An annotation tool for academic literature processing

Molnár Zsolt[1], Polgár Tímea[1], Vincze Veronika[1,2,3]

[1]ScienceBoost Kft.

[2]Szegedi Tudományegyetem, Informatikai Tanszékcsoport

[3]MTA-SZTE Mesterséges Intelligencia Kutatócsoport

zsolt.molnar@hubscience.com; timea.polgar@hubscience.com

vinczev@inf.u-szeged.hu

Abstract: In this paper, we present our annotation tool that facilitates research and annotation work by quick, yet efficient literature processing. Our tool helps users create a unique and refined collection of linked information, which can lead to more effective and faster decisions in research. The tool is currently optimized for biomedical domain, but it can adapted to other academic fields with minimal efforts.

1 Introduction

Medical institutes usually store considerable amount of valuable information (patient data) as free text. Such information has a great potential in aiding research related to diseases or improving the quality of medical care. The size of document repositories makes automated processing in a cost-efficient and timely manner an increasingly important issue. The intelligent processing of clinical texts is the main goal of Natural Language Processing (NLP) [1] for medical texts.

In order to provide supervised NLP solutions for medical and clinical text mining, there is an intense need for annotated texts. There already exist a number of annotation tools within the community and on the market, which help the researcher collect and annotate relevant data. For instance, [4] and [5] provide a comparison of annotation tools for the biomedical domain, while [3] lists several annotation tools and compares them among parameters such as type of client, content of annotation, applied tags and attributes, annotation format etc.

Some of these annotation tools contain built-in machine learning (ML) methods e.g. for automatically annotating drugs in medical reports [6] or recognizing genes in biomedical articles. These annotators excel on the specific field they were developed for, but would provide poor performance on general texts. These are mostly based on supervised ML methods, in other words, training of the ML model requires domain-specific corpora manually annotated by experts, which can be very expensive. The high costs associated with this approach has led to a shift towards unsupervised or semi-supervised ML methods that, instead of manually labeled data, rely on human expertise encoded in expert-curated knowledge bases [2].

In this paper, we present our annotation tool that facilitates research and annotation work by quick, yet efficient literature processing. With our tool, users can create a unique and refined collection of linked information, which can yield more effective and faster decisions in research.

2 The annotation tool

Our work is designed to facilitate research by quick, yet efficient literature processing. This is the driving force behind our work to provide a teamwork-driven, AI-powered literature survey primarily in life sciences but our solutions can be adapted to any academic field.

The tool can be used online from a browser and it helps researchers build up their own knowledge graph over a specific topic while reading scientific publications. In order to reach this goal, an expert needs to do the first steps manually (manual annotation), and based on these annotated data, the algorithm can be automatically trained, hence the rest of the documents are annotated automatically. Documents and the manual annotation task can be shared with other colleagues (see below) to accelerate the training procedure.

In the following subsections, we report the typical annotation process and functionalities of our tool.

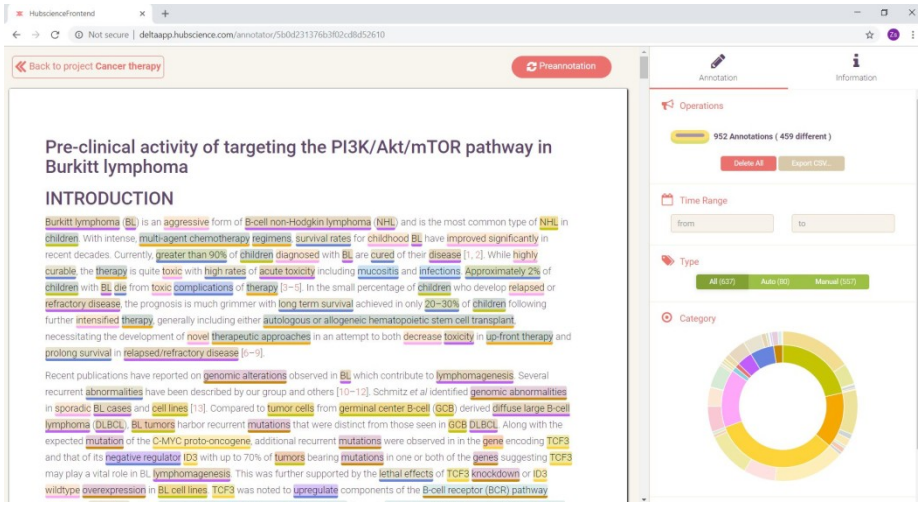

Fig. 1: The annotation interface.

2.1 Document Management

In order to be able to annotate, first there is a need for available documents. Documents can be added to the tool in multiple formats:

- Online Journal: The website reference (URL from the browser address bar) can be added and the system will download and insert the full text article for annotation.
- Upload: local PDF or HTML files can also be uploaded from the local computer.
- New Document: new notes or copy-paste fragments from any source can be also added to the user interface for annotation.

2.2 Annotation

Documents can be annotated manually and automatically as well. During manual annotation, the annotation categories are visualized above the selected text, and the appropriate one can be chosen by clicking on it.

The preannotation icon will allow users to tag all the text elements, which have been already added to the system. The system can also manage dictionaries containing lists of text elements, these can also be added to the system.

Dictionaries are built up from wordcards and wordpacks (see Fig. 2). The wordcard contains the term, its synonyms, its abbreviations as the most important properties. Additional properties can also be added if needed. Wordcards can be organized and grouped into wordpacks, i.e. lists of word cards that belong together. Typically a wordpack is domain specific. Users can also define their own wordpacks but the tool inherently contains some biomedical wordpacks.

Within the texts, an annotated item can be connected to several other annotations (see Fig. 3). They do not necessarily need to be in the same text, they can be linked to each other in different documents too. The system is able to build up a knowledge graph over a topic (see Fig. 4). The knowledge graph is a bunch of connected information that will help the researcher derive new knowledge on the basis of existing annotations.

2.3 Category system

Our tool offers an annotation category system for biomedical text annotation. However, this category system is fully customizable and can be modified according to the needs and the topic that user is interested in. New categories and new subcategories can be added, edited, or removed (see Fig. 5).

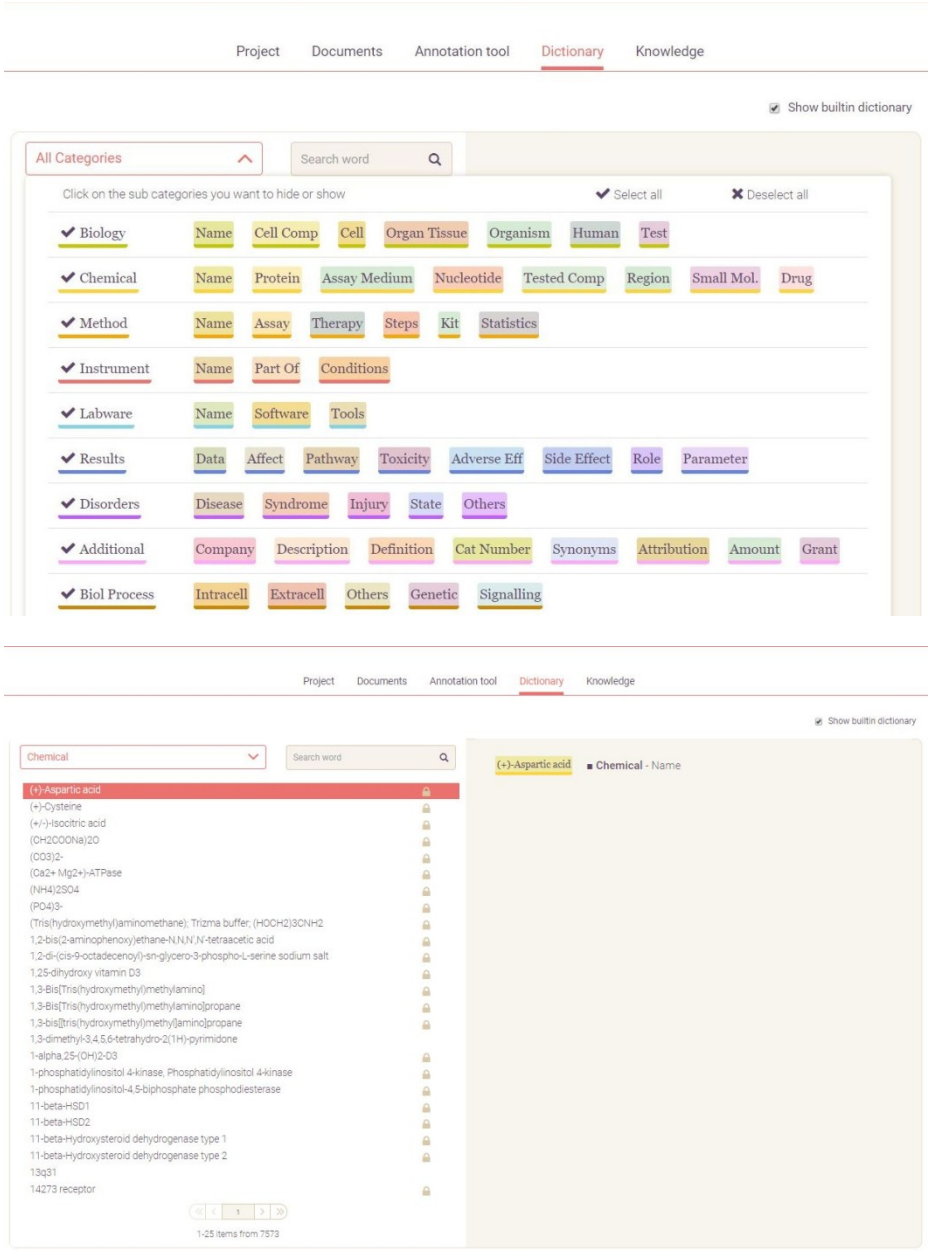

Fig. 2: Dictionary panels.

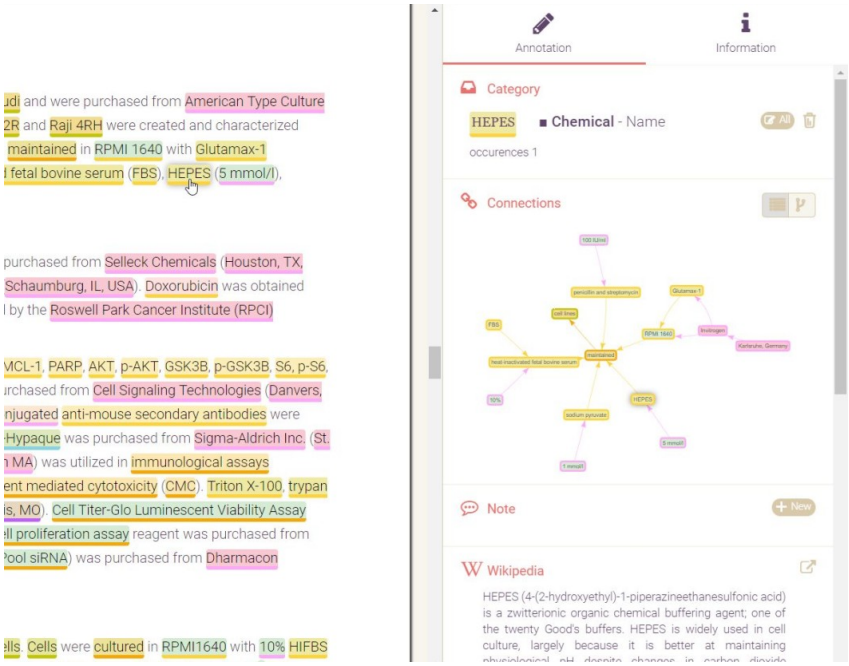

Fig. 3: Connections in an annotated documents.

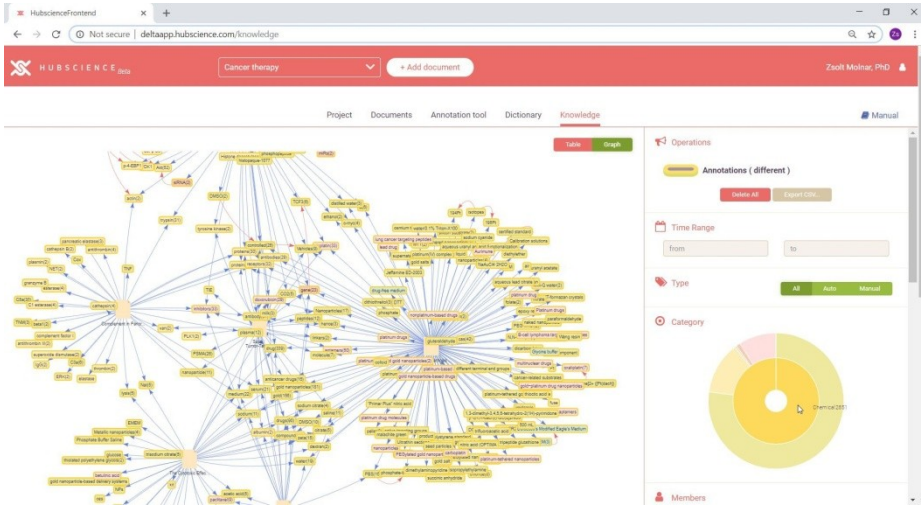

Fig. 4: A knowledge graph.

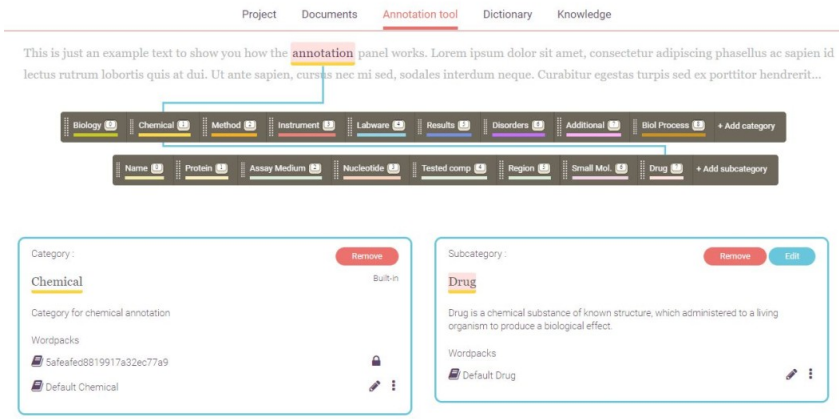

Fig. 5: A category panel.

2.4 Working with projects

Documents can be organized into projects. Working with projects allows the users to perform the analysis in an organized way and the machine learning algorithm can be trained directly on topics.

The main advantages of using projects can be summarized as follows:

- Projects allow collaborative work, i.e. several colleagues can work together on the same set of documents
- Documents belonging to the same topic can be organized within one project
- The annotation category system can be customized for each project
- Special dictionaries can be employed for each project
- Annotation statistics/analytics can be aggregated for the whole project, i.e. documents belonging to the same topic

Within projects, members can have multiple roles: owner, admin, or member role. The owner can do everything, admin can invite others and edit the category system, the member can only annotate.

2.5 Info Panel

The info panel shows relevant general statistical and meta-information of the document (see Fig. 6). In addition, during annotation the selected word or text element is looked up in Wikipedia or Wikidata providing more information about the specific terms helping students or users inexperienced at a given topic (see Fig. 7).

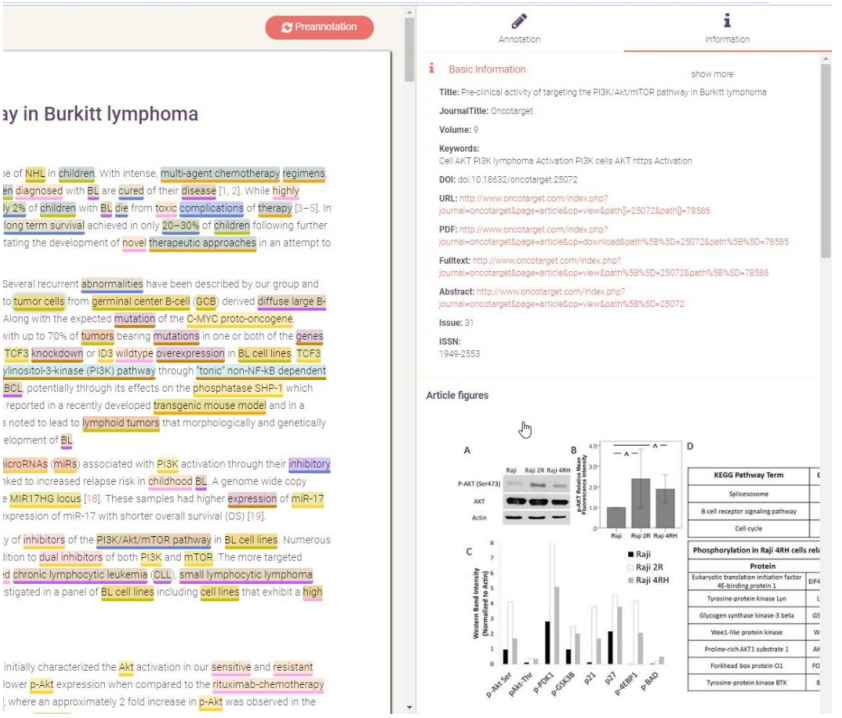

Fig. 6: An info panel with basic information on the article.

2.6 Other functionalities

In addition to token-level annotations, users can also apply labels for the whole document. For instance, users can rate documents on a scale of 1 to 10 by relying on their own experience in evaluating scientific information (see Fig. 8). This might help other users to decide whether they need that document or not.

Relevance scores can also be added to the documents (High, Medium or Low relevance) by relying on the user's criteria such as how relevant they can be for them in a given project or topic.

Moreover, any other notes, comments or reminders can be directly added to the annotated information and web contents may also be linked to it.

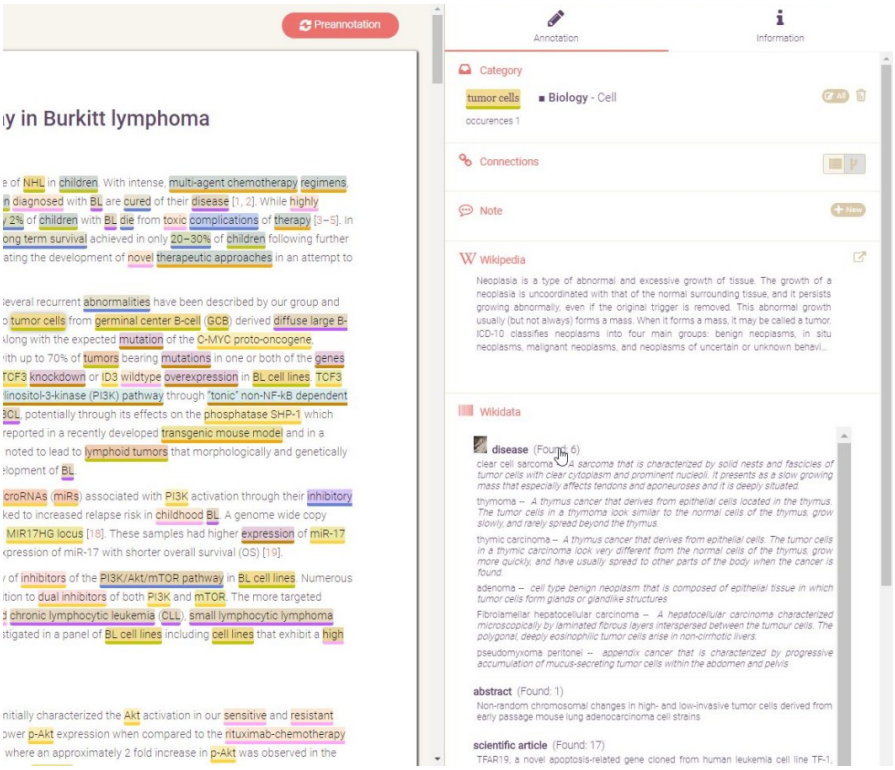

Fig. 7: An info panel with information from Wikipedia.

3 Advantages of the tool

We believe that our annotation tool can be fruitfully exploited in several fields of research, and as such, several groups of users can profit from it. For instance, it can be used as supporting material for annotation for researchers and students to stay competitive at universities and research institutions. The tool can also provide an effective way for sharing an annotated literature survey between team members of the same research group, leveraging the power of teamwork.

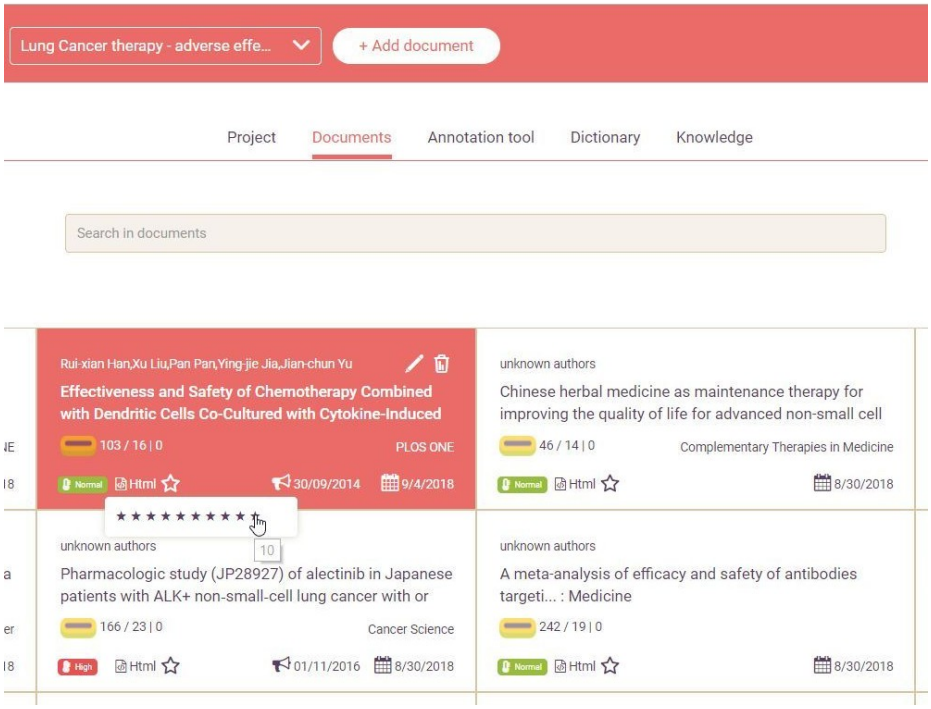

Fig. 8: Rating an article.

The tool facilitates smart ontology building for specific areas. It also offers a way to personalize the annotation labels, hence data curation may become easier and faster. Finally, it offers an easy to use software interface for visualizing annotations and their relations, thus enabling the discovery of novel academic achievements.

4 Availability

The basic version of the tool is available for everyone free of charge at our website (www.hubscience.com).

The tool is currently optimized for biomedical domain, but it can be adapted to other academic fields with only minimal efforts.

References

1. Ananiadou, S., Mcnaught, J.: Text Mining for Biology and Biomedicine. Artech House, Inc., Norwood, MA, USA (2005)

2. Chasin, R., Rumshisky, A., Uzuner, O., Szolovits, P.: Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. Journal of American Medical Informatics Association, Vol. 21 (2014) 842-849
3. http://knot.fit.vutbr.cz/annotations/comparison.html
4. Médigue, C., Moszer, I.: Annotation, comparison and databases for hundreds of bacterial genomes. Research in microbiology, Vol. 158, No. 10 (2007) 724-736
5. Neves, M., Leser, U.: A survey on annotation tools for the biomedical literature. Briefings in bioinformatics, Vol. 15, No. 2 (2012) 327-340
6. Tikk, D., Solt, I.: Improving textual medication extraction using combined conditional random fields and rule-based systems. Journal of American Medical Informatics Association, Vol. 17 (2010) 540-544