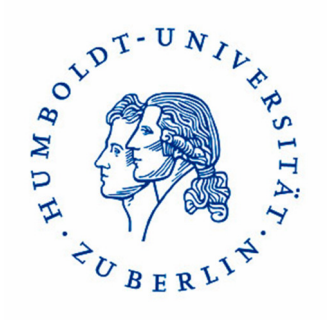

Exposé for the diploma thesis

Gene Ontology concept recognition

Christoph Jacob

Supervising tutors: Prof. Dr. Ulf Leser, Philippe Thomas

May 2, 2011

1. Motivation/Introduction

In the past years the increasing usage of electronic publication has led to an enormous growth of databases and repositories containing articles relevant to scientifically different domains. Of these databases, the electronically available database PubMed [Putnam, 1998] has emerged as the most relevant one for the biomedical community. As of today, PubMed consists of more than 20 million abstracts [NLM Systems, 2011] and about 2 million full-text articles (accessible via PubMed Central [Vastag, 2000]) containing a huge amount of information on various biomedical topics. While PubMed is freely available, its size and rapid growth and the unstructured nature of text written in natural language call for tools and methods that assist in the process of information extraction in order to become more accessible and usable. Ontologies provide controlled vocabularies which may assist this process if a text passage can be mapped to an ontology concept.

In the biomedical domain the Gene Ontology (GO) [The Gene Ontology Consortium, 2004] has evolved as the de facto standard for providing a controlled and structured vocabulary of terms describing attributes of genes and gene products. The GO is organized as a directed acyclic graph and can be divided into three sub-ontologies: cellular component, molecular function and biological process, overall containing approximately 34.000 so-called GO terms.

The GO is being used for the process of gene function annotation. GO annotation involves two tasks: identifying genes/gene products and gene functions in free text and the process of associating both using GO terms. This process is carried out to maintain the Gene Ontology Annotation (GOA) database [Camon et al., 2004] which contains associations of genes and gene products to GO terms that also reference the PubMed article supporting the annotation. Typically, GO annotation is performed via manual curation by a domain expert, the so-called curator, who has to read the entire article to extract relevant genes/gene products and their functional description. He then has to map this information to a pair of gene-identifier/GO term. While it has been shown that the first task of recognizing genes and gene products can be carried out with the help of high-precision tools [Hirschman et al., 2005; Morgan et al., 2008], the latter task of identifying GO terms in free text has yet to be solved in a satisfying manner and will be subject of this thesis.

Since GO terms are concepts organized in an ontology, the problem of identifying GO terms can be described as the problem of concept recognition. While there have been efforts in developing concept recognition tools for GO terms in the past, most of these tools lack a detailed evaluation. Evaluation is often carried out on very small corpora of text, usually manually annotated. Due to the size of the GO, a large-scale evaluation on a realistically sized corpus would be appropriate. A thorough evaluation is a requirement to keep pace with the constant growth of PubMed and the dynamic nature of the GO.

2. Related Work

Several solutions exist for the problem of finding GO terms or ontology concepts in general in full text, of which the most relevant ones can be split into two major groups: participants of the first BioCreative challenge and tools for building the Open Biomedical Annotator (OBA).

2.1. BioCreative

First, approaches and tools developed in the course of the second task of the first BioCreative challenge in 2004 should be considered. Hirschman et al. provide an overview of the challenge in [Hirschman et al., 2005], describing the first task of gene and protein name extraction and mapping to a unique identifier and the second task of finding text passages which support a given GO annotation. Both tasks were the result of examining the process of gene function annotation along the "curation pipeline". Results of the challenge show that while the best systems in the gene normalization task achieved precision and recall values of 0.8, results for the annotation task were significantly lower. A detailed evaluation [Blaschke et al., 2005] revealed three different groups of approaches taken by the eight participants of task 2.

The first group of approaches can be characterized by various textual matching techniques which involved mainly the GO terms themselves and their definition as well as words derived from full text supporting existing annotations. This approach, adopted by five participants, yielded the best results in terms of the absolute number of true positives with a precision of the best methods of 0.29. The most promising results were those of [Krallinger et al., 2005], [Couto et al., 2005] and [Verspoor et al., 2005]. Krallinger et al. adopted a sentence-sliding window approach using constructed sub-tags for each GO term as search terms within a fixed window of 4 sentences. For their tool FiGo, Couto et al. used the information content for each word contained in or related to the GO term and scored sentence passages accordingly. Verspoor et al. constructed word proximity networks using co-occurrence matrices for words in documents and representative words from GO terms returning a single sentence. Co-occurrence was implemented by using words which frequently occurred together with the GO term in question in free text, assuming that these words could be used to further describe the GO term itself.

The second and third groups adopt machine learning and hybrid approaches which were developed by three participants. While the hybrid approach [Chiang et al., 2004] scored the highest precision of 0.8 the focus on precision resulted in a very low absolute number of predictions. Machine learning approaches yielded lower results than textual matching approaches. In their evaluation Blaschke et al. conclude, that the lack of a decent-sized, high quality training set constituted a problem for these approaches.

Camon et al. drew further conclusions from the perspective of the GOA team [Camon et al., 2005]. Common mistakes were identified and analyzed and suggestions for improving the performance of GO term recognition were given. For example, they suggest disregarding certain parts of the GO (e.g. GO terms marked as obsolete) as well as certain sections of full text articles as adopted during manual curation. Furthermore, since some participants submitted entire paragraphs, evidence passages should be limited to a maximum number of five lines to better reflect manual curation. Despite the fact that the results were evaluated by expert curators from

GOA, an Inter-Annotator Agreement showed only an overlap of 39% in the assessment of the submissions. In addition, the relatively small training and test sets flawed the evaluation.

The results of BioCreative have contributed to the development of a tool for finding GO terms in biomedical articles by Damyanliev. In his diploma thesis [Damyanliev, 2010] he describes his approach of constructing a bag-of-words (BOW) for each GO term consisting of the words the GO term comprises of and supplementary words derived from parent- and child-terms. He then scores a sliding window with a fixed size of three sentences by searching for words from the BOW within the window. His findings include that using child-terms as well as bigrams during BOW-construction improves the recognition of GO terms and stemming produces slightly worse results. Since evaluation was carried out on the results of BioCreative, values for precision and recall values could only be estimated (about 0.7 each). Damyanliev concludes that with few enhancements such as implementing co-occurrence as in [Verspoor et al., 2005], his approach could yield a promising direction in solving the problem of recognizing GO terms in full text.

2.2. Open Biomedical Annotator

Second, approaches revolving around the development of the Open Biomedical Annotator (OBA), which is being maintained by the National Center for Biomedical Ontology (NCBO), deserve further investigation. OBA has been developed to assist in the process of annotation [Jonquet et al., 2009] by presenting a Web service which takes a full text as input and finds concepts from the Unified Medical Language System (UMLS) Metathesaurus therein. The UMLS Metathesaurus contains the entire GO; therefore the OBA can be used for the purpose of finding GO terms in full text. The OBA derives annotation in two phases: the first phase uses a concept recognition tool to create direct annotations which are then expanded by multiple components.

During the development of the OBA, two concept recognition tools have been tested and undergone a comparative assessment [Shah et al, 2009]. They compared mgrep [Dai et al., 2008], and MetaMap [Aaronson, 2001], and found that mgrep served the purpose of the OBA better, mainly because of its speed. Part of the comparison was carried out on 2800 PubMed abstracts which were annotated to the Biological Process part of the GO. Both tools showed fair precision values of 0.77 (mgrep) and 0.76 (MetaMap). Due to the fact that evaluation was carried out manually and thus no gold standard existed, recall could not be determined. A textual matching approach using dictionaries to find and map concepts was adopted by both tools. While mgrep is actively being used by the OBA and MetaMap remains a state-of-the art tool for "general purpose" concept recognition, both tools have yet to undergo a more thorough evaluation.

The successful integration of the concept recognition tool mgrep into a Web service which has been integrated into the UIMA tool chain and workflow [Roeder et al., 2010], shows the high relevance of tools and methods to identify concepts such as GO terms in free text for annotation purposes.

2.3. GNAT

Since this thesis will focus in the second part of the annotation procedure, recognizing GO terms in free text, the first part, finding genes and gene products, will be performed using the tool GNAT [Hakenberg et al., 2008; Solt et al., 2010]. GNAT is an inter-species gene normalization (ISGN) tool capable of finding and mapping gene occurrences to species-specific Gene IDs from Entrez Gene [Maglott et al., 2005]. It therefore overcomes problems such as synonymy and ambiguity of gene mentions. An evaluation on benchmarking data from BioCreative 1 and 2 has shown that GNAT can perform the ISGN with a precision of 0.91 and a recall of 0.74. GNAT will be used in the course of this thesis for evaluation purposes.

3. Goals

The analysis of related work and recent developments in the field of concept recognition leads to the conclusion that due to the lack of decent-sized training and test sets, a dictionary based, textual matching approach should be given precedence over machine learning approaches. This decision is also strengthened by the speed advantage of textual matching implementations which will be necessary when working on entire PubMed. Furthermore, many tools capable of finding GO terms lack a sophisticated evaluation which impedes conclusions towards which methods truly help in GO concept recognition and to what extent.

Therefore, this thesis serves two purposes. First, it aims at developing a fast concept recognition tool for GO terms which can be used to find text passages resembling a GO term. The concept recognizer will adopt a BOW approach similar to the one developed in [Damyanliev, 2010]. This task can also be described by the following sentence: "Given a certain article, find every contained GO term and return a passage as evidence".

Second, instead of using a small corpus of annotated text as the gold standard, this concept recognizer will be evaluated on a gold standard derived from currently existing annotations in the GOA database, to ensure its effectiveness with respect to prevailing curation standards. This will allow for a large scale evaluation, however, since these annotations do not contain evidence passages this evaluation might not be as accurate as manual one. In addition, evaluation will feature a comparison to the state-of-the-art concept recognition tools mgrep and MetaMap which will also be evaluated on the derived gold standard.

The first task mainly requires the integration and usage of different existing methods. The focus of this thesis will be on evaluation in order to assess which methods and techniques are best suited for a fast and accurate recognition of GO terms and which of the existing GO concept recognition tools performs best in solving that problem.

4. Approach

As already mentioned, the concept recognizer developed in this thesis will adopt an approach, which involves constructing BOWs for each GO term. Each BOW will consist of words from the GO term itself and words which have a high co-occurrence with the GO term in question. These words will be derived from existing annotations. Furthermore, words from child GO terms should be considered as well for constructing the BOW. Each of the words in the BOW will be assigned a score which represents the evidence strength. The score will be low for words occurring frequently in other GO terms and high for very distinctive terms and implemented using the tf-idf measure.

A sentence sliding window approach will then be used to score text passages by searching for words of the BOWs within the window. Variable sized windows will account for differences in the size of BOWs. A small BOW will contain only few words; hence the search for these words should be broadened by extending the window size in order to gain a decent score at all. On the other hand, a large BOW might result in many hits within the window; thus the search can be further narrowed down to a few or even one sentence within the window. The calculated score should respect the evidence content of the words in the BOWs as well as proximity of words within the window.

Additionally, standard techniques such as stop word removal and stemming will be evaluated and applied where necessary. In the case of stemming, a differentiated usage is required since it might be counterproductive in some cases such as during the stemming of the two GO terms "protein transport" and "protein transporter" [Chiang et al., 2003].

Since a comparison between this approach and the two concept recognition tools mgrep and MetaMap is to be conducted, it will be necessary to use both tools in the same manner in order to

derive mappings from text passages to GO terms. mgrep can currently only be accessed online via the OBA Web service, while MetaMap can either be accessed online via a Web service or used on a local installation. Depending on the performance, one of these methods will be used. Both tools require different input formats and produce structurally different output data, which requires pre- and post-processing steps to use the tools and to make the results comparable.

5. Evaluation

Due to the fact that no large corpus of annotated abstracts or full texts containing evidence passages exists, this thesis will generate a gold standard based on existing annotations from GOA to conduct a large scale evaluation on all PubMed articles referenced by GOA. Each annotation contains a triplet of PubMed ID/Entrez GeneID/GO ID. The set of all triplets constitutes the gold standard used in this thesis. The goal of the concept recognition tools is to derive these triplets.

Each tool returns a pair of PubMed ID/GO ID together with the evidence passage. This information will then be combined with each pair of PubMed ID/Entrez GeneID from the results of gene normalization using GNAT. The so-constructed triplet can then be checked against the gold standard. If the gold standard includes the found triplet, it is evaluated as a true positive; else it is evaluated as a false positive. GOA-triplets which are not found will be regarded as false negatives. Standard Information Retrieval measures such as Precision, Recall and F-measure can then be calculated to allow for a comparison between the three concept recognition tools.

Evaluation can then be conducted in further detail. The results can be analyzed structurally (e.g. length of GO terms recognized or location of recognized terms within the GO) as well as semantically (e.g. are there differences in recognizing GO terms for certain species such as Arabidopsis or Drosophila). Since some of the PubMed articles are available in full text, the difference between recognizing GO terms in abstracts versus full texts could be analyzed as well.

Literature

Aronson AR: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.** *Proc AMIA Symp* 2001, 17–21.

Blaschke C, Krallinger M, Leon EA, Valencia A: **Evaluation of Bio-CreAtIvE assessment of task 2.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S16.

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 Res* 2004, 32(Database issue):262-266.

Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Mslen J, Binns D, Apweiler R: **An evaluation of GO annotation retrieval for BioCreAtIvE and GOA.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S17.

Chiang JH, Yu HC: **Extracting Functional Annotations of Proteins Based on Hybrid Text Mining Approaches.** *Proc BioCreAtIvE Challenge Evaluation Workshop* 2004.

Chiang J, Yu HC: **MeKE: discovering the functions of gene products from biomedical literature via sentence alignment.** *Bioinformatics* 2003, **19**:1417-1422.

Couto FM, Silva MJ, Coutinho PM: **Finding genomic ontology terms in text using evidence content.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S21.

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.

Damyanliev N: **Finding Gene Ontology Terms in Biomedical Articles.** 2010

Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization of gene mentions with GNAT.** *Bioinformatics* 2008, **24**(16):i126-i132.

Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S1.

Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreAtIvE task 1B: normalized gene lists.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S11.

Jonquet C, Shah NH and Musen MA: **The Open Biomedical Annotator.** *AMIA Summit on Translational Bioinformatics San Francisco* 2009.

Krallinger M, Padron M, Valencia A: **A sentence sliding window approach to extract protein annotations from biomedical articles.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S19.

Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids* 2005, **33**(Database issue):D54-D58.

Morgan A, et al.: **Overview of BioCreative II Gene Normalization.** *Genome Biology* 2008, **9** (Suppl 2):S3.

NLM Systems: **Data, News and Update Information. PubMed Update.** Internet: http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update [18.04.2011].

Putnam NC: **Searching MEDLINE free on the Internet using the National Library of Medicine's PubMed.** *Clin Excell Nurse Pract* 1998, **2**(5):314-316.

Roeder C, Jonquet C, Shah NH, Baumgartner WA Jr, Verspoor K, Hunter L: **A UIMA wrapper for the NCBO annotator.** *Bioinformatics* 2010, **26**(14):1800-1801.

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):S14.

Solt I, Gerner M, Thomas P, Nenadic G, Bergman CM, Leser U, Hakenberg J: **Gene mention normalization in full texts using GNAT and LINNAEUS.** *Proceedings of the BioCreative III Workshop* 2010, 143-148.

The Gene Ontology Consortium: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, 32(Database issue):D258-D261.

Vastag B: **NIH launches PubMed Central.** *J Natl Cancer Inst* 2000, **92**(5):374.

Verspoor K, Cohn J, Joslyn C, Mniszewski S, Rechtsteiner A, Rocha LM, Simas T: **Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks.** *BMC Bioinformatics* 2005, **6**(Suppl 1):S20.