Term Analysis for Ontology Quality Assurance

Karin Verspoor*, Daniel Dvorkin, K. Bretonnel Cohen, and Lawrence E. Hunter
Center for Computational Pharmacology, University of Colorado Denver

1 Introduction

The biomedical domain has seen tremendous growth in the size and coverage of community ontologies. The National Center for Biomedical Ontologies (NCBO) BioPortal, the primary repository of these ontologies, currently contains 129 ontologies and structured hierarchies primarily developed by biologists for use in supporting consistent labeling (annotation) of attributes of entities stored in databases. This growth has been achieved through coordinated efforts across diverse research groups. For instance, the Gene Ontology (GO) [6] is under development by over 20 groups around the world, who collectively have created over 27,000 concepts related to gene products over close to 10 years, and described the *is-a* and *part-of* relations among them. With such large-scale effort, even in one as rigorously curated as the GO, it is inevitable that some inconsistencies in term expression are introduced. In our work, described in [8] and summarized here, we explore strategies for identifying such inconsistencies, and work specifically with the GO.

We call the consistency of expression of concepts in an ontology *univocality*, inspired by the philosophical term referring to a shared interpretation of the nature of reality [5]. In this work, we take univocality to be a primary goal for assuring ontology quality and search for univocality violations.

It has been previously noted that terms in structured controlled vocabularies, such as the Gene Ontology (GO), often have a highly regular, even compositional, linguistic structure and that this structure can be exploited for the purposes of accessing those terms computationally and reasoning over them [7, 3, 4]. This regularity is particularly important now that there are efforts to perform intra- or inter-ontology enrichment by linking terms [1], because the tools that are used to support these efforts analyze the formal structure of the terms and take advantage of patterns of expression. The more consistency in expression there is, the more terms will be able to be appropriately and automatically linked. It is also intuitively important for human usability of the ontologies – the more consistently concepts are phrased, the easier the resource should be to search and augment.

The most basic case of a univocality violation is the occurrence of redundant terms, terms expressing the same meaning with two distinct forms (e.g. "regulation of transcription" and "transcription regulation"), and would certainly indicate an error in the ontology. However, in our work on the GO we were not able to identify any such cases. Rather the univocality failures occurred in *semantically similar*, rather than identical, concepts expressed using different forms. We thus generalize the notion of univocality in this work to apply to similar concepts and thereby assess the ontology quality more broadly. We describe an automated methodology for identifying potential failures of term univocality and apply the method to the Gene Ontology to discover a small but significant number of terms that should be rephrased to improve the overall quality of the ontology.

2 Approach

Our goal is to detect sets of terms within a controlled vocabulary that express similar concepts using different surface forms and are therefore not univocal. We approach this problem through *term transformation* and *clustering*. We hypothesize that pairs of terms which are not univocal

[*] to whom correspondence should be addressed, *Karin.Verspoor@ucdenver.edu*

Table 1: Example term variations indicating a lack of univocality

{*GTERM embryo sac*}		**{*CTERM CTERM oxidati*}**	
		GO:0019327	**oxidation** of lead sulfide
GO:0009562	**embryo sac** nuclear migration	GO:0018158	protein amino acid **oxidation**
GO:0009558	cellularization of the **embryo sac**	GO:0019604	toluene **oxidation** to catechol
		GO:0019479	L-alanine **oxidation** to propanoate
		GO:0019602	toluene **oxidation** via 3-hydroxytoluene

will be transformational variants of one another, such that when they are normalized to a common representational form they will cluster together. We automatically apply transformations to the terms in the vocabulary in order to normalize their form and group terms that have the same form as a result of these transformations together into a cluster. We then utilize an automated heuristic search over the term clusters to identify term occurrences that are expressed non-uniformly, as compared to similar terms.

A pair of terms which are not univocal in the GO appears on the left of Table 1. For consistency, one of these terms should be rephrased, e.g. GO:0009558 could be rephrased "embryo sac cellularization" in order to align not only with 0009562, but also *inter alia* GO:0009553, "embryo sac development".

[4] showed that a large proportion (65.3% in their study) of GO terms contain another GO term as a proper substring. Here, we draw on that insight and perform substitution of the embedded GO term with a generic label in order to better capture the overall structure of the (larger) term. We similarly search for embedded occurrences of terms from the Chemical Entities of Biological Interest (ChEBI) ontology [2] and substitute them with a distinct generic label.

The three transformations we perform are as follows:

- **abstraction**: identification of Gene Ontology or ChEBI terms embedded in a longer GO term, and replacement of this embedded term with a generic *GTERM* token, for an embedded GO term, or *CTERM* token, for an embedded ChEBI term.
- **stopword removal**: elimination of stopwords, or words which do not normally carry semantic content, such as *the*, *of*, etc.
- **reordering**: alphabetic ordering of the tokens within the term.

For example, all of the terms on the right of Table 1 collapse to the form **{*CTERM CTERM oxidati*}** after all three transformations have been applied.

After the transformation and clustering steps, we apply a heuristic search over the generated clusters to identify potential univocality violations. This is an automated search over the full set of clusters (for all combinations of transformations) that draws on the intuition that the abstraction transformation is fundamental to identifying univocality violations – without it we are limited to only considering terms that are near identical at the string level – and that this transformation is the necessary starting point to generalize our univocality analysis to *semantically similar* terms. Thus we only consider clusters for which abstraction has been applied in our search. The algorithm specifically looks for terms which appear in distinct clusters after only the abstraction transformation, but merge upon application of one of the other two transformations. This cluster merging indicates that these terms are semantically similar, but that they differ in terms of either their word order or the stopwords they contain. These differences may indicate a univocality failure. The clusters identified automatically in this way are then examined manually for terms which appear to violate univocality, such as GO:0019327 above, which should be phrased "lead sulfide oxidation" for consistency, and categorized as either a true positive cluster (containing a univocality violation) or a false positive cluster (not containing a univocality violation).

As shown in Table 2, the application of the automated heuristic search to the clusters identified 237 clusters that potentially contain non-univocal terms. Of these, 47 were redundant – e.g. two

Table 2: Results of heuristic search for univocality violations

	clusters
Total candidates	237
Identical	47
False Positive (FP)	123
True Positive (TP)	67

Table 3: Breakdown of false positives (FP)

	clusters	%
semantic import of stopword	61	50%
non-parallel structure	33	27%
semantic import of stemming	21	17%
syntactic variation	6	5%
(other)	2	1%

combinations of transformations resulted in the identical set of terms in a given cluster. The second occurrence of a cluster of terms was identified automatically and discarded in the analysis. This left 190 clusters to be manually reviewed by the first author. 67, or 35%, of these clusters were identified as containing one or more terms that were not univocal with other terms in the cluster. This number of true positive clusters represents only 0.03% of the source GO terms. The total number of terms in these 67 clusters is 374. Many of the terms in each cluster are in the 'correct' (standard) form, while at least one would be in the non-univocal form. We did not specifically count the number of *terms* that were not univocal, as the decision as to which form is standard and which is non-univocal is a curation decision for the ontology curators.

Table 3 provides a breakdown of the false positives in our analysis. The primary source of false positives is that removing stopwords may remove words that in fact indicate important semantic relationships. *Non-parallel structure* corresponds to clusters in which the member terms do not have an obvious common structure on which to evaluate univocality. The false positive category *syntactic variation* is similar to the *non-parallel structure* but refers more specifically to terms which are mostly parallel but show some semantically relevant syntactic variation in expression. Finally, we find that stemming may incorrectly conflate multiple terms. Further details can be found in [8].

3 Conclusion

We have introduced an automated method for identifying violations of univocality among a set of controlled vocabulary terms that reduces the set of terms that need to be examined manually to a manageable size. Using the method, we were able to identify 67 examples of univocality violations in the Gene Ontology that can be addressed in order to improve the quality of that ontology.

References

[1] Michael Bada and Lawrence Hunter. Identification of OBO nonalignments and its implications for OBO enrichment. *Bioinformatics*, 12:1448–1455, 2008.

[2] Kirill Degtyarenko. Chemical vocabularies and ontologies for bioinformatics. In *Proc 2003 Itnl Chem Info Conf*, 2003.

[3] Christopher J. Mungall. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5(6-7), August 2004.

[4] Philip V. Ogren, K. Bretonnel Cohen, George K. Acquaah-Mensah, Jens Eberlein, and Lawrence Hunter. The compositional structure of Gene Ontology terms. *Pacific Symposium on Biocomputing*, pages 214–225, 2004.

[5] Baruch Spinoza. *Ethica Ordine Geometrico Demonstrata (Ethics).* The Collected Works of Spinoza, Volume I, ed. by Edwin Curley (Princeton, 1985), 1677.

[6] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000.

[7] Karin Verspoor. Towards a semantic lexicon for biological language processing. *Comparative and Functional Genomics*, 6:61–66, 2005.

[8] Karin Verspoor, Daniel Dvorkin, K. Bretonnel Cohen, and Lawrence Hunter. Ontology quality assurance through analysis of term transformations. *Bioinformatics*, ISMB 2009 Proceedings, In press 2009.