

Protein Annotation as Term Categorization in the Gene Ontology

Karin Verspoort[†], Tiago Simas[†], Cliff Joslyn[†], Andreas Rechtsteinert[†],
Sue Mniszewski[†], Judith Cohn[†], Luis Rocha[†], Andy Fulmer[‡]

[†]Los Alamos National Laboratory
PO Box 1633, MS B256
Los Alamos, NM 87505

[‡]Procter & Gamble Company
Cincinnati, OH

We addressed BioCreAtIvE Task 2, the problem of annotation of a protein with a node in the Gene Ontology (GO). We approached the task as a problem of categorizing into the GO based on the term neighborhood of occurrences of the protein in the document. The system incorporates NLP components such as a morphological normalizer, a named entity recognizer, and a statistical term frequency analyzer. The categorization methodology utilizes the structure of the GO to select nodes that serve as apparent cluster heads.

Pre-processing

Since we were given only a Swiss-Prot or TrEMBL ID as the input identifier for the protein, we needed to establish a set of names by which that protein could be referenced in the text. We made use of both the gene name and protein names that are in Swiss-Prot itself, when available, and a collection of synonyms constructed by Procter & Gamble Company. The fallback case was to use the name filed from the EBI TrEMBL human. A script was applied to these names that generated variants so that we wouldn't be limited to matching to the sometimes unusual strings containing mismatched punctuation and parentheticals such as "(precursor)" or "(fragment)" which were felt not to be likely to occur directly in the text. The resulting database tables were used to construct a gazetteer list which was dynamically loaded from the database into a GATE (Cunningham et al. 2002) gazetter processing module (which in turn compiles it into a finite state recognizer for the terms).

Additional pre-processing was performed on the document corpus. First, the original SGML documents were parsed to extract the core Title, Abstract, and Body components, to normalize the SGML character entities to their corresponding ASCII characters (for instance, converting "′" to an apostrophe), and to remove all formatting tags apart from the paragraph markers. Subsequently, we morphologically normalized the documents using a tool called "BioMorpher"¹. We performed frequency analysis on the resulting terms, and selected representative terms for each a document using a TFIDF (term frequency inverse document frequency) filter.

Training

The {protein, document, GO id} triples were used to determine sets of terms related to GO ids. After document pre-processing, we divided each document in a set of paragraphs, and calculated a matrix of the paragraphs and the frequency of each term in the paragraph (PxW matrices for Paragraph by Word). With the PxW matrix, we calculated a *proximity* matrix according to Rocha (2002) for word-word proximity, using the probabilistic function at right.

With these matrices, the training triples, and the terms derived from each GO node label, we were able to recommend new terms for each GO node.

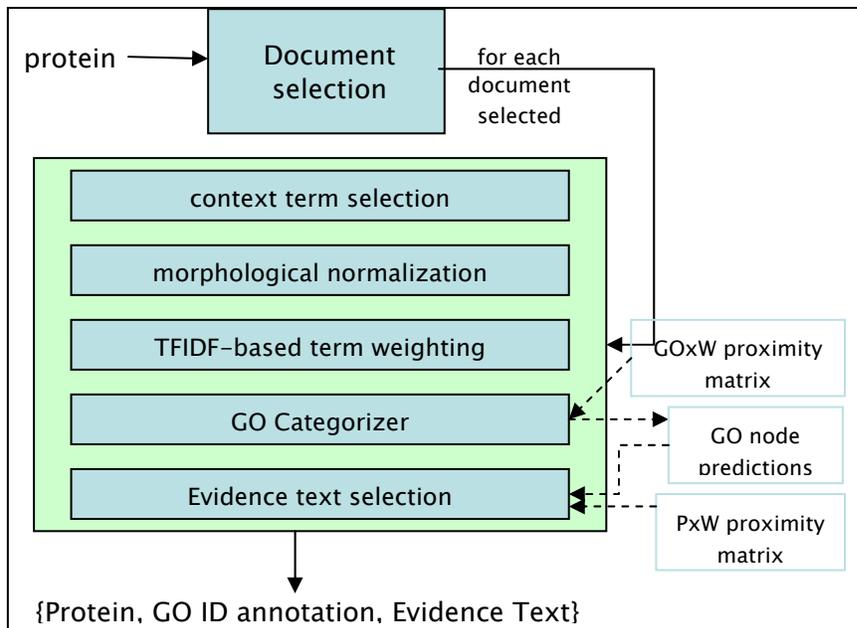
$$pro(w_i, w_j) = \frac{\cap(w_i, w_j)}{\cup(w_i, w_j)}$$

Using the expanded set of terms associated with each GO node and the PxW matrices, we were then able to recommend a set of paragraphs for each {GO, document} pair based on the relationship between terms in each paragraph and terms associated with the GO node.

Dynamic processing

The architecture of the system can be seen below.

¹ BioMorpher is a morphological analysis tool built on the Morph tool originally developed at the University of Sheffield by Kevin Humphreys and Hamish Cunningham for general English, extended to include large exception lists for biological text as well as to handle some morphological patterns not handled by the original tool.



The system is built around a technology called the GO Categorizer (GOC, Joslyn et al. 2003), which utilizes the structure of the Gene Ontology to find the best covering nodes given a set of node "hits". It is based on pseudo-distances between comparable nodes, with rank ordering of nodes balancing **coverage** – covering as many inputs as possible – and **specificity** – covering inputs at the lowest level possible. For BioCreAtIvE, we

submitted **terms** as inputs to GOC rather than genes. Terms are collected through analysis of the sentential context of the given protein, morphologically normalized, and weighted using a normalized TFIDF value derived during pre-processing. The weights represent the contentfulness of each term. Internally, GOC looks for overlaps between the input term set and (morphologically normalized) terms associated with each individual node in the Gene Ontology. A match between an input term and a term associated with a GO node counts as a "hit" on that node. The strength of that hit is determined by the weight of the term in the input set. Terms are associated with GO nodes via one of three mechanisms – (a) **Direct**: the term occurs in the node label of GO node, (b) **Definitional**: the term occurs in the definition text associated with GO node, (c) **Proximity**: additional terms are identified as closely related to each GO node based on proximity as described above (Rocha 2002). Direct and indirect associations are counted as distinct "hits" on a node and can be weighted differently. After transforming the input query into a set of node hits, GOC traverses the structure of the Gene Ontology, percolating hits upwards, and calculating scores for GO nodes (see Joslyn et al 2003). GOC returns a set of GO nodes representing cluster heads for the weighted term input set, as well as data on which of the input terms contributed to the selection of each cluster head. This information is used to select the **evidence text** for the GO assignment associated with the cluster head. To address this, we again bring in **proximity measurement** – in this case, the proximity of terms to individual paragraphs in the document. The set of terms which contributes to an annotation is judged to be close to one or more paragraphs in the document; the closest match is selected as the evidence.

References

- H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.
- Joslyn, C., S. Mniszewski, A. Fulmer, G. Heaton (2003). "Structural Classification in the Gene Ontology". In Proceedings of the Sixth Annual Bio-Ontologies Meeting (Bio-Ontologies 2003), Brisbane, Australia, June 28, 2003.
- Rocha, Luis M. (2002). "Semi-metric Behavior in Document Networks and its Application to Recommendation Systems". In: Soft Computing Agents: A New Perspective for Dynamic Information Systems. V. Loia (Ed.) International Series Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 137-163.
<http://www.c3.lanl.gov/~rocha/semimetricIOS.html>