Collaborative Semi-Automatic Annotation of the Biomedical Literature.

C. Cano, A. Labarga, A. Blanco
Dept. Computer Science and A.I.
University of Granada
Granada, Spain
ccano, armando@decsai.ugr.es, alabarga@googlemail.com

L. Peshkin
Dept. Systems Biology
Harvard Medical School
Boston, MA
peshkin@gmail.com

Abstract—The increasing availability of whole human genomes due to the improvements in high-throughput sequencing technologies makes the interpretation and annotation of data at a whole-genome scale more and more feasible. However, these tasks critically depend on the availability of knowledge already stored in databases or published in the scientific literature. This scenario requires new reliable and integrative information extraction systems to be available to the biomedical community.

In this work we present a hybrid approach for mining the wealth of knowledge stored in the scientific literature. This approach is based on the use of efficient text-mining tools in combination with highly accurate collaborative human curation. BioNotate-2.0 is an open source, modular friendly tool which implements a collaborative semi-automatic annotation platform in which human and automated annotations are efficiently combined. BioNotate-2.0 also builds upon the Semantic Web, facilitating the dissemination of annotated facts into other resources and pipelines.

BioNotate-2.0 allows any interested user to run his own annotation efforts or to contribute to an existing annotation project. To access and contribute to the BioNotate-2.0 annotation platform, please check: http://genome2.ugr.es/bionotate2/. BioNotate source code is available at: http://sourceforge.net/projects/bionotate/.

Keywords—Collaborative annotation, semi-automatic curation, text-mining, biomedical literature.

I. INTRODUCTION

Efficient access to information contained in the scientific literature is of great relevance for biomedical research. For example, the availability of raw whole-genome data in public repositories is quickly shifting the focus of the community towards the interpretation and annotation of whole genomes, which in turn cannot be accomplished without an efficient use of the wealth of knowledge in the literature. The biological literature also constitutes the main source of information for manually populating biological databases and developing bio-ontologies [1].

Extracting facts from the literature and making them accessible may be approached from two directions: first, manual curation efforts; second, text mining and Natural Language Processing (NLP) tools [2].

Some well known community-wide efforts for the evaluation of text-mining and information extraction systems in the biological domain are BioCreative (Critical Assessment of Information Extraction systems in Biology) challenge [3] and the NAACL BioNLP’09 Shared Task on Event Extraction [4]. As a direct result of these competitive evaluations, platforms integrating the best participant’s servers have been released (BioCreative Meta-Server [5] and U-Compare [6], respectively). A related initiative is the Collaborative Annotation of a Large Biomedical Corpus (CALBC, http://www.calbc.eu), which addresses the automatic generation of a very large, community-wide shared text corpus annotated with a dozen types of biomedical entities by different named entity recognition systems.

In theory, text mining is the perfect solution to transforming factual knowledge from publications into database entries. But computational linguists have not yet developed tools that can analyse more than 30% of English sentences correctly and transform them into a structured formal representation [7].

On the other hand, manually curated data is precise, because the curator is trained to inspect literature and databases, select only high-quality data, and reformat the facts according to the schema of the database. Additionally, curators select citations from the text as evidence for the identified fact.

The problem with curation of data is that it is time consuming and costly, and therefore has to focus on the most relevant facts. This compromises the completeness of the curated data, and curation teams are doomed to stay behind the latest publications. Therefore, an environment where manual curation and text mining can effectively and efficiently work together for rapid retrieval and analysis of facts with precise post-processing and standardization of the extracted information is highly desirable [7].

To partially overcome the limitations of manual curation, the notion of community annotation has recently started to be adopted in the biomedical field. The rationale behind this is to build knowledge upon the Social Web, allowing any interested user to collaborate in the annotation task and share their findings with the community. For instance, WikiProteins [8] and WikiGene [9] deliver appropriate environments in which it is possible to address the annotation of genes and proteins. GET-
Evidence http://evidence.personalgenomes.org/about is another wiki-based environment enabled by the Personal Genome Project (PGP) in which human curators can contribute to the annotation of functional roles to DNA variations in human genomes. CBioC [10] proposes a similar framework for annotating relationships between different biomedical entities (gene/gene, gene/disease and gene/biological process) using automated annotations as a starting point. However, CBioC does not allow the users to access the whole corpus of annotations until it is made publicly available by the CBioC team. In a previous work, we presented a collaborative annotation tool called BioNotate [15], which allows any interested user to run his own distributed annotation efforts, gathering contributions from curators worldwide and providing immediate access to the obtained annotations in stand-off format [15].

Another key point to consider is the current need of information integration in life sciences. This domain has turned into a very data-intensive field with huge amounts of very heterogeneous information being spread over many different databases and resources. Therefore, protocols and tools that contribute to the assembly, integration and publication of biomedical knowledge available at Web scale are highly desirable. Some efforts in this direction include the Bio2RDF project, aiming at applying Semantic Web technologies to publicly available databases [11]; and the concept of nanopublication [12] which aims to standardize a representation for core scientific statements with associated annotations.

In this work, we present an integrated approach to concept recognition in biomedical texts, which builds upon both the Semantic Web, which values the integration of well-structured data, and the Social Web, which aims to facilitate interaction amongst users by means of user-generated content. This approach extends our previous work since it combines automated named entity recognition tools with manual collaborative curation and normalization of the entities and their relations for a more effective identification of biological relations of interest. Furthermore, identified facts are converted to a standardized representation for making connections to other datasets and resources on the Semantic Web.

II. PROPOSED SOLUTION

The system we propose is called BioNotate-2.0. BioNotate-2.0 is an open-source modular friendly system which combines automated text annotation with distributed expert curation and serves the resulting knowledge in a Semantic-Web-accessible format. This system is composed of different modules which cover the different stages of the problem at hand: from the search of relevant texts, to the automatic identification of biomedical entities of interest (genes, proteins, diseases, SNPs, etc.), the distributed curation of facts and the publication of the annotations.

Figure 1. System architecture showing the different modules of BioNotate-2.0 and their interactions: 1) Corpora creation, 2) Automated annotation using text-mining tools, 3) Collaborative manual curation of annotated facts, 4) Publication of curated facts via a SPARQL endpoint.

Figure 1 shows the implemented pipeline and how the tools are interconnected.

The description of the pipeline and its composing modules follows.

Administration module.

The administration module allows users to generate the problem definition, the annotation schema and the format for the snippets that will be employed in the annotation and curation tasks. As part of the problem definition, administrators can also provide the references to the bio-ontologies or terminological resources which will be used to normalize the entities of interest in the annotation task. Finally, they are also allowed to upload their own corpus or create it by providing query terms and making use of the automatic retrieval module.

Automatic retrieval and annotation.

To generate a collection of relevant literature associated to the topic of interest, administrators can send a query to
Pubmed or Citexplore and load the retrieved abstracts to
the system for constituting an initial corpus. This set of
abstracts can be further refined, and additions and deletions
can be manually made to the set. This base collection is then
automatically annotated using Named Entity Recognition
(NER) tools and text-mining systems. This first annotation
eases latter manual curation efforts by providing fast and
moderately accurate results, and enables textual semantic
markup to be undertaken efficiently over big collections. Our
system currently uses Whatizit web services [13] to annotate
genes and diseases, and MutationFinder [14] to identify
SNPs. However, the modular-friendly design of BioNotate
allows to easily incorporate more tools and methods to the
pipeline.

**Collaborative manual curation**

The distributed manual curation module is based on
the first version of the tool: BioNotate-1 [15]. This tool
implements a client/server system: the server logic serves
snippets to the annotators and keeps track of the annotations
made by every user; the client logic displays the snippets and
their associated annotations and allows the curator to add,
change and remove annotations. Latter versions of the tool
allow users to run their own annotation efforts implementing
custom annotation schemas [16].

Figures 1 and 2 show screenshots of different implementa-
tions of the client annotation interface, where curators
can easily add and remove annotations on a given abstract.
Annotated pieces of text may be further refined or corrected
by different users, being the final annotation the results of
this collaborative process.

The annotation task should be conceived to be as sim-
ple as possible for the user, and may amount to voting
(agree/disagree) with respect to the previous automatic an-
notation. More sophisticated schemas may be easily im-
plemented from the administration panel, but may also
require the participation of curators with higher knowledge
on the problem addressed. Participation may be restricted to
specific curators or may be open to any interested user.

**Publication of annotated facts**

The implemented system allows to export the curated facts
as RDF or to publish it as a SPARQL endpoint. Resource
Description Framework (http://www.w3.org/RDF/) is a core
technology for the World Wide Web Consortium’s Semantic
Web activities that allows data to be disseminated across
different resources. The RDF data model represents arbitrary
information as a set of simple statements of the form
subject-predicate-object, where each entity has a (globally)
unique identifier. These identifiers allow everybody to make
statements about a given resource and, together with the
simple structure of the RDF data model, make it easy to
combine the statements from different resources and thus to
integrate information. SPARQL is a query language able to
retrieve and manipulate data stored in RDF.

The publication module allows to publish the RDF data to
any available RDF server, such as Sesame or Joseki. The ref-
ence implementation uses Joseki (http://www.joseki.org),
an HTTP engine that supports the SPARQL Protocol and
the SPARQL RDF Query language. It is part of Jena, a
Java framework for building Semantic Web applications
developed at HP labs.

In order to expose the data as Linked data, we use
Pubby http://www4.wiwiß.fu-berlin.de/pubby/, a system
developed by the Free University of Berlin that makes it
easy to turn a SPARQL endpoint into a Linked Data server.

**III. DISCUSSION**

Our platform provides web based tools for: 1) the creation
and automatic annotation of biomedical literature, 2) the
distributed manual curation of these pieces of text with
biomedical entities or facts of interest 3) the publication of
curated facts in semantically enriched formats, so annotated
facts can be easily incorporated into any other resource or
pipeline using Semantic Web standards.

Our implementation is based on several open-source
projects and allows disparate research groups to perform
literature annotation to suit their individual research needs,
while at the same time contributing to the large-scale ef-
fort. There are multiple levels of integration built into the
system. At one level, several annotators could collaborate
on processing statements from a single corpus on their own
server. At another level, multiple corpora could be created
different servers, and the resulting corpora could be
integrated into a single overarching resource.

The system we propose uses BioNotate-1 as the collabo-
rate annotation platform, which was shown be effective for
the annotation of a small corpus on interacting genes related
to autism, with averaged inter-annotator agreement rates over
75% [15]. Later versions of our modular-friendly tool has
shown improved versatility, leading to a wide interest in
diverse user communities. Two visible success stories for
BioNotate are the integration with:

- The Autism Consortium research efforts (see Figure
  2 or http://bionotate.hms.harvard.edu/autism), in which
  our system was used to ease the distributed curation of
  the results of gene-taggers for detecting genes involved
  in autism. In this instance, BioNotate is integrated with
  a different automated component, the LingPipe gene
  identification tool [17].
- The Personal Genome Project (PGP), in which our
  system is being used for the annotation of evidences
  associated to whole-genome variants from the GET-
  Evidence database: http://genom2.ugr.es/bionotate2/
  GET-Evidence/. Figure 3 shows a screenshot of this
  system.
IV. CONCLUSION

We have presented BioNotate-2.0, an open-source system which takes advantage of both manual and automated curation methods to generate reliable biological facts.

Our approach roughly consists on leaving the machine the aspects of this task where automated approaches do well, while the aspects where domain expertise and human skill are unsurpassed by machines are left to human annotation. Another challenge we addressed is to keep the generated knowledge available in machine-readable formats, so it can be disseminated into other tools and resources.

By assisting curators with automated annotations we expect their work to be considerably reduced in terms of time and complexity, since they have to correct previous annotations rather than create them from scratch.

Further work includes to quantify this time reduction in curation tasks to prove the convenience of this type of hybrid approaches and to implement “active learning” methods for prioritizing the annotation process, so that, at every point of the process, curated text contributes the most to improve the state of information about the rest of the knowledge base.

ACKNOWLEDGMENT

C. Cano and A. Blanco are supported by the projects P08-TIC-4299 of J. A., Sevilla and TIN2009-13489 of DGICT, Madrid. C. Cano is also supported by the GREIB project for young researchers (GREIB.PYR_2010_02).

REFERENCES


References:


