–26, Toronto, CA, 1986. ACM.

[5] G. Papcun, K. Sentz, A. Fulmer, J. Xu, O. Lubeck, and M. Wolinsky. A construction grammar approach to extracting regulatory relationships from biological literature. In *Proceedings of the Pacific Symposium on Biocomputing*, Kauai, Hawaii, 2003.

[6] Pubmed. http://www.ncbi.nih.gov/entrez/query.fcgi.

[7] C. M. Verspoor, C. Joslyn, and G. Papcun. Interactions between the gene ontology and a domain corpus for a biological natural language processing application. In *Proceedings of the ISMB'03 Workshop on Bioontologies*, Brisbane, Australia, 2003.