

# ToxiCat: Hybrid Named Entity Recognition services to support curation of the Comparative Toxicogenomic Database

Dina Vishnyakova<sup>1,2,4,\*</sup>, Julien Gobeill<sup>1,3,4</sup>, Emilie Pasche<sup>1,2,3,4</sup> and Patrick Ruch<sup>1,3,4</sup>

<sup>1</sup> Bibliomics and Text Mining (BiTeM) Group: <http://bitem.hesge.ch>

<sup>2</sup> Division of Medical Information Sciences, University and University Hospitals of Geneva

<sup>3</sup> Information Science Department, HES-SO/University of Applied Science Geneva

<sup>4</sup> SIBtex, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

\*Corresponding author: SIMED; University Hospitals of Geneva; 4, rue Gabrielle-Perret-Gentil; CH-1211 Geneva 14; Tel: +41 22 372 61 99; email: [dina.vishnyakova@hcuge.ch](mailto:dina.vishnyakova@hcuge.ch)

## Abstract

We report on the original implementation of named entity recognition (NER) modules based on an automatic text categorization pipeline, so-called ToxiCat (Toxicogenomic Categorizer), developed to perform biomedical documents classification and prioritization for the previous Biocreative campaign in order to speed up the curation of the Comparative Toxicogenomics Database (CTD). ToxiCat NER modules are a group of components that analyse text for enclosed information. These modules are based on an information retrieval engine for MEDLINE (EAGLi), a gene normalization (GN) service (NormaGene) developed for a previous BioCreative campaign, gene ontology categorizer (GOCat) and finally an entity recognizer for diseases and chemicals. The NER services are publically available as RESTful web services at <http://pingu.unige.ch:8080/Toxicat>.

## Introduction

The recognition of biomedical concepts in texts is a key technology for automatic or semi-automatic analysis of textual resources. Most of applications are based on Named Entity Recognition (NER) tools in information retrieval, information extraction and document classification tasks. In recent years, NER systems development has reached great attention in the bioinformatics community. Multiple systems and algorithms have been developed and implemented. These systems and algorithms can be roughly split into 3 categories: rule-based and dictionary-based systems, fully automatic machine-learning systems and hybrids approaches, combining first two categories. Most tools require the user to specify certain configuration settings, like choosing a dictionary or creating an appropriate corpus of annotated texts in order to perform a reliable assessment where the operation to find or to design such a dataset could be time-consuming.

The work we present here is focused on the construction of some NER tools for the curation of the CTD (1), where the main accents are put on the identification of gene/protein, chemical, disease, and chemical/gene-specific action term mentions, each within the context of CTD's controlled vocabulary structure. We should notice that there are several information available about the development of the NER systems for gene/protein, chemical or disease concepts while the identification of a chemical/gene-specific action term is covered only within the framework of CTD. The representation of a chemical/gene-specific action term in a text is often not implicit. We have used components such as EAGLi's Keyword extractor (2) and NormaGene (3) in order to ease the process of systems configuration or to avoid time-consuming dataset-couple processes and finally GOCat (4) to solve the problem with the chemical/gene-specific action term recognition.

## Data and Methods

### Data overview

The CTD track of BioCreative IV proposes to focus on the interoperability, e.g. to explore how text-mining methods can successfully be applied to practically help biocuration of a large molecular biology knowledge base. The main objective of the Track-3 task is to provide Web Services for concepts annotations to maintain the Comparative Toxicogenomics Database (CTD) with the interacting entities (small molecules and gene products) and the pathologies likely to reflect the toxicity of the chemical compound.

The organizers provided a learning corpus in BioC XML format of 1,112 abstracts for training; all curated gene/protein, chemical, and disease actors, and associated chemical/gene-specific action terms. Each curated interaction associated with the article is also provided. Testing set consisted of 510 documents.

### Methods

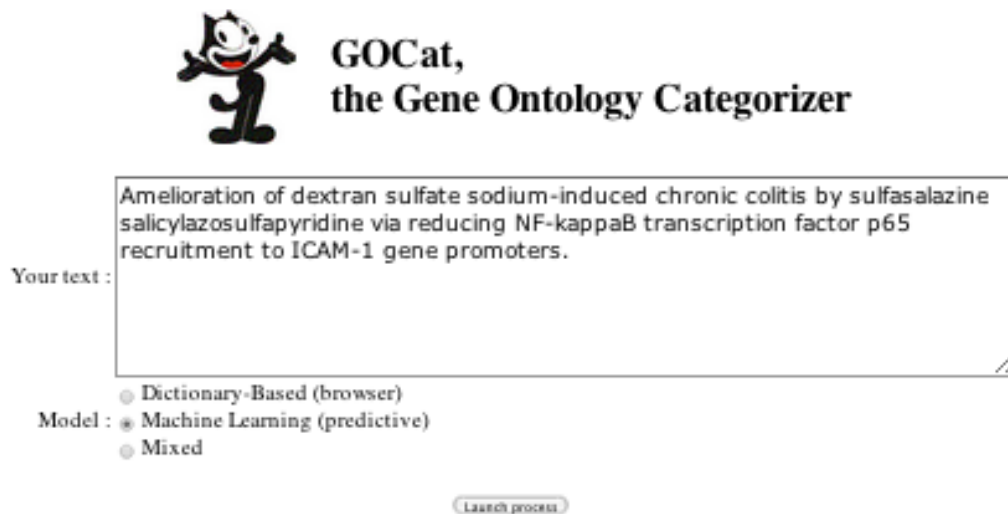
We designed NER services for each category of interest, i.g. for gene/protein, disease, chemical and chemical/gene action term as following:

- **Gene/Protein NER:** We based our NER service for gene/protein concepts on the NormaGene named-entity normalizer (3,5). This gene and protein named-entity recognizer was developed for the BioCreative III task to address the gene normalization task (3). Like other named-entity recognizer, it identifies the patterns of the gene and protein name as well as it attempts to assign a unique identifier. Thus, NormaGene also attempts to recognize, when possible, what organisms is mentioned in the text to link properly a gene/protein name with a unique sequence. Internally, NormaGene is able to recognize all gene candidates stored in the Gene and Protein Synonyms DataBase (GPSDB) (6), as well as all species stored in NEWT ([www.ebi.ac.uk/newt/](http://www.ebi.ac.uk/newt/))), which is appropriate to annotate contents for UniProt/SwissProtKB but which does exceed the

coverage of CTD (7). The internal dictionaries of NormaGene are therefore reduced to curate CTD. Finally, results returned by NormaGene are compared to the CTD genes controlled vocabulary to further reduce the list of results. The controlled vocabulary of CTD contains over 257.000 NCBI genes' identifiers and over 479.000 genes' names including synonyms (5). If the entities recognized by NormaGene are found in the CTD genes' vocabulary then we extract all synonyms based on the approved genes ID and match them against the abstract. Indeed, gene and protein identifiers suggested by NormaGene cannot always be explicitly found in the body of the input document as NormaGene uses a generative model, which exploits also functional similarities (3) and not only textual similarities. Additionally, as a final check all candidates selected by NormaGene NER tool are matched against synonyms from a provided dictionary of CTD.

- **Disease/Chemical NERs:** We created an ad-hoc keyword recognizer for diseases and chemicals. This keyword recognizer is based on the controlled vocabularies provided by CTD. Unlike the previous results of CTD Triage task in Biocreative 2012, where systems showed high results based on Recall the current task (Track-3-CTD) is taking into account the Precision of the system, see (5) for more details. Disease/Chemical NER relies on the UMLS Metathesaurus. For both chemical and disease entities, a Word-Sense Disambiguator (WSD) is created, based on the UMLS Semantic Types (5, 8). In (5) we have described in details which types of chemicals and diseases were eliminated from the final result. Further, in order to avoid common English words in the list of candidates, we created a common English word recognizer based on a general-purpose English corpora. Unspecific disease and chemical names were thus discarded.
- **Chemical/Gene action term NER:** CTD curates specific chemical–gene and protein interactions in vertebrates and invertebrates from the published literature. Most interactions are binary, involving one chemical and one gene or protein. After exploring the corpus provided by the organizers we found that the information about chemical/gene action term is not represented explicitly in the text of the provided corpus. Since the concept of action term identification in the text is not widely covered by the bioinformatics community, it makes the task especially complex. We assumed that to use Gene Ontology could help to identify action terms. Ontological approaches rely on formal ontological principles to formalize the relations expected between biological entities according to general theories specified in some upper-level ontologies (9). In the Gene Ontology, we can observe that several chemical entities are found in GO descriptors and synonyms (9). Consequently we attempted to assign some GO concepts to the input text using the Gene Ontology Categorizer – GOCat (4), see Fig.1 and Fig 2. GOCat is a state-of-the-art thesaurus-based system combined with a machine learning system (4). The output of GOCat is a ranked list of candidate GO terms, which are the most likely to characterize the functional profile of a given abstract. Next, we process GOCat results with the developed NER based on a dictionary where all action terms

provided by organizers (<http://ctdbase.org/help/ixnQueryHelp.jsp?actionType>) were included.



**GOCat,**  
**the Gene Ontology Categorizer**

Your text :  
Amelioration of dextran sulfate sodium-induced chronic colitis by sulfasalazine salicylazosulfapyridine via reducing NF-kappaB transcription factor p65 recruitment to ICAM-1 gene promoters.

Model : ☒ Dictionary-Based (browser) ☒ Machine Learning (predictive) ☐ Mixed

[Launch process](#)

**Figure 1.** The GOCat interface where as an input the user can provide a text, e.g. an abstract of the document and choose the processing of results between three models: Dictionary-Based, Machine learning and Mixed.



← → ↺ eagl.unige.ch/GOCat/result.jsp

Apple Yahoo! New folder YouTube Википедия Новости Популярн

🏠

**This is the output of the Machine Learning model.**

Why are some extracted passages irrelevant ? [+/-](#)

Go to : [molecular\\_function](#) (13) / [biological\\_processes](#) (32) / [cellular\\_components](#) (5)

**All concepts**

#	Score	GO ID	Name
1	1.00	GO:0005515	protein binding
2	0.75	GO:0005634	nucleus
3	0.47	GO:0051059	NF-kappaB binding <a href="#">+/-</a>
4	0.28	GO:0005737	cytoplasm
5	0.25	GO:0051092	positive regulation of NF-kappaB transcription factor activity <a href="#">+/-</a>
6	0.25	GO:0032088	negative regulation of NF-kappaB transcription factor activity <a href="#">+/-</a>
7	0.17	GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB cascade <a href="#">+/-</a>
8	0.13	GO:0003700	sequence-specific DNA binding transcription factor activity <a href="#">+/-</a>
9	0.13	GO:0008134	transcription factor binding <a href="#">+/-</a>
10	0.10	GO:0006468	protein phosphorylation
11	0.10	GO:0042493	response to drug
12	0.10	GO:0000122	negative regulation of transcription from RNA polymerase II promoter <a href="#">+/-</a>
13	0.10	GO:0014070	response to organic cyclic compound

**Figure 2.** An example of the GOCat output returned for the input of “Amelioration of dextran sulfate sodium-induced chronic colitis by sulfasalazine salicylazosulfapyridine via reducing NF-kappaB transcription factor p65 recruitment to ICAM-1 gene promoters.” Here, the output is a list of the most associated GO concepts, which are split into the three GO axes: molecular functions, biological processes and cellular components.

## Results and Conclusion

The results of ToxiCat (Group 183), computed on the official data provided by BioCreative 2012's organizers using official metrics, are shown in Table 1, where:

- Curated Actors - the terms curated for the current document by CTD in the respective NER category.
- Text Mined Actors - the text mined terms returned from the NER Web Service on a provided document
- Text Mined Actors Hits - provides an explanation of how matches between the curated terms and the text mined terms were determined.

**Table1.** Toxicat NER services results of the Track-3.

	disease	chemical	gene/protein	action term
<b>Records Processed</b>	510	510	510	510
<b>Text Mined Actors</b>	795	1156	1062	1763
<b>Text Mined Actor Hits</b>	366	685	370	450
<b>Curated Actors</b>	943	1192	1122	966
<b>Micro-Average Recall Aggregate Curated Actors</b>	0.388	0.57	0.32	0.46
<b>Macro-Average Recall Aggregate Curated Actors</b>	0.396	0.56	0.35	0.45
<b>Micro-Average Precision</b>	0.46	0.59	0.348	0.255
<b>Macro-Average Precision</b>	0.40	0.55	0.342	0.259
<b>Average Seconds Processed</b>	0.43	0.83	4.40	24.22

In BioCreative IV, Track 3 was investigated to interoperability and efficiency aspects; therefore the ability to integrate a particular workflow and the processing time were assessed. In Table 1, disease and chemical NER services show quite acceptable average response time compared to gene and action term NER services. This result suggests that a general-purpose gene normalizer and the NER service based on gene ontology categorizer are time-consuming for a specific database curation task. However, action term NER service is competitive regarding the recall compare to disease and gene/proteins NERs. At the same time such text processing tools (NormaGene and GOCat) can particularly address situations where training data are not available.

On the training data, our NER services obtained a precision of 77% for chemicals and 72% for disease and a recall 74% and 69% respectively. Then, we applied these settings to the official data. The results in Table 1 showed some overfitting phenomena, e.g entities detected by NER were rejected by the WSD from the final results.

Although current results seem suggesting that text mining can effectively help curators' tasks by providing access to more relevant contents, it is worth noticing that the effectiveness of some NERs can be obtained by specializing some of the general-purpose text mining tools. Finally, we plan to further investigate text-mining tools, which can be integrated into a biocuration process and can decrease time-consuming factor in situations where training data are not available.

## References

1. Wiegers T., (2012) Collaborative Biocuration-Text Mining Development Task for Document Prioritization for Curation. Proceedings of BioCreative 2012.
2. Ruch P. (2006) Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*. **22**(6):658-64
3. Lu Z., Wilbur W J., et al. (2011) The gene normalization task in BioCreative III. *BMC Bioinformatics*, **12**(Suppl 8):S2
4. Gobeill J, Pasche E., Vishnyakova D. and Ruch P (2013), Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. *Database*, doi: 10.1093/database/bat041
5. Vishnyakova D, Pasche E, Ruch P. (2012) Using binary classification to prioritize and curate articles for the Comparative Toxicogenomics Database. *Database (Oxford)*.;2012:bas050.
6. Pillet V, Zehnder M, Seewald AK, Veuthey AL, Petrak J. (2005) GPSDB A new database for synonyms expansion of gene and protein names. *Bioinformatics*. **21**(8):1743-4.
7. Davis AP, Wiegers TC, Murphy CG, and Mattingly CJ. (2011). The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database*, Oxford
8. Jimeno-Yepes A., McInnes B., Aronson A. (2011) Collocation analysis for UMLS knowledge-based word sense disambiguation. *BMC Bioinformatics*, **12**(Suppl 3):S4
9. Burgun, A., & Bodenreider, O. (2005, April). An ontology of chemical entities helps identify dependence relations among Gene Ontology terms. *In First Symposium on Semantic Mining in Biomedicine*.