Assisted editing in the biomedical domain: motivation and challenges

Rinaldi, Fabio

DOI: https://doi.org/10.1145/2494266.2494268

Originally published at:
Rinaldi, Fabio (2013). Assisted editing in the biomedical domain: motivation and challenges. In: ACM Symposium on Document Engineering, Florence, Italy, 10 September 2013 - 12 September 2013. ACM, 197-200.
DOI: https://doi.org/10.1145/2494266.2494268

Assisted Editing in the Biomedical Domain: Motivation and Challenges

Fabio Rinaldi
Institute for Computational Linguistics
University of Zurich, Switzerland
fabio.rinaldi@uzh.ch

ABSTRACT

One of the characteristics of biomedical scientific literature is the high ambiguity of the domain-specific terminology which can be used to describe technical concepts and specific objects of the domain. This is partly due to the very broad scope of the domain of interest and partly to inherent properties of the terminology itself. There are simply very large numbers of genes, proteins, organs, cell lines, cellular phenomena, experimental methods, and so on. For example, UniProt, the most authoritative protein database, currently contains more than 33 million entries. Clearly, the names which are typically used to refer to proteins are polysemic and might refer to hundreds of different entries in a reference database.

Such a large and extensive terminology necessarily makes it difficult to derive from the literature a simplified representation of the entities and relationships described in the articles, despite considerable efforts by the text mining community. In this paper we propose to complement such efforts with editing tools that can assist the authors in efficiently adding to their publications a minimal semantic annotation so that much of the ambiguity is avoided.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

biomedical literature processing, text mining

1. INTRODUCTION

An emerging trend in information processing is based upon the usage of **text mining** tools for the extraction of tailored information from textual reports such as newspaper articles or scientific publications. The constantly growing amount of information needs to be properly managed in order to be a support in everybody's daily life rather than a burden. One area where this problem is particularly evident is that of research in **molecular biology**. Text Mining tools aim at supporting the process of knowledge gaining from the literature, by supporting the search for relevant articles, the semi-automated extraction of relevant passages, and the transformation of the information from the textual format to some suitable semantic format.

There are multiple possible scenarios for the application of text mining tools in biology. The most basic scenario, and the one currently most frequently pursued, is the creation of tools for supporting the professional end-user (e.g. a researcher in molecular biology), who autonomously wishes to browse the existing literature in search of information relevant to a particular information need. Another possible usage is within the process of literature curation, which is the activity performed by professionals who are paid to read the literature in search of particular items of information (e.g. newly detected protein interactions), and store such information in public databases, which can in turn be accessed later by the biologists. One example is IntAct [4], a database of protein-protein interactions maintained at the European Bioinformatics Institute. A third even more advanced usage scenario would be within advanced authoring tools for the authors of scientific literature. Novel text mining tools can be used to suggest candidate *semantic annotations* to the author or curators, depending on the scenario of application. The semantic annotations (formal descriptions of the main entities and relationships discussed in the paper), manually confirmed (or modified) by the author or curator, can then be stored together with the electronic version of the article, using one of the standard formats developed by the semantic web community, allowing a much more efficient information retrieval and processing.

The Semantic Web movement aims at enriching web resources with semantic annotations which will allow remote agents to easily find and use them. However the problem of creating these annotations is seldom addressed. Manual creation of the annotations is not a feasible option, except in a few experimental cases. We believe that Natural Language Processing techniques are mature enough to help addresses this issue, at least for textual resources (which still constitute the vast majority of the material available on the web). Documents can be analized fully automatically and converted into a semantic annotation, which can then be stored together with the original documents. It is this annotation that constitutes the machine understandable resource that remote agents can query. A semi-automatic approach is also considered, in which the system suggests candidate

DocEng'13, September 10–13, 2013, Florence, Italy.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-1789-4/13/09 ...$15.00.
http://dx.doi.org/10.1145/2494266.2494268 .

annotations and the user simply has to approve or reject them.

The benefits of the semantic web should come for free to most of the users: semantic markup should be a by-product of normal computer use. There is a real need to lower the barrier of entry: the vast majority of the users cannot be expected to understand and use formal ontologies. In order to achieve interoperability between software agents, a lot of human understandability has been sacrificed: precise ontologies and formally defined semantics are foreign concepts to the average users.

One of the motivations behind the semantic web movement was that computers are not powerful enough to process (and understand) natural language. Therefore machine understandable information should be added to web resources. This is still true: it would be unfeasible to process the enormous amounts of textual resources that are added to the web every day (let alone process all the existing web content). However, it is technically possible (and practically conceivable) to have specialised editors that process (in a transparent fashion) textual resources as the users publish them on the web, and add semantic annotations automatically extracted from the documents. In other words, the idea is to move the problem from the consumer to the producer of the information.

The OntoGene team at the University of Zurich has developed advanced text mining applications based on a combination of deep-linguistic analysis and machine learning techniques [14, 2, 11]. In the rest of this position paper, after describing in Section 2 the overall architecture of our text mining system, we illustrate our integrated assisted editing environment (Section 3), and we provide a short discussion on previous evaluations of the system through participation in community-organized text mining tasks (Section 4).

2. THE ONTOGENE TEXT MINING SYSTEM

In this section we provide a brief description of the OntoGene text mining environment. The first step in order to process a collection of biomedical literature consists in the annotation of names of relevant domain entities in biomedical literature (we consider in particular proteins, genes, species, experimental methods, and cell lines) and grounding them to widely accepted identifiers (IDs) such as those assigned by the UniProt Knowledge Base or the Cell Line Knowledge Base (CLKB). The term annotation uses a large term list that is compiled on the basis of the entity names extracted from the mentioned knowledge bases. This resulting list covers the common expression of the terms. A term normalization step is used to match the terms with their actual representation in the text, taking into account a number of possible surface variations. Finally, a disambiguation step tries to resolve the possible ambiguity of the matched terms [11].

A marked-up term can be ambiguous for two reasons. First, the term can be assigned an ID from different term types, e.g. a UniProtKB ID and a PSI-MI Ontology ID. This situation does not occur often and usually happens with terms that are probably not veru interesting. We disambiguate such terms by removing all the UniProtKB IDs, similarly to what done in [17]. Second, the term can be assigned several IDs from a single type. This usually happens with UniProtKB terms and is typically due to the fact that the same protein occurs in many different species. One way to disambiguate such protein 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 [5] 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.

Using the information concerning mentions of relevant domain entities, derived as described above, and their corresponding unique identifiers obtained by the process of disambiguation, it is possible to create candidate interactions. We use an approach based on a combination of syntactic parsing and machine learning [12, 2]. From the results of syntactic analysis we can derive a number of possible entity interactions, such as protein-protein interactions [15] or drug-gene-disease interactions [14]. These candidate interactions can then be manually validated by the target users, be they expert database curators or the original authors of the article.

The end result of processing by the OntoGene pipeline is a richly annotated version of the original document, where annotations are organized into three levels:

- **Structural Annotations**
- **Lexical Annotations**
- **Semantic Annotations**

Structural annotations are used to define the physical structure of the document, it's organization into head and body, into sections, paragraphs and sentences. Lexical annotations identify lexical units that have some relevance for the project. Semantic annotations are meant to represent the propositional content of the document (the "meaning"). While structural annotations apply to large text spans, lexical annotations apply to smaller text spans (sub-sentence) and semantic annotations are not directly associated to a specific text span, however, they are linked to text units by co-referential identifiers. All annotations are required to have an unique ID and thus will be individually addressable, this allows semantic annotations to point to the lexical annotations to which they correspond. Semantic Annotations themselves are given a unique ID, and therefore can be elements of more complex annotations.

The structure of the documents is marked using an intuitively appropriate scheme based on the TEI recomendations. Broadly speaking, structural annotations are concerned with the organization of documents into sub-units, such as section, title, paragraphs and sentences.

Lexical Annotations are used to mark any text unit (smaller than a sentence), which can be of interest in the application. They include (but are not limited to): Named Entities in the classical MUC sense, new domain-specific Named Entities, Terms, Temporal Expressions, Events.

The relations that exist between lexical entities are expressed through the semantic annotations. So lexically identified entities can be linked to other entities in case they are in some interesting relationship, such a protein-protein interaction or a drug-gene correlation.

3. ODIN

Despite the significant improvements in the last couple of years, most experts agree that, at least for the time being, it is unrealistic to expect fully automated text mining systems to perform at a level acceptable for tasks that require high accuracy. However, existing systems can already achieve results which are sufficiently good to be used in a semi-automated context, where a human expert validates the output of the system. One application where this support is badly needed is biomedical literature curation. Our ODIN system was originally developed starting in 2008 as a visualization interface for the OntoGene text mining system (see Figure 1). It was later modified to serve as a literature curation tool, and in this new role it was first presented in 2010 [9]. We now plan to extend the usage of ODIN to authors of scientific paper, who, even better than curators, can easily disambiguate ambiguous terms, since they are in possession of the primary knowledge that drove their editing decisions.

In the past couple of years a few similar tools have been described in the literature. REFLECT [6] can be used to annotated publications with gene, protein, or small molecules. It can be operated either through a browser plugin or remotely via web interface. DOMEO [1] is a more recent literature curation tools which supports several types of curation paradigms (e.g. highlighting, adding notes, adding semantic tags). It is also notable for its strong integration with ontology services such as those provided by the NCBO (National Center for Biomedical Ontology). However, DOMEO does not offer text mining capabilities on its own, relying instead on external services for this purpose.

ODIN is unique in that it integrates advanced text mining capabilities with a user friendly interface. In case of ambiguity, the curator or author is offered the opportunity to correct the choices made by the system, at any of the different levels of processing: entity identification and disambiguation, organism selection, interaction candidates. The user can access all the possible readings given by the system and select the most accurate. Candidate interactions are presented in a ranked order, according to the score assigned by the system. The user can, for each of them, confirm, reject, or leave undecided. The results of the curation process can be fed back into the system, thus allowing incremental learning.

The documents and the annotations are represented consistently within a single XML file, which also contains a record of the user interaction, thus allowing advanced logging support. The annotations are selectively presented, in a ergonomic way through CSS formatting, according to different view modalities, While the XML annotations are transparent to the annotator (who therefore does not need to have any specialized knowledge beyond his biological expertise), his/her verification activities result in changes at the DOM of the XML document through client-side JavaScript. The use of modern AJAX methodology allows for online integration of background information, e.g. information from different term and knowledge bases, or further integration of foreign text mining services.

The presence of the raw XML in the browser document gives the flexibility to compile dynamically tabular grid views of terms and relations including filtering, reordering, and editing the annotations in a spreadsheet-like way (this includes also chart visualizations). To keep the implementation effort feasible, the use of a dedicated JavaScript application framework is crucial. Among several available JavaScript frameworks, ExtJS, jQuery and Prototype were our main candidates. We then decided for ExtJS because of its compact and coherent architecture covering all kinds of GUI widgets. The advantage of a client-side presentation logic is the flexibility for the end user and the data transparency. For text mining applications, it is important to be able to link back curated metainformation to its textual evidence.

4. EVALUATION

As a way to verify the quality of the core text mining functionalities of the OntoGene system, we have participated in a number of text mining evaluations campaigns [8, 3, 12, 13]. Some of most interesting results include best results in the detection of protein-protein interactions in BioCreative 2009 [15], top-ranked results in several tasks of BioCreative 2010 [16], best results in the triage task of BioCreative 2012 [8].

The usage of ODIN as a curation tool has been tested in a few collaborations with curation groups, including PharmGKB [9], CTD [7], RegulonDB [10]. The next challenge is to test it in a suitable assisted editing scenario.

5. CONCLUSIONS

We have presented an advanced text mining architecture (OntoGene Text Miner), which is embedded within a user-friendly curation interface (ODIN). Currently ODIN is meant to support the activities of expert database curators, however we are planning to further develop it into a tool that will assist authors in creating semantic annotations to be added to papers at time of creation by the authors themselves.

6. ACKNOWLEDGMENTS

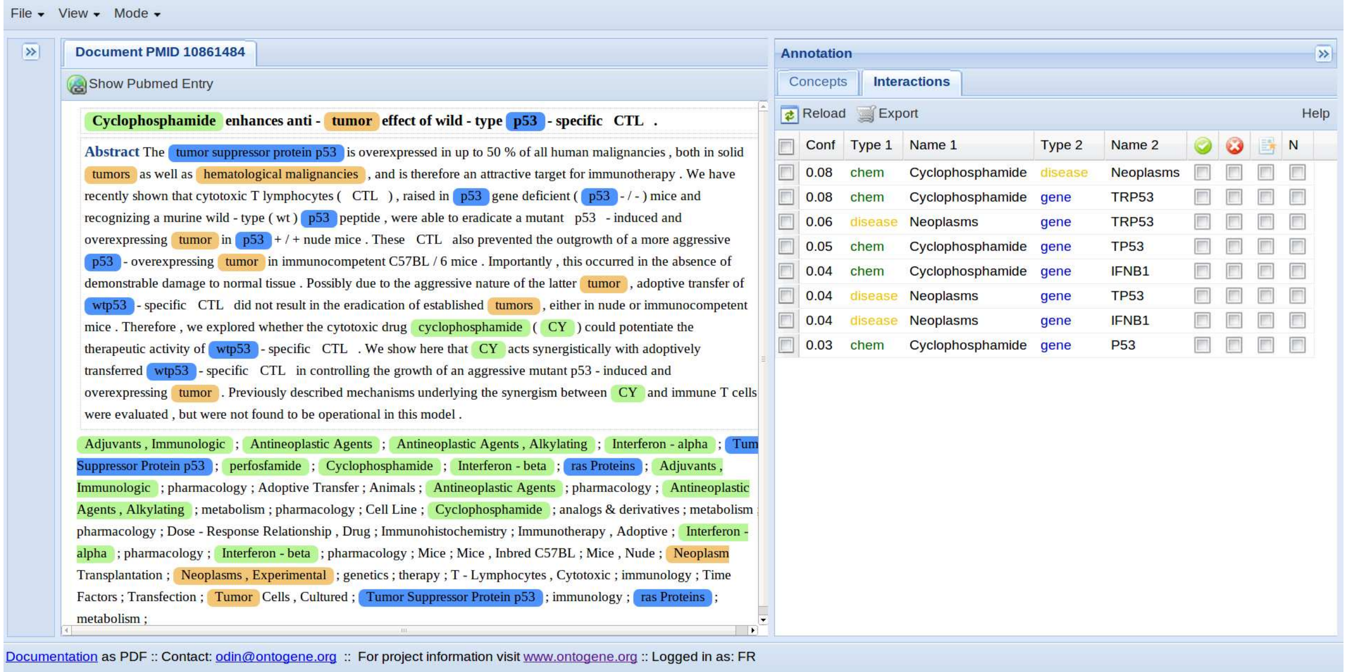
The OntoGene group is partially supported by the Swiss National Science Foundation (grants $100014-118396/1$ and $105315-130558/1$) and by NIBR/IT, Text Mining Services, Novartis Pharma AG, Basel, Switzerland. The author would like to thank the anonymous reviewers for their useful comments.

7. REFERENCES

[1] P. Ciccarese, M. Ocana, and T. Clark. Open semantic annotation of scientific publications using DOMEO. *J Biomed Semantics*, 3 Suppl 1:S1, 2012.

[2] S. Clematide and F. Rinaldi. Ranking relations between diseases, drugs and genes for a curation task. *Journal of Biomedical Semantics*, 3(Suppl 3):S5, 2012.

[3] S. Clematide, F. Rinaldi, and G. Schneider. Ontogene at calbc ii and some thoughts on the need of document-wide harmonization. In *Proceedings of the CALBC II workshop, EBI, Cambridge, UK, 16-18 March*, 2011.



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 ; Tum Suppressor Protein p53 ; perfosfamide ; Cyclophosphamide ; Interferon - beta ; ras Proteins ; Adjuvants , Immunologic ; pharmacology ; Adoptive Transfer ; Animals ; Antineoplastic Agents ; pharmacology ; Antineoplastic Agents , Alkylating ; metabolism ; pharmacology ; Cell Line ; Cyclophosphamide ; analogs & derivatives ; metabolism pharmacology ; Dose - Response Relationship , Drug ; Immunohistochemistry ; Immunotherapy , Adoptive ; Interferon - alpha ; pharmacology ; Interferon - beta ; pharmacology ; Mice ; Mice , Inbred C57BL ; Mice , Nude ; Neoplasm Transplantation ; Neoplasms , Experimental ; genetics ; therapy ; T - Lymphocytes , Cytotoxic ; immunology ; Time Factors ; Transfection ; Tumor Cells , Cultured ; Tumor Suppressor Protein p53 ; immunology ; ras Proteins ; metabolism ;

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

Documentation as PDF :: Contact: odin@ontogene.org :: For project information visit www.ontogene.org :: Logged in as: FR

Figure 1: A screenshot showing the ODIN interface

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

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

[6] E. Pafilis, S. I. O'Donoghue, L. J. Jensen, H. Horn, M. Kuhn, N. P. Brown, and R. Schneider. Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, 27(6):508–510, Jun 2009.

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

[8] F. Rinaldi, S. Clematide, S. Hafner, G. Schneider, G. Grigonyte, M. Romacker, and T. Vachon. Using the ontogene pipeline for the triage task of biocreative 2012. *The Journal of Biological Databases and Curation, Oxford Journals*, 2013.

[9] F. Rinaldi, S. Clematide, G. Schneider, M. Romacker, and T. Vachon. ODIN: An advanced interface for the curation of biomedical literature. In *Biocuration 2010, the Conference of the International Society for Biocuration and the 4th International Biocuration Conference.*, page 61, 2010. Available from Nature Precedings `http://dx.doi.org/10.1038/npre.2010.5169.1`.

[10] F. Rinaldi, S. Gama-Castro, A. López-Fuentes, Y. Balderas-Martínez, and J. Collado-Vides. Digital curation experiments for regulondb. In *BioCuration 2013, April 10th, Cambridge, UK*, 2013.

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

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

[13] F. Rinaldi, T. Kappeler, K. Kaljurand, G. Schneider, M. Klenner, M. Hess, J.-M. von Allmen, M. Romacker, and T. Vachon. OntoGene in Biocreative II. In *Proceedings of the II Biocreative Workshop*, 2007.

[14] F. Rinaldi, G. Schneider, and S. Clematide. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*, 45(5):851–861, 2012.

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

[16] G. Schneider, S. Clematide, and F. Rinaldi. Detection of interaction articles and experimental methods in biomedical literature. *BMC Bioinformatics*, 12(Suppl 8):S13, 2011.

[17] L. Tanabe and W. Wilbur. Tagging gene and protein names in biomedical text. *bioinformatics*, 18(8):1124–32, 2002.