AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

Notes about the OntoGene Pipeline

Fabio Rinaldi, Simon Clematide, Gerold Schneider, Gintarė Grigonytė
Institute of Computational Linguistics, University of Zurich, Switzerland

Abstract

In this paper we describe the architecture of the OntoGene Relation mining pipeline and some of its recent applications. With this research overview paper we intend to provide a contribution towards the recently started discussion towards standards for information extraction architectures in the biomedical domain.

Our approach delivers domain entities mentioned in each input document, as well as candidate relationships, both ranked according to a confidency score computed by the system. This information is presented to the user through an advanced interface aimed at supporting the process of interactive curation.

Introduction

As a way to cope with the constantly increasing generation of results in molecular biology, some organizations maintain various types of databases that aim at collecting the most significant information in a specific area. For example, UniProt/SwissProt (UniProt Consortium 2007) collects information on all known proteins. MINT (Zanzoni et al. 2002) and IntAct (Hermjakob et al. 2004) are databases collecting protein interactions. PharmGKB (Klein et al. 2001; Sangkuhl et al. 2008) curates knowledge about the impact of genetic variation on drug response for clinicians and researchers. The Comparative Toxicogenomics Database collects interactions between chemicals and genes in order to support the study on the effects of environmental chemicals on health (Mattingly et al. 2006). Most of the information in these databases is derived from the primary literature by a process of manual revision known as "literature curation". Text mining solutions are increasingly requested to support the process of curation of biomedical databases.

The work presented here is part the OntoGene project[1], which aims at improving biomedical text mining through the usage of advanced natural language processing techniques. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, term recognition, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences (Rinaldi et al. 2006; 2008). The results of the entity detection feed directly into the process of identification of interactions. The syntactic parser (Schneider 2008) takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the entity annotation has a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

In the context of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature), the OntoGene group has developed a user-friendly interface (ODIN: OntoGene Document INspector) which presents the results of the text mining pipeline in an intuitive fashion, and allows a better interaction of the curator with the underlying text mining system (Rinaldi et al. 2012).

In the rest of this paper we describe the OntoGene pipeline architecture, the ODIN interface for assisted curation and briefly survey some of the applications.

Information Extraction

In this section we describe the OntoGene Text Mining pipeline which is used to (a) provide all basic preprocessing (e.g. tokenization) of the target documents, (b) identify all mentions of domain entities and normalize them to database identifiers, and (c) extract candidate interactions. We also briefly describe machine learning approaches used to obtain an optimized scoring of candidate interactions based upon global information from the set of interactions existing in the original database.

Preprocessing and Detection of Domain Entities

Several large-scale terminological resources are used in order to detect names of relevant domain entities in biomedical literature (proteins, genes, chemicals, diseases, etc.) and ground them to widely accepted identifiers assigned by the original database, such as UniProt Knowledgebase, National Center for Biotechnology Information (NCBI) Taxonomy, Proteomics Standards Initiative Molecular Interactions Ontology (PSI-MI), Cell Line Knowledge Base (CLKB), etc.

Terms, i.e. preferred names and synonyms, are automatically extracted from the original database and stored in a common internal format, together with their unique identifiers (as obtained from the original resource). An efficient lookup procedure is used to annotate any mention of

[1] http://www.ontogene.org/

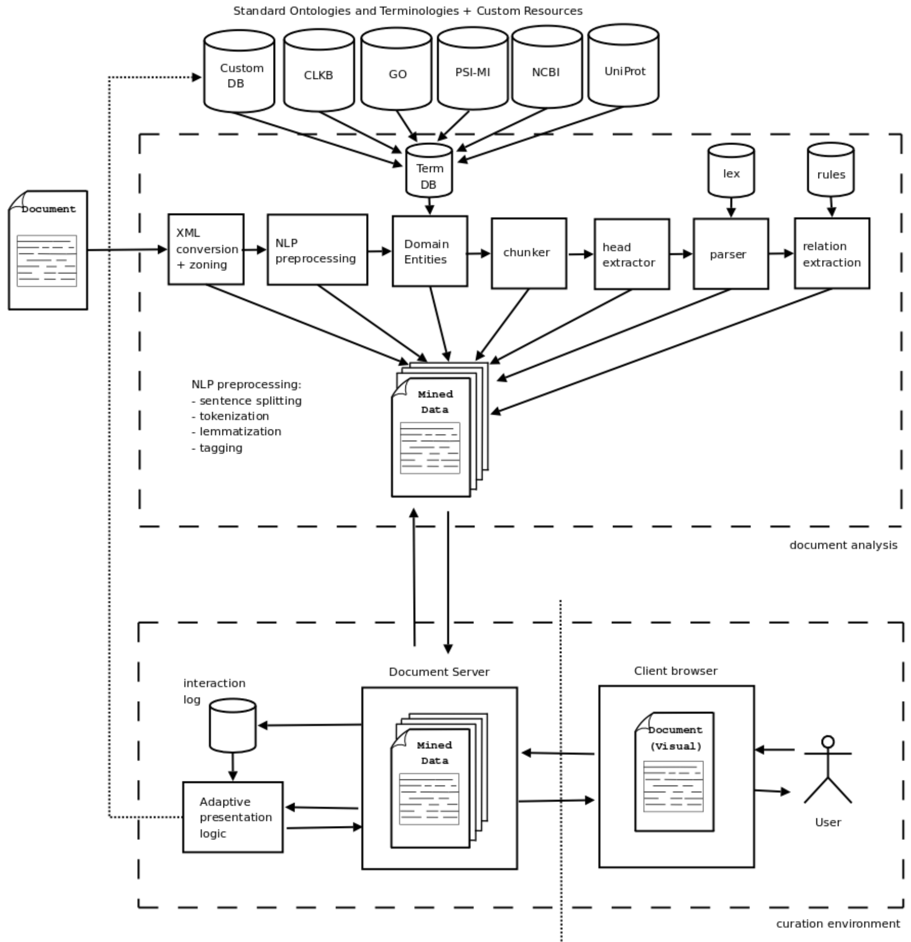


Figure 1: General architecture of the OntoGene system.

a term in the documents with the ID(s) to which it corresponds. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings (Rinaldi et al. 2008). Our normalization rules are similar to the rules reported in (Hakenberg et al. 2008; Wang and Matthews 2008). In case the normalized strings match exactly, the input sequence is annotated with the IDs of the reference term, and no further disambiguation on concepts is done. For more technical details of the OntoGene term recognizer, see (Rinaldi, Kaljurand, and Saetre 2011).

Using the described term list, we can annotate biomedical texts in a straightforward way. First, in a preprocessing stage the input text is transformed into a custom XML format, and sentences and tokens boundaries are identified. For this task, we use the LingPipe tokenizer and sentence splitter which have been trained on biomedical corpora. The tokenizer produces a granular set of tokens, e.g. words that contain a hyphen (such as ‘Pop2p-Cdc18p’) are split into several tokens, revealing the inner structure of such constructs which would allow to discover the interaction mention in “Pop2p-Cdc18p interaction”. Tagging of terms is performed by sequentially processing each token in a sentence and, if it can start a term, annotate the longest possible match (partial overlaps are excluded). In the case of success, all the possible IDs (as found in the term list) are assigned to the candidate term. The annotator ignores certain common English function words, as well as figure and table references (e.g. ‘Fig. 3a’ and ‘Table IV’).

Some of the annotated terms can be very ambiguous, i.e. possibly refer to several database identifiers. This is particularly true in the case of proteins and genes which are typically ambiguous in relation to several species (Tanabe and Wilbur 2002). One way to disambiguate protein and gene names is to apply knowledge about the organisms that are most likely to be the focus of the experiments described in the articles. We have described in (Kappeler, Kaljurand, and Rinaldi 2009) an approach to create a ranked list of ’focus’ organisms. We use such a list in the disambiguation process by removing all the IDs that do not correspond to an organism present in the list. Additionally, the scores provided for each organism can be used in ranking the candidate IDs for each entity. Such a ranking is useful in a semi-automated curation environment where the curator is expected to take the final decision. However, it can also be used in a fully automated environment as a factor in ranking any other derived information, such as interactions where the given entity participates.

Detection of Interactions (baseline)

The information about mentions of relevant domain entities (and their corresponding unique identifiers) can be used to create candidate interactions. In other words, the cooccurrence of two entities in a given text span (typically one or more sentences, or an even larger observation window) is a low-precision, but high-recall indication of a potential relationship among those entities. In order to obtain better precision it is possible to take into account the syntactic structure of the sentence, or the global distribution of interactions in the original database. In this section we describe in detail how candidate interactions are ranked by our system, according to their relevance for the original database.

Depending on the detail in which the interactions in the target database have been curated, a suitable context window needs to be selected. For some applications a single sentence might be sufficient, while for some others this would be too restrictive. In applications dealing with protein-protein interactions (Rinaldi et al. 2008) we found that a context of one sentence could deliver the best results, while in recent applications based upon the PharmGKB database (Rinaldi, Schneider, and Clematide 2012) (drugs, genes, diseases) and CTD database (Clematide and Rinaldi 2012) (chemicals, genes, diseases), we found that one sentence is too restrictive. In an evaluation limited to those PubMed articles from CTD with explicit evidence for at most 12 relations we found the following distribution: For about 32% of all relations from the CTD, where our term recognizer was able to detect both participating entities, there was no sentence containing both entities in the PubMed abstract.

An initial ranking of the candidate relations can be generated on the basis of frequency of occurrence of the respective entities only:

$$relscore(e_1, e_2) = (f(e_1) + f(e_2))/f(E)$$

where $f(e_1)$ and $f(e_2)$ are the number of times the entities e_1 and e_2 are observed in the abstract, while $f(E)$ is the total count of all identifiers in the abstract. An additional zone-based boost might be used in some cases (e.g. for entities mentioned in the title).

The initial set of ranked candidate interactions delivered by the simple approach described above is further refined using a combination of several techniques. In particular we use information delivered from a syntactic analysis of sentences and a machine learning approach which tries to optimize the probability of given entities to participate in an interaction

Detection of Interactions (syntax based)

In this section we describe how to make use of the syntactic analysis components of the OntoGene pipeline in order to extract the interaction type. All sentences in the input documents are parsed using our dependency parser Pro3Gres (Schneider 2008). After parsing, we collect all syntactic connections that exist between all the terms as follows. For each term-coocurrence, i.e. two terms appearing in the same sentence, a collector traverses the tree from one term up to the lowest common parent node, and down the second term, recording all intervening nodes . An example of such a traversal can be seen in Figure 2. Such traversals, commonly called tree walks or paths, have been used in many PPI applications (Kim, Yoon, and Yang 2008). If one records all the information that an intermediate node contains, for example its lexical items and subnodes, the path would be extremely specific, which leads to sparse data and hence a recall problem for most applications. If one only records the grammatical role labels, the paths are too general, which leads often to a precision problem. We record the head lemma of the top node, and the grammatical labels plus prepositions connecting all intervening nodes.

The candidate interactions are ranked according to a combination of several features, including: (1) *Syntactic path*: This feature depends on the syntactic path between two proteins A and B belonging to a candidate interaction. For more details see (Schneider et al. 2009; Rinaldi et al. 2010); (2) *Known interaction*: Interactions that are already reported in the IntAct and MINT databases receive a low score. The older the entry data in the database, the lower the score; (3) *Novelty score*: On the basis of linguistic clues (e.g. *"Here we report that..."*) we attempt to distinguish between sentences that report the results detected by the authors from sentences that report background results. Interactions in 'novelty' sentences are scored higher than interactions in 'background' sentences; (4) *Zoning*: The abstract and the conclusions are the typical places for mentioning novel interactions, the introduction and methods section are less likely and get lower scores; (5) *Pair salience*: Proteins that are mentioned frequently in an article are more likely to participate in a relevant interaction than proteins that are mentioned only once.

The scores of each feature are multiplied, and the total score of a protein-protein interaction is the sum of its occurrences. The result is then normalized to the range `[0,1]` with the following formula: $log(score)/log(maxscore)$. The value of this score is then used for ranking the candidate interactions. A low threshold can then be used to remove the least promising candidates, leading to an increase in precision at the cost of a minimal loss of recall.

The head words of the syntactic paths have a high correspondence to the trigger words used in annotation tasks which use relation labels, such as BioNLP (Cohen et al. 2009). For example, *bind, inhibit, reduce, block, downregulate, metabolize, expression, activate, regulate, express* map to CTD action codes or BioNLP labels. Many heads refer to the investigator's conclusion (*demonstrate, show, assess, find, reveal, explain, suggest*) or to methodology (*treat, exhibit*). Some are underspecified (e. g. *play* which comes from 'play a role in'), and some are only syntactic operators (e.g. *appear, ability*). Some are sematically ambiguous: for example, *contribute* can equally be part of an investigator's conclusion or a syntactic operator (e.g. 'contributes to

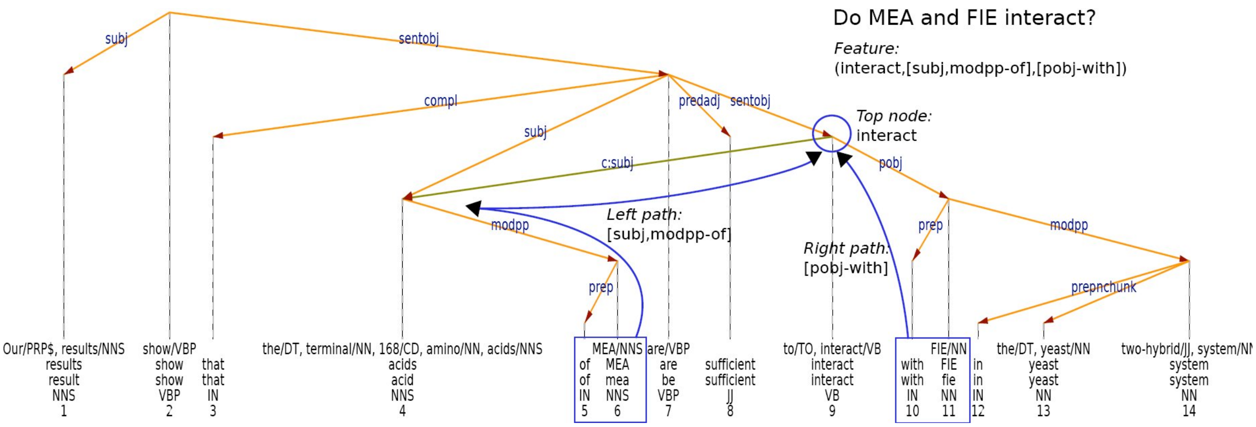


Figure 2:

the activation'). The process of mapping these values into CTD action codes will require biological expertise for completion.

Detection of Interactions (machine learning)

Additionally we apply a supervised machine learning method for scoring the probability of an entity to be part of a relation which was manually curated as an important association in the original database. There are two key motivations for this approach. First, we need to lower the scores of false positive relations which are generated by too broad entities (frequent but not very interesting). The goal is to model some global properties of the curated relations in the original database. Second, we want to penalize false positive concepts that our term recognizer detects. In order to deal with such cases, we need to condition the entities by their normalized textual form t. The combination of a term t and one of its valid entities e is noted as $t\!:\!e$. Due to lack of space we cannot describe in sufficient detail the maximum entropy approach that we use to generate an optimized reranking of candidate interactions. Additional information can be found in (Clematide and Rinaldi 2012).

For example, according to the term database of the CTD the word 'PTEN' (phosphatase and tensin homolog) may denote 9 different diseases (Autistic Disorder; Carcinoma, Squamous Cell; Glioma; Hamartoma Syndrome, Multiple; Head and Neck Neoplasms; Melanoma; Prostatic Neoplasms; Endometrial Neoplasms; Craniofacial Abnormalities) apart from denoting the gene "PTEN". Using background information from the manually curated CTD database we can automatically derive the relevancy of the concepts related to the word "PTEN". Doing so leads to a result which clearly prefers the interpretation of "PTEN" as a gene.

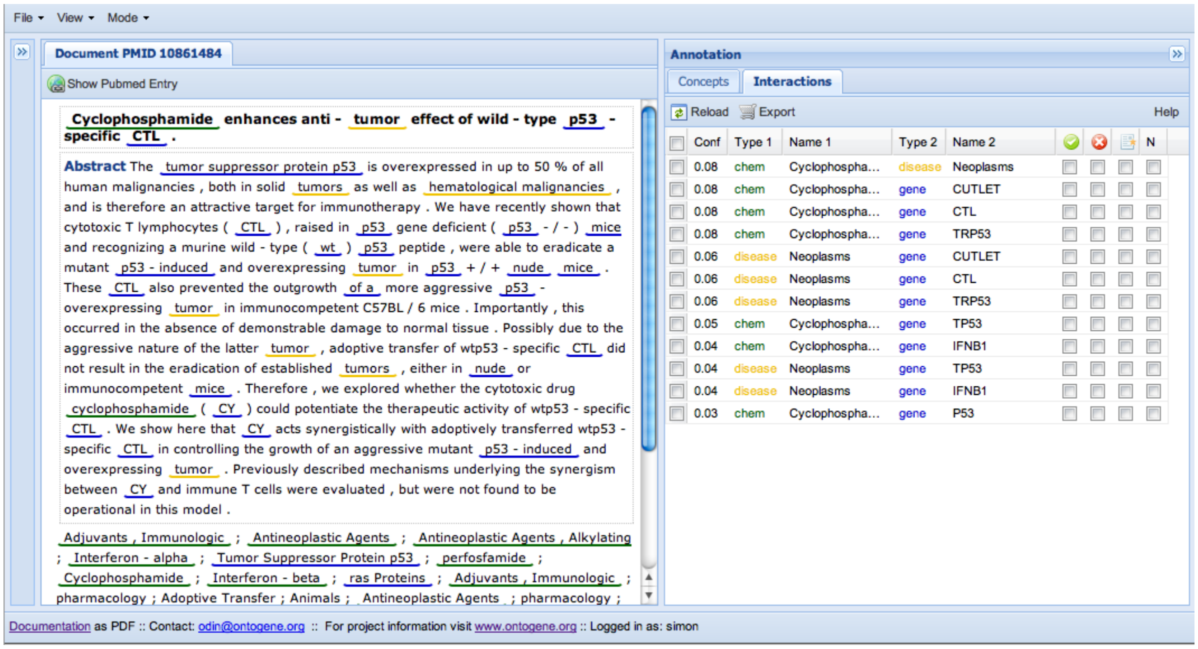
The ODIN Interface

The results of the OntoGene text mining system are made accessible through a curation system called **ODIN** ("OntoGene Document INspector") which allows a user to dynamically inspect the results of their text mining pipeline. A previous version of ODIN was used for participation in the 'interactive curation' task (IAT) of the BioCreative III competition (Arighi et al. 2011). This was an informal task without a quantitative evaluation of the participating systems. However, the curators who used the system commented positively on its usability for a practical curation task. An experiment in interactive curation has been performed in collaboration with curators of the PharmGKB database (Klein et al. 2001; Sangkuhl et al. 2008). The results of this experiment are described in (Rinaldi, Schneider, and Clematide 2012), which also provides further details on the architecture of the system.

More recently, we adapted ODIN to the aims of CTD curation, allowing the inspection of PubMed abstracts annotated with CTD entities and showing the interactions extracted by our system.[2] Once an input term is selected, the system will generate a ranking for all the articles that might be relevant for the target chemical. The pubmed identifier and the title of each article are provided, together with the relevancy score as computed by the sytem. The pubmed identifier field is also an active link, which when clicked brings the user to the ODIN interface for the selected article. Figure 3 shows a screenshot of this interface.

At first access the user will be prompted for a "curator identifier", which can be any string. Once inside ODIN two panels are visible: on the left the article panel, on the right the results panel. The panel on the right has two tabs: concepts and interactions. In the *concept* tabs a list of terms/concepts is presented. Selecting any of them will highlight the terms in the article. In the *interactions* panel the candidate interactions detected by the system are shown. Selecting any of them will highlight evidence in the document.

[2] An entry page for the results of task 1 (training and test sets combined) can be found at the following URL: http://kitt.cl.uzh.ch/kitt/bcms/bc4/ALL/.


File ▾
View ▾
Mode ▾
Document PMID 10861484
Show Pubmed Entry
Annotation
Concepts
Interactions
Reload
Export
Help
Documentation as PDF :: Contact: odin@ontogene.org :: For project information visit www.ontogene.org :: Logged in as: simon

Cyclophosphamide enhances anti - tumor effect of wild - type p53 - specific CTL .

Abstract The tumor suppressor protein p53 is overexpressed in up to 50 % of all human malignancies , both in solid tumors as well as hematological malignancies , and is therefore an attractive target for immunotherapy . We have recently shown that cytotoxic T lymphocytes (CTL) , raised in p53 gene deficient (p53 - / -) mice and recognizing a murine wild - type (wt) p53 peptide , were able to eradicate a mutant p53 - induced and overexpressing tumor in p53 + / + nude mice . These CTL also prevented the outgrowth of a more aggressive p53 - overexpressing tumor in immunocompetent C57BL / 6 mice . Importantly , this occurred in the absence of demonstrable damage to normal tissue . Possibly due to the aggressive nature of the latter tumor , adoptive transfer of wtp53 - specific CTL did not result in the eradication of established tumors , either in nude or immunocompetent mice . Therefore , we explored whether the cytotoxic drug cyclophosphamide (CY) could potentiate the therapeutic activity of wtp53 - specific CTL . We show here that CY acts synergistically with adoptively transferred wtp53 - specific CTL in controlling the growth of an aggressive mutant p53 - induced and overexpressing tumor . Previously described mechanisms underlying the synergism between CY and immune T cells were evaluated , but were not found to be operational in this model .

Adjuvants , Immunologic ; Antineoplastic Agents ; Antineoplastic Agents , Alkylating ; Interferon - alpha ; Tumor Suppressor Protein p53 ; perfosfamide ; Cyclophosphamide ; Interferon - beta ; ras Proteins ; Adjuvants , Immunologic ; pharmacology ; Adoptive Transfer ; Animals ; Antineoplastic Agents ; pharmacology ;

□	Conf	Type 1	Name 1	Type 2	Name 2	✓	✗	📝	N
□	0.08	chem	Cyclophospha...	disease	Neoplasms	□	□	□	□
□	0.08	chem	Cyclophospha...	gene	CUTLET	□	□	□	□
□	0.08	chem	Cyclophospha...	gene	CTL	□	□	□	□
□	0.08	chem	Cyclophospha...	gene	TRP53	□	□	□	□
□	0.06	disease	Neoplasms	gene	CUTLET	□	□	□	□
□	0.06	disease	Neoplasms	gene	CTL	□	□	□	□
□	0.06	disease	Neoplasms	gene	TRP53	□	□	□	□
□	0.05	chem	Cyclophospha...	gene	TP53	□	□	□	□
□	0.04	chem	Cyclophospha...	gene	IFNB1	□	□	□	□
□	0.04	disease	Neoplasms	gene	TP53	□	□	□	□
□	0.04	disease	Neoplasms	gene	IFNB1	□	□	□	□
□	0.03	chem	Cyclophospha...	gene	P53	□	□	□	□

Figure 3: Entity annotations and candidate interactions on a sample PubMed abstract

All items are active. Selecting any concept or interaction in the results panel will highlight the supporting evidence in the article panel. Selecting any term in the article panel prompts the opening of a new panel on the right (annotation panel), where the specific values for the term can be modified (or removed) if needed. It is also possible to add new terms by selecting any token or sequence of tokens in the article.

Adaptation for a triage task

The triage task is the first step of the curation process for several biological databases: it amounts to selecting and prioritizing the articles to be curated in the rest of the process. In BioCreative 2012 (task 1) we implemented a solution to this problem using the assumption that articles should be considered relevant if they contain the target entity provided as input, and their relevance score is increased by the presence of interactions in which the target chemical is involved.

A conventional IR system (Lucene) is used to provide a baseline document classification and ranking. Information derived from the OntoGene pipeline, and from the ranking process described in the previous section, is then added as additional features in order to improve the baseline ranking generated by the IR system. The only significant technical change to Lucene preprocessing is the replacement of the "StandardAnalyzer" component (which is the default analyzer for English, responsible for tokenization, stemming, etc.) with our own tokenization results, as delivered by the OntoGene pipeline. The advantage of this approach is that we can flexibly treat recognized technical terms as individual tokens, and map together their synonyms (Rinaldi et al. 2002). In other words, after this step all known synonyms of a term will be treated as identical by the IR system.

Synonymous terms (as identified by the pipeline) are mapped to their unique identifiers (for this experiment the term identifier provided by the CTD database). The initial search is conducted by mapping the target chemical to the corresponding identifier, which is then used as a query term for the IR system application. All relations which involve a term equivalent to the target (the target or one of its synonyms) are detected. From these relations we extract the interacting entity (the second term in those interactions). An expanded query is created combining the original search term with all other entities which are seen to interact with it in the target abstract. The additional query terms are weighted according to the normalized score of the interactions from which they are extracted.

In the search process, Lucene compares the expanded query with all the entities that are found in any given document. We have experimentally verified on the training data that this query expansion process improves the average MAP scores from 0.622 to 0.694. In the BioCreative 2012 shared task 1 the OntoGene pipeline proved once again its flexibility and efficiency by delivering very effective entity recognition. In particular our system had the best recognition rate for genes and diseases, and the 2nd best for chemicals.

Acknowledgments

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1) and Novartis Pharma AG, NIBR-IT, Text Mining Services, CH-4002, Basel, Switzerland.

References

Arighi, C.; Roberts, P.; Agarwal, S.; Bhattacharya, S.; Cesareni, G.; Chatr-aryamontri, A.; Clematide, S.; Gaudet, P.;

Giglio, M.; Harrow, I.; Huala, E.; Krallinger, M.; Leser, U.; Li, D.; Liu, F.; Lu, Z.; Maltais, L.; Okazaki, N.; Perfetto, L.; Rinaldi, F.; Saetre, R.; Salgado, D.; Srinivasan, P.; Thomas, P.; Toldo, L.; Hirschman, L.; and Wu, C. 2011. Biocreative iii interactive task: an overview. *BMC Bioinformatics* 12(Suppl 8):S4.

Clematide, S., and Rinaldi, F. 2012. Ranking interactions for a curation task. *Journal of Biomedical semantics*. accepted for publication.

Cohen, K. B.; Demner-Fushman, D.; Ananiadou, S.; Pestian, J.; Tsujii, J.; and Webber, B., eds. 2009. *Proceedings of the BioNLP 2009 Workshop*. Boulder, Colorado: Association for Computational Linguistics.

Hakenberg, J.; Plake, C.; Royer, L.; Strobelt, H.; Leser, U.; and Schroeder, M. 2008. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology* 9(Suppl 2):S14.

Hermjakob, H.; Montecchi-Palazzi, L.; Lewington, C.; Mudali, S.; Kerrien, S.; Orchard, S.; Vingron, M.; Roechert, B.; Roepstorff, P.; Valencia, A.; Margalit, H.; Armstrong, J.; Bairoch, A.; Cesareni, G.; Sherman, D.; and Apweiler, R. 2004. IntAct: an open source molecular interaction database. *Nucl. Acids Res.* 32(suppl 1):D452–455.

Kappeler, T.; Kaljurand, K.; and Rinaldi, F. 2009. TX Task: Automatic Detection of Focus Organisms in Biomedical Publications. In *Proceedings of the BioNLP workshop, Boulder, Colorado*, 80–88.

Kim, S.; Yoon, J.; and Yang, J. 2008. Kernel approaches for genic interaction extraction. *Bioinformatics* 9:10.

Klein, T.; Chang, J.; Cho, M.; Easton, K.; Fergerson, R.; Hewett, M.; Lin, Z.; Liu, Y.; Liu, S.; Oliver, D.; Rubin, D.; Shafa, F.; Stuart, J.; and Altman, R. 2001. Integrating genotype and phenotype information: An overview of the PharmGKB project. *The Pharmacogenomics Journal* 1:167–170.

Mattingly, C.; Rosenstein, M.; Colby, G.; Forrest Jr, J.; and Boyer, J. 2006. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology* 305A(9):689–692.

Rinaldi, F.; Dowdall, J.; Hess, M.; Kaljurand, K.; Koit, M.; Vider, K.; and Kahusk, N. 2002. Terminology as Knowledge in Answer Extraction. In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE02)*, 107–113.

Rinaldi, F.; Schneider, G.; Kaljurand, K.; Hess, M.; and Romacker, M. 2006. An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA. *BMC Bioinformatics* 7(Suppl 3):S3.

Rinaldi, F.; Kappeler, T.; Kaljurand, K.; Schneider, G.; Klenner, M.; Clematide, S.; Hess, M.; von Allmen, J.-M.; Parisot, P.; Romacker, M.; and Vachon, T. 2008. OntoGene in BioCreative II. *Genome Biology* 9(Suppl 2):S13.

Rinaldi, F.; Schneider, G.; Kaljurand, K.; Clematide, S.; Vachon, T.; and Romacker, M. 2010. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7(3):472–480.

Rinaldi, F.; Clematide, S.; Garten, Y.; Whirl-Carrillo, M.; Gong, L.; Hebert, J. M.; Sangkuhl, K.; Thorn, C. F.; Klein, T. E.; and Altman, R. B. 2012. Using ODIN for a PharmGKB re-validation experiment. *Database: The Journal of Biological Databases and Curation*.

Rinaldi, F.; Kaljurand, K.; and Saetre, R. 2011. Terminological resources for text mining over biomedical scientific literature. *Journal of Artificial Intelligence in Medicine* 52(2):107–114.

Rinaldi, F.; Schneider, G.; and Clematide, S. 2012. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*.

Sangkuhl, K.; Berlin, D. S.; Altman, R. B.; and Klein, T. E. 2008. PharmGKB: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews* 40(4):539–551. PMID: 18949600.

Schneider, G.; Kaljurand, K.; Kappeler, T.; and Rinaldi, F. 2009. Detecting Protein/Protein Interactions using a parser and linguistic resources. In *Proceedings of CICLing 2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics*, 406–417. Mexico City, Mexico: Springer LNC 5449.

Schneider, G. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Tanabe, L., and Wilbur, W. J. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics* 18(8):1124–1132.

UniProt Consortium. 2007. The universal protein resource (uniprot). *Nucleic Acids Research* 35:D193–7.

Wang, X., and Matthews, M. 2008. Distinguishing the species of biomedical named entities for term identification. *BMC Bioinformatics* 9(Suppl 11):S6.

Zanzoni, A.; Montecchi-Palazzi, L.; Quondam, M.; Ausiello, G.; Helmer-Citterich, M.; and Cesareni, G. 2002. MINT: a Molecular INTeraction database. *FEBS Letters* 513(1):135–140.