Comprehensive Benchmark of Gene Ontology Concept Recognition tools

Christoph Jacob, Philippe Thomas, Ulf Leser ({thomas,leser}@informatik.hu-berlin.de)

Knowledge Management in Bioinformatics, Institute for Computer Science,
Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Abstract

The Gene Ontology has evolved as the de facto standard for describing gene function in the biomedical domain. Information about gene function can be often found in written articles. In this work we evaluate three tools capable of recognizing Gene Ontology concepts in text on an automatically generated gold standard of 88,573 articles. The analysis reveals differences in concept recognition for these tools. An ensemble based approach is implemented to exploit idiosyncrasies between different tools and substantially improves recognition quality.

Introduction

In the biomedical domain, the Gene Ontology (GO) has evolved as the de facto standard for providing a controlled and structured vocabulary of terms describing attributes of genes. The Gene Ontology is used for the process of gene function annotation. GO annotation involves two tasks: identifying genes and gene functions in free text and associating both. Gene function annotations are collected and stored by the Gene Ontology Annotation (GOA) project (Camon et al., 2004).

In the past years, the amount of available biomedical literature has been growing at an estimated double-exponential pace (Hunter and Cohen 2006), rendering manual curation of all relevant publications as too time consuming. Therefore, concept recognition tools like mgrep (Dai et al., 2008) and MetaMap (Aronson, 2001) have been proposed. Shah et al. (2009) compared mgrep and MetaMap in terms of runtime and evaluated recognition of terms from the biological process GO-branch on a set of 2,827 PubMed abstracts. Both tools performed equally well in terms of precision. Due to the lack of a gold standard, recall has not been evaluated.

The first evaluation of different GO recognition tools has been performed in the first BioCreative competition (Blaschke et al., 2005). Evaluation proved to be difficult, as curation of predictions was too time consuming. The recently published CRAFT corpus (Bada et al., 2012) provides annotations for nine different concept types for 67 full text articles.

However, it is currently unclear what performance can be expected when evaluating concept recognition tools in a real world scenario. Some terms might not be relevant for GO-curators as corpora are annotated with respect to specific guidelines. To overcome this, we provide a comprehensive benchmark of three different concept recognition tools in the context of GO terms. In difference to previous approaches, the tools are evaluated using a large automatically generated corpus, which we believe is a valuable complementary approach to evaluating on manually annotated corpora, since it better resembles the real annotation process. Furthermore, this corpus reflects the curation approach more closely, as only relevant GO terms are annotated (e.g. the GO term *cell* may appear literally in almost every PubMed article, but is rarely used for annotation purposes).

Materials and Methods

Gold Standard(s)

In this work, we derive a gold standard corpus by exploiting existing annotations from the GOA database. The GOA project collects and integrates annotations from various sources and thus provides the most comprehensive collection of GO-annotations. GOA entries consist of three elements. The gene of interest, the GO term, and the supporting PubMed article. In this work we simplify this data to distinct binary tuples consisting of GO term and PubMed-ID. This allows us to evaluate GO concept recognition independent of errors in gene name recognition and subsequent relationship extraction. This strategy allows us to generate a corpus which is about two orders of magnitude larger than those used in previous works.

MetaMap

MetaMap is a general purpose tool for concept recognition. It currently recognizes concepts contained in the Unified Medical Language System (UMLS) Metathesaurus. For this work, we used a local version of MetaMap 2010 using the UMLS2010AB dataset. We selected parameters which seemed to be appropriate for the task of GO term recognition in order to create a realistic setting environment. In particular, we used the strict matching model and activated word sense disambiguation.

mgrep

The other tool considered in our work is mgrep. The basic idea of mgrep is to search dictionary terms in a supplied text passage. It is left to the user to define a proper dictionary which incorporates synonyms and lexical variations for each concept. This means, that the quality of the results is primarily determined by the quality of the dictionary – mgrep only ensures that the input text is searched in an efficient manner by implementing a radix tree based string search algorithm.

The current workflow for generating a basic dictionary as suggested by the authors consists of the following steps: GO terms and synonyms are extracted from the respective OBO file. Subsequently, lexical variations of the extracted terms are built using NCBI's lvg tool. This allows for the detection of small lexical variations.

t4rgot

We propose an additional method which we further refer to as "tool for recognition of GO terms" (t4rgot). In contrast to classical approaches to the problem of term searching, which are primarily dictionary based, the functionality of t4rgot is inspired by information retrieval techniques.

A schematic representation of t4rgot is shown in Figure 1. The tool consists of 2 components: the Indexer and the Recognizer. As a preprocessing step, the Indexer builds a bag of words (BOW) representation for each GO-term separately. Each bag contains all terms associated with one GO-term. The words in the bag are not stored in any particular order. Thus, positional information is given up for the purpose of efficiently dealing with variations in word order during the search. To alleviate this issue, bigrams are added to the BoWs and a score is assigned to each word/bigram in a bag. In our setting, we score bigrams 10 times as high as single words. Finally, the Indexer stores all BoWs in an index (2) which can later be processed and used by the Recognizer (3).

The Recognizer searches for GO terms in a provided input text (3). The search features the same preprocessing steps used by the Indexer. For each sentence, a set of GO-candidates is derived by matching the BOWs with words from the sentence. These candidates are then scored using the cosine similarity measure to determine which GO terms best describe the contents of the sentence. Depending on the use case, a variable number of highest scoring terms

for each sentence or article may be returned to the user. In addition, a threshold based on the determined score may be introduced to further filter the results (4).

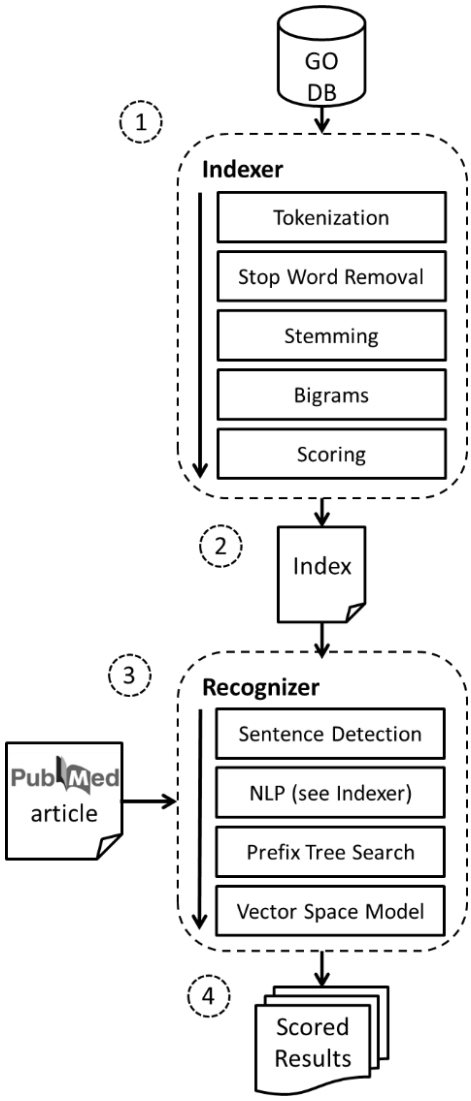

Figure 1: Schematic diagram of the t4rgot components

Results

Gold Standard corpus

The generated gold standard corpus contains 248,847 distinct GO-ID, PubMed-ID tuples collected from 88,573 PubMed articles. For 3,683 articles the full text version was available in the PubMed-Central open access subset. The remaining 84,890 articles could only be retrieved as abstracts. Approximately 18,500 of all 33,311 GO terms (55.5%) can be found in current GOA-annotations and only 5,744 GO terms are contained in the available full texts. A ratio of tuples per article of 4.12 per full text article and 2.75 per abstract can be observed. These differences can be partially explained by two observations. First, for both sets a strong positive correlation between publication year and number of annotations is observed. Second, articles in PMC tend to be more recent (1992) than articles where no full-text is found (1979). Detailed information is shown in Table 1.

Table 1: Contents of the gold standard

	PubMed articles	GO term count	Tuple count	Tuples per article
Overall	90,700	18,523	252,578	2.78
Evaluation	88,573	18,500	248,847	2.81
Full text	3,683	5,744	15,159	4.12
Abstracts	84,890	18,105	233,688	2.75

Index size and runtime

We first investigated time and hard drive requirements to set up the three tools. The results are shown in Table 2. In case of MetaMap, determining the size of the index is not trivial. MetaMap uses a proprietary dataset derived from the UMLS 2010AB release and GO constitutes only a portion of this dataset. The entire UMLS dataset is about 3.6 gigabytes large. The dictionary used by mgrep is by far the largest compared to the other two tools. It occupies about 183 gigabytes of disk space which is remarkably high considering the fact that all of this data is derived from only 33,311 terms of the Gene Ontology. The smallest index of all three tools is used by t4rgot. Only about 20 megabytes of disk space is needed to store the BOW information. The main reason for the dictionary size for mgrep is, that lexical word variations and variations in word order are stored separately. In difference, the BOW approach of t4rgot saves every token exactly once and implicitly handles variations in word order.

Runtime was assessed on a 2.25 GHz machine with 60 gigabytes of main memory. The longest execution time can be observed for MetaMap with 2 days for all 84,890 articles on a single core. In contrast, execution time for mgrep and t4rgot is much lower for both abstracts and full texts. T4rgot is capable of recognizing GO terms in all abstracts from PubMed (21 million) within 18 days on a single core.

Table 2: Index sizes and runtimes of t4rgot, mgrep and MetaMap

	t4rgot	mgrep	MetaMap
Index size	20.34 MB	187,874.84 MB	3,665.79 MB
Abstract	41 min	104 min	2,979 min
Full text	48 min	109 min	4,308 min

Precision, recall and F measure

Concept recognition results for all three tools on abstracts are shown in Table 3.

Table 3: Comparison of results from t4rgot, mgrep and MetaMap on abstracts

	t4rgot	mgrep	MetaMap
Count	843,959	843,469	663,612
True Pos.	36,456	40,655	37,634
False Pos.	807,503	802,814	625,978
False Neg.	197,232	193,033	193,402
Precision	4.32%	4.82%	5.67%
Recall	15.60%	17.40%	16.29%
F measure	6.77%	7.55%	8.41%

The highest recall on abstracts is achieved by mgrep. The tool correctly identifies more than 40,000 GO terms, leading to a recall of 17.40%. Of the 2.75 tuples which are on average contained in each abstract, mgrep is capable of correctly finding 0.48 tuples. MetaMap finds fewer true positives but also returns overall fewer terms, achieving the highest precision of 5.67% and subsequently the highest F-measure with 8.4%. T4rgot is not able to achieve the results of the other two tools for abstracts, as precision, recall and F measure fall about 2 percentage points (pp) short of the respective best result.

Table 4 shows the results for the three different tools when applied to full-text articles.

Table 4: Comparison of results from t4rgot, mgrep and MetaMap on full texts

	t4rgot	mgrep	MetaMap
Count	220,980	238,916	213,242
True Pos.	4,688	4,551	4,233
False Pos.	216,292	234,365	209,009
False Neg.	10,471	10,608	10,713
Precision	2.12%	1.90%	1.99%
Recall	30.93%	30.02%	28.32%
F measure	3.97%	3.58%	3.71%

Different observations can be made for full texts in comparison to abstracts. Here, t4rgot leads all tools in all three measures and achieves the highest rates of precision (2.12%), recall (30.93%) and F measure (3.97%). The tool extracts 4,688 correct GO terms. The lowest precision of 1.90% is achieved by mgrep. MetaMap returns the fewest amounts of tuples and true positives which results in the lowest recall of 28.32%.

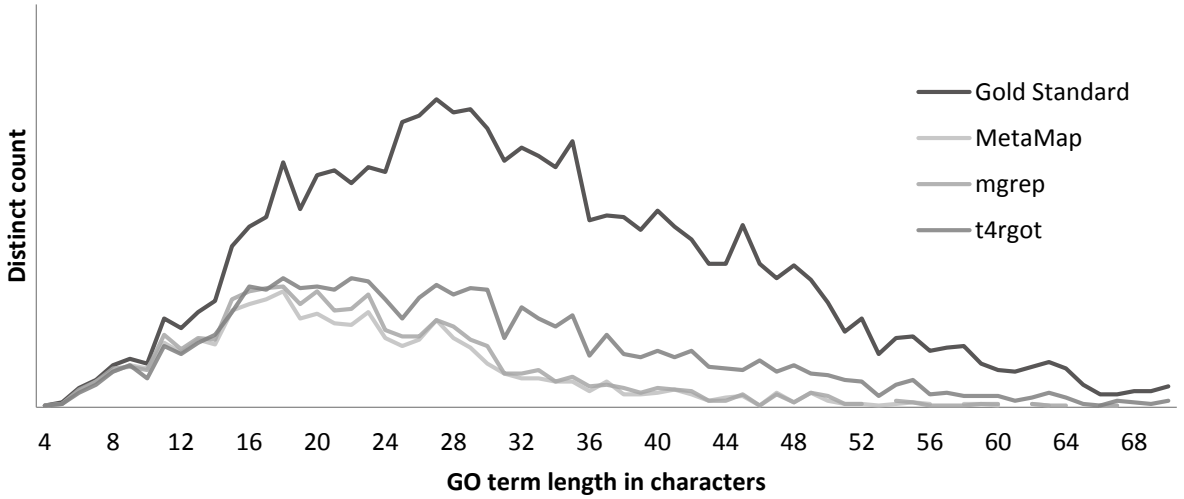

Figure 2: Length distribution of distinct GO terms of the gold standard and distinct GO terms contained in true positives returned by each tool for full texts

It is also remarkable that the observed recall rates for full texts are generally twice as high as for abstracts. This is probably caused by the fact that abstracts only constitute a small portion of the entire article. During the process of manual annotation, curators typically use the entire full text to derive appropriate GO terms. It can be suspected that in many cases these GO terms could not be extracted from the abstract because the text passage containing the term was missing.

Finally, the small variation in the number of true positives returned by all three tools raises the question whether all tools find approximately the same terms or if the tools find entirely different terms. This question is further investigated in the following subsection, by analyzing the true positives returned by all three tools within full text articles in more detail.

We first analyzed the distribution of distinctly found GO terms in comparison to their character length. The results are depicted in Figure 2. GO terms from the gold standard are displayed by the dark grey curve. Ideally, a tool should follow this curve. All three tools correctly identify a large portion of short terms consisting of fewer than 20 characters. The curves of MetaMap and mgrep look very similar and steadily decline for GO terms with more than 28 characters. In contrast, t4rgot seems to better correlate with the profile of the gold standard. Indeed, we observe a significant correlation between gold standard and t4rgot predictions (Kendall's tau = 0.75; p-value < 0.01). For the other two tools we observe significant correlations of 0.55 in comparison to the gold standard.

T4rgot is apparently able to identify more of the longer GO terms compared to MetaMap and mgrep which is an indication that the approach of t4rgot works better for identifying longer GO terms. This suspicion is also confirmed by comparing the average lengths of all GO terms contained in all true positives for each tool. While MetaMap and mgrep extract terms with an average length of 18 characters, the terms found by t4rgot contain on average 24 characters which is much closer to the average length of a 29 characters in the gold standard. An analysis of missed GO terms revealed an average in length of 34 characters which supports the observation that longer GO terms are harder to identify.

The average depth of distinct GO terms returned by t4got is 4.4 and 4.0 for mgrep and MetaMap. This difference is significant according to the Wilcoxon signed-rank test. In comparison, the average node depth in the gold standard is 5.4. For all three tools we observe that terms from the sub-ontology *cellular component* achieve with 4.9% the highest F1 score, followed by *biological process* with 2.9% and *molecular function* with 2.3%. Similar results have been reported for the BioCreative competition where cellular component terms had the highest fraction of correct predictions.

Ensemble systems

The previous analysis of true positive predictions shows, that t4rgot produces different results in comparison to MetaMap and mgrep. This suggests that a hybrid system, consisting of two complementing systems, might lead to superior performance. To test this hypothesis we generated a hybrid system by combining t4rgot with mgrep. We evaluated two different combinations: first, to increase recall, we built the union between predictions of the two tools. Second, to increase precision, we built the intersection. Experiments are restricted to t4rgot and mgrep, as ensembles using MetaMap produce highly similar results (data not shown). Results of the two different ensembles are shown in Table 5.

Table 5: Results for the ensemble system of t4rgot and mgrep on abstracts and full texts

	Union t4rgot & mgrep		Intersection t4rgot & mgrep	
	abstract	full text	abstract	full text
Count	1,546,285	411,841	141,144	48,055
True Pos.	58,733	6,431	18,379	2,808
False Pos.	1,487,552	405,410	122,765	45,247
False Neg.	174,955	8,728	215,259	12,351
Precision	3.80%	1.56%	13.02%	5.84%
Recall	25.13%	42.42%	7.87%	18.52%
F measure	6.60%	3.01%	9.81%	8.88%

As expected, the union between t4rgot and mgrep achieves very high recall rates of 25.13% for abstracts and 42.42% for full texts. In comparison to the individual results, recall increases up to 12 pp. The ensemble system using the intersection between the two tools, achieves superior precision rates of 13.0% and 5.8% for abstracts and full texts respectively. While this system may not return a large amount of GO terms for each article, it may be more suitable for the purpose of manual annotation for two reasons. First, the system processes a full text article within two or three seconds and could thus be used in a real-time application environment. Second, curators usually use only one or two GO terms during the process of annotation. If the system were to return high precision results, the curator could quickly choose between 10 to 15 proposed GO terms for each article without having to go through a long list of possible candidates.

In conclusion, the proposed ensemble system may be of relevance for the current annotation efforts undertaken by the GOA project. While the individual systems each suffer from various problems, their combination is able to return results of higher quality. The results obtained from t4rgot present a valuable addition to the results of mgrep and can be used to construct a system with either high recall or high precision, depending on the use case. The intersection of the results of both tools leads to an increase in precision of 8.2 percentage points for abstracts and 3.72 percentage points for full texts compared to the individual systems. In contrast, the union of both results is able to increase recall by 7.73 percentage points for abstracts and 11.49 percentage points for full texts compared to the highest recall obtained by either t4rgot or mgrep.

Discussion

A drawback of the automatically generated gold standard is that not all GO terms have been annotated. In fact, only a very small fraction of the contained terms may have been used by curators. The reason is that curators follow specific guidelines when selecting terms relevant for curation. Therefore, unspecific top level terms such as *cell* are very rarely used as curators tend to annotate the most specific term. However, this is also one of the biggest advantages as this corpus reflects actual annotation behavior. We assume that all tools would achieve better results on a corpus manually annotated for all GO concepts. However, this corpus would not represent curator's behavior, who are only interested in informative terms. We therefore believe that such a corpus is a realistic scenario to evaluate real world capabilities of GO concept recognition tools.

In summary, all tools achieve very low rates of precision. For abstracts, precision ranges between 4% and 6% while for full texts all tools achieve a precision of about 2%. Considering the results achieved by participants of the BioCreative contest (Blaschke et al., 2005) (20% to 80% precision) or results from previous comparisons of mgrep and MetaMap (Shah et al., 2009) (above 70% precision), the figures obtained here clearly contradict these results. However, evaluation in these cases was conducted manually, achieving almost complete coverage of all GO terms contained in the used corpora of text.

Conclusion

In this publication we present the first large scale evaluation of concept recognition tools. The collected large-scale gold standard reflects the curators' needs better than a small manually annotated corpus. We also provide a detailed evaluation of three different tools and conclude that the combination of t4rgot and mgrep provide better results than a single system.

Acknowledgements

The authors would like to thank Fan Meng and Manhong Dai, the creators of mgrep from the Molecular and Neuroscience Institute at the University of Michigan for their support for the evaluation of mgrep and development of t4rgot. We also would like to thank the anonymous reviewers and the TAIR database curators for their helpful comments on the manuscript.

References

Aronson AR: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.** *Proceedings of the AMIA Symposium* 2001, 17-21.

Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, Baumgartner WA, Cohen KB, Verspoor K, Blake JA, Hunter LE: **Concept Annotation in the CRAFT Corpus.** *BMC Bioinformatics* 2012, 13:161.

Blaschke C, Krallinger M, Leon EA, Valencia A: **Evaluation of Bio-CreAtIvE assessment of task 2.** *BMC Bioinformatics* 2005, 6(Suppl 1):16.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Research* 2004, 32(Database issue):262-266.

Dai M, Shah NH, Xuan W, Musen MA, Watson SJ, Athey B, Meng F: **An Efficient Solution for Mapping Free Text to Ontology Terms.** *AMIA Summit on Translational Bioinformatics San Francisco* 2008.

Hunter L, Cohen KB: **Biomedical language processing: what's beyond PubMed?.** *Molecular Cell* 2006, 21(5):589-594.

Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA: **Comparison of concept recognizers for building the open biomedical annotator.** *BMC Bioinformatics* 2009, 10(Suppl 9):14.