Answering Gene Ontology terms to proteomics questions by supervised macro reading in Medline

Julien Gobeill[1]✉, Emilie Pasche[2], Douglas Teodoro[2], Anne-Lise Veuthey[3], Patrick Ruch[2]

[1]University of Applied Sciences, Information Sciences, Genevay
[2]Hospitals and University of Geneva, Geneva
[3]Swiss-Prot group, Swiss Institute of Bioinformatics, Geneva

Motivation and Objectives

Biomedical professionals have at their disposal a huge amount of literature. But when they have a precise question, they often have to deal with too many documents to efficiently find the appropriate answers in a reasonable time. Faced to this literature overload, the need for automatic assistance has been largely pointed out, and PubMed is argued to be only the beginning on how scientists use the biomedical literature (Hunter and Cohen, 2006).

Ontology-based search engines began to introduce semantics in search results. These systems still display documents, but the user visualizes clusters of PubMed results according to concepts which were extracted from the abstracts. GoPubMed (Doms and Schroeder, 2005) and EBIMed (Rebholz-Schuhmann et al, 2007) are popular examples of such ontology-based search engines in the biomedical domain. Question Answering (QA) systems are argued to be the next generation of semantic search engines (Wren, 2011). QA systems no more display documents but directly concepts which were extracted from the search results; these concepts are supposed to answer the user's question formulated in natural language. EAGLi (Gobeill et al, 2009), our locally developed system, is an example of such QA search engines.

Thus, both ontology-based and QA search engines, share the crucial task of efficiently extracting concepts from the result set, i.e. a set of documents. This task is sometimes called macro reading, in contrast with micro reading – or classification, categorization – which is a traditional Natural Language Processing task that aims at extracting concepts from a single document (Mitchell et al, 2009).

This paper focuses on macro reading of MEDLINE abstracts. Several experiments have been reported to find the best way to extract ontology terms out of a single MEDLINE abstract, i.e. micro reading. In particular, (Trieschnigg et al, 2009) compared the performances of six classification systems for reproducing the manual Medical Subject Headings (MeSH) annotation of a MEDLINE abstract. The evaluated systems included two morphosyntactic classifiers (sometimes also called thesaurus-based), which aim at literally finding ontology terms in the abstract by alignment of words, and a machine learning (or supervised) classifier, which aims at inferring the annotation from a knowledge base containing already annotated abstracts. The authors concluded that the machine learning approach outperformed the morphosyntactic ones. But the macro reading task is fundamentally different, as we look for the best way to extract then combine ontology terms from a set of MEDLINE abstracts.

The issue investigated in this paper is: to what extent the differences observed between two classifiers for a micro reading task are observed for a macro reading one? In particular, the redundancy hypothesis claims that the redundancy in large textual collections such as the Web or MEDLINE tends to smoothe performance differences across classifiers (Lin, 2007). To address this question, we compared a morphosyntactic and a machine learning classifiers for both tasks, focusing on the extraction of Gene Ontology (GO) terms, a controlled vocabulary for the characterization of proteins functions. The micro reading task consisted in extracting GO terms from a single MEDLINE abstract, as in the Trieschnigg et al's work; the macro reading task consisted in extracting GO terms from a set of MEDLINE abstracts in order to answer to proteomics questions asked to the EAGLi QA system.

Methods

We evaluated two statistical classifiers which-were both studied in the Trieschnigg et al's work. The morphosyntactic classifier was EAGL. It is described comprehensively in (Ruch, 2006). It showed very competitive results when it was compared to other state-of-the-art morphosyntactic

classifiers, as during the official BioCreative I evaluation (Blaschke et al, 2005) or in the Trieschnigg et al's work against Metamap (Aronson and Lang, 2010). The machine learning classifier was a k-NN. The k-NN is a remarkably simple and scalable algorithm which assigns to a new abstract the GO terms that are the most prevalent among the k most similar abstracts contained in a knowledge base (Manning and Schütze, 1999). The knowledge base was designed from the GOA database, which contains 85'000 manually curated abstracts and is available at http://www.ebi.ac.uk/GOA/, Last accessed on August 1st, 2012). These abstracts were indexed with a classical Information Retrieval engine (Ounis et al, 2006) and, for each input text, the k=100 most lexically similar ones were retrieved in order to infer the GO terms.

For the micro reading task, we designed a so called GOA benchmark of one thousand MEDLINE abstracts sampled from the GOA database; the classifiers were evaluated on their ability to extract the GO terms that were manually associated with these abstracts by the GOA experts. For the macro reading task, we designed two benchmarks of fifty questions by exploiting two biological databases: the Comparative Toxicogenomics Database (CTD) contains more than 2'800 chemicals annotated with GO terms, and is available at http://ctdbase.org/ (Last accessed on August 1st, 2012); the UniProt database contains millions of proteins annotated with GO terms, and is available at http://www.uniprot.org/ (Last accessed on August 1st, 2012). Questions were sampled from these databases and dealt with molecular functions and a given chemical compound, such as "what molecular functions are affected by Aminophenols ?", or cellular components and a given protein, such as "what cellular component is the location of NPHP1?". The classifiers were successively embedded in the EAGLi's QA engine for extracting GO terms from a set of one hundred MEDLINE abstracts retrieved by EAGLi for each question. The most prevalent GO terms extracted from these abstracts were then proposed as answers by the QA engine. Please refer to (Gobeill et al, 2009) for a deeper description of EAGLi. Thus, their evaluation was extrinsic and was based on their ability to extract GO terms from a set of abstracts and then provide to EAGLi the answers contained in the databases.

There were on average 2.8 GO terms per abstract to return in the GOA benchmark, and 30/1.3 GO terms per question to find (literally to answer) for respectively the CTD/UniProt benchmark. As both categorizers output a ranked list of candidate GO term, we chose metrics from the Information Retrieval domain that were well-established during the TREC campaigns (Voorhees et al, 2001). For precision considerations, we computed the Mean Reciprocal Rank (MRR) which is the multiplicative inverse of the rank of the first correct outputted GO term.

Results and Discussion

For the micro reading task (i.e. extracting GO terms from a single abstract), as in the Trieschnigg et al's work with MeSH classification, the machine learning classifier (k-NN) outperforms the morphosyntactic one (EAGL). For the macro reading task (i.e. extracting GO terms from a set of abstracts), for both benchmarks, the k-NN also outperforms EAGL, and the observed differences in top-precision are similar and consistent with the micro-reading task. These results weaken the redundancy hypothesis, as the performance of classifiers for micro reading tasks appears to be of importance for macro reading tasks.

It is worth observing that, unlike other text mining tasks, Information Retrieval and Question Answering have been largely resisting to machine learning advances (Athenikosa and Hanb, 2009). Ontology-based search engines powered with morphosyntactic classifiers could benefit from such a new component, as it allows to inject knowledge contained in curated databases in the result set. This could provide promising research pathways for the biomedical data mining community.

Beyond comparisons, our QA engine with supervised macro reading in MEDLINE achieved a top-precision ranging from 0.58 to 0.69 to answer

Table1: top-precision for both GO classifiers observed in micro reading then macro reading tasks, along with the percentage of improvement with the k-NN.

	Micro reading task	**Macro reading task**	
	GOA benchmark	**CTD benchmark**	**UniProt benchmark**
EAGL	0,23	0,34	0,33
k-NN	.48 +109%	.69 +103%	.58 +76%

proteomics questions. This performance allows its users to save time on consulting the literature, as well as to automatically produce function predictions for massive proteomics datasets, such as in (Anton et al, 2012). EAGLi is available at http://eagl.unige.ch/EAGLi/ (Last accessed on August 1st, 2012).

Acknowledgements

Work supported by the Swiss National Fund for Scientific Research [BiND project 3252B0-105755].

References

1. Anton BP, Chang YC, et al (2012) COMBREX: Design, Methodology, and Initial Results. Manuscript submitted for publication.
2. Aronson AR and Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 17(3), 229. doi:10.1136/jamia.2009.002733
3. Athenikosa S and Hanb H (2009) Biomedical question answering: A survey. Comput Methods Programs Biomed. 99(1), 1. doi:10.1016/j.cmpb.2009.10.003
4. Blaschke C, Leon EA, et al (2005) Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics 6(Suppl 1):S16. doi:10.1186/1471-2105-6-S1-S16
5. Doms A and Schroeder M (2005) GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 1(33), 783. doi:10.1093/nar/gki470
6. Gobeill J, Pasche E, et al (2009) Question answering for biology and medicine. Information Technology and Application in Biomedicine, Larnaca, Cyprus.
7. Hunter L and Cohen KB (2006) Biomedical language processing: what's beyond PubMed? Mol Cell. 21(5), 589. doi: 10.1016/j.molcel.2006.02.012
8. Lin J (2007) An exploration of the principles underlying redundancy-based factoid question answering. ACM Trans. Inf. Syst. 25(2). doi: 10.1145/1229179.1229180
9. Manning CD and Schütze H (1999) Foundations of Statistical Natural Language Processing. Cambridge, MA, MIT Press. doi:10.1023/A:1011424425034
10. Mitchell TM, Betteridge J, et al (2009) Populating the Semantic Web by Macro-reading Internet Text. Proceedings of the 8th Intern. Semantic Web Conf. doi: 10.1007/978-3-642-04930-9_66
11. Ounis I, Amati G, et al (2006) Terrier: A High Performance and Scalable Information Retrieval Platform. Proceedings of ACM SIGIR'06 Workshop.
12. Rebholz-Schuhmann D, Kirsch H, et al (2007) EBIMed--text crunching to gather facts for proteins from Medline. Bioinformatics 23(2), 237. doi:10.1093/bioinformatics/btl302
13. Ruch P (2006) Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 22(6), 658. doi:10.1093/bioinformatics/bti783
14. Trieschnigg D, Pezik P, et al (2009) MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics 25(11), 1412. doi:10.1093/bioinformatics/btp249
15. Voorhees E (2001) Overview of the QA Track. In Proceedings of the TREC-10 Conference. NIST, Gaithersburg. 2001:157-165.
16. Wren JD (2011) Question answering systems in biology and medicine--the time is now. Bioinformatics 27(14). doi:10.1093/bioinformatics/btr327