Supplementary Materials: Machine learning models for automatic Gene Ontology annotation of biological texts*

Jayati H. Jui[0000−0002−9718−6387] and Milos Hauskrecht[0000−0002−7818−0633]

University of Pittsburgh, Pittsburgh, PA 15260, USA
{jaj146,milos}@pitt.edu

1 Dataset Distribution

Table 1. Distribution of the three GO Categories in the Benchmarking Corpus.

Corpus	GO Categories		
	Biological Process (BP)	Molecular Function (MF)	Cellular Component (%)
Train	62%	27%	10%
Test	60%	28%	11%
All	63%	26%	10%

2 Evaluation Metrics

Recall at rank *n*: Recall at rank n measures the exact recall which denotes the proportion of GO terms correctly predicted by the model's top n predictions [2]. Let T denote the set of true annotations of a test article and P_n denote the top n predictions of the model, then Recall at rank n is given by:

$$R_n(T, P_n) = \frac{|T \cap P_n|}{|T|} \tag{1}$$

Mean Reciprocal Rank: Mean reciprocal rank was used for evaluation in TREC question answering track [3]. The reciprocal rank is calculated as the reciprocal (multiplicative inverse) of the rank of the first correctly predicted GO term. For a query Q, let the set of true annotations and top n predictions are given by T and P_n, the reciprocal rank RR_n for top n predictions is given by:

$$RR_n(T, P_n) = \frac{1}{min(i \mid P_i \in T;\ i \in \{1, 2, \ldots, n\})} \tag{2}$$

*Supported by the Defense Advanced Research Projects Agency (DARPA) through Cooperative Agreement D20AC00002 awarded by the U.S. Department of the Interior, Interior Business Center. The content of the article does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Then, Mean Reciprocal Rank (MRR_n) for top n predictions is computed as the arithmetic mean of the reciprocal ranks.

Hierarchical measures at rank *n*: Hierarchical measures of precision, recall and F-scores were used at BioCreAtIve IV competition [1]. The hierarchical measures are more applicable for evaluating on ontologies because they make use of the GO hierarchy in the computation. Let $\mathcal{A}(X)$ be a function that returns the set of ancestors of a GO term X, and T and P_k be the set of true and predicted GO annotations, then hierarchical recall at rank n is given by:

$$hR_n(T, P_n) = \frac{|\mathcal{F}(T) \cap \mathcal{F}(P_n)|}{|\mathcal{F}(P_n)|} \tag{3}$$

where,

$$\mathcal{F}(T) = \bigcup_{t \in T} \mathcal{A}(t) \tag{4}$$

and

$$\mathcal{F}(P_n) = \bigcup_{p \in P_n} \mathcal{A}(p) \tag{5}$$

Similarly, hierarchical precision at rank n is defined as:

$$hP_n(T, P_n) = \frac{|\mathcal{F}(T) \cap \mathcal{F}(P_n)|}{|\mathcal{F}(T)|} \tag{6}$$

and the hierarchical F-score is:

$$hF_n = 2 \cdot \frac{hP_n \cdot hR_n}{hP_n + hR_n} \tag{7}$$

3 Additional Results

Table 2. Performance comparison of Word2Vec (Doc + Topic) in three GO categories

GO Category	R_{10}	TREC	BioCreAtIve		
		MRR_{10}	hP_{10}	hR_{10}	hF_{10}
Biological Process (BP)	0.32	0.38	0.29	0.59	0.36
Molecular Function (MF)	**0.61**	0.42	0.23	**0.72**	0.29
Cellular Component (CC)	0.37	**0.49**	**0.30**	0.62	**0.36**

References

1. Arighi, C., Cohen, K., Hirschman, L., Lu, Z., Tudor, C., Wiegers, T., Wilbur, W., Wu, C.: Proceedings of the fourth biocreative challenge evaluation workshop (2013)

2. Lena, P.D., Domeniconi, G., Margara, L., Moro, G.: Gota: Go term annotation of biomedical literature. BMC bioinformatics **16**, 1–13 (2015)
3. Voorhees, E.M., Buckland, L.: Overview of the trec 2003 question answering track. In: TREC. vol. 2003, pp. 54–68 (2003)