An Open-Source Annotation Tool for Collaboratively Annotating Biomedical Documents*

Ornella Irrera, Fabio Giachelle and Gianmaria Silvello

Department of Information Engineering, University of Padua, Padua, Italy

Abstract

In recent years there has been a growing interest in developing techniques to effectively extract knowledge from biomedical textual documents. Many solutions rely on Named Entity Recognition and Linking (NER+L) which consists in detecting entities in text and disambiguating them through the use of knowledge bases. Despite its potential, applying this approach to the biomedical domain is limited by the lack of large annotated corpora useful to train Machine Learning (ML) models. Nowadays, it is difficult to find large sets of annotated data covering a wide range of biomedical sub-domains: the creation of annotated corpora in fact, is an expensive and time-consuming task usually performed by experts.
To address this problem and ease and speed up the annotation process, we propose MedTAG, a web-based, collaborative, customizable annotation tool for biomedical documents; it is platform-independent and it provides a straightforward installation procedure.

Keywords

Bio-medical annotation tool, Annotated corpora creation, Digital health, Semantic annotation

1. Introduction

The availability of biomedical data stored in electronic form faced an exponential growth over the last decade. As a consequence, the need to preserve and curate biomedical data is growing [2]. The knowledge contained in biomedical documents in fact, is a central asset to derive new scientific insights in research domains such as pathology, genetics and epidemiology [3, 4].

Extracting knowledge and relevant information from medical data is not a trivial task. First, the largest part of biomedical data are stored in an unstructured textual format not easily machine-readable. Second, biomedical documents have plenty of abbreviations, symbols and terms that need to be disambiguated [5]. In this context, NER+L techniques are central to effectively process biomedical data. NER+L is a specific Natural Language Processing (NLP) task whose aim is to extract entities from the textual document and link them to concepts belonging to a knowledge base [6].

* This is an extended abstract of [1]

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy

ornella.irrera@unipd.it (O. Irrera); fabio.ghiachelle@unipd.it (F. Giachelle); gianmaria.silvello@unipd.it (G. Silvello)

http://www.dei.unipd.it/~irreraorne/ (O. Irrera); http://www.dei.unipd.it/~giachell/ (F. Giachelle); http://www.dei.unipd.it/~silvello/ (G. Silvello)

0000-0003-2284-5699 (O. Irrera); 0000-0001-5015-5498 (F. Giachelle); 0000-0003-4970-4554 (G. Silvello)

A common limitation related to the application of NER+L techniques to the biomedical domain is the lack of large annotated corpora needed to evaluate systems and training ML models. Creating large annotated corpora, able to cover several biomedical sub-domains, is a complex and demanding task which requires the supervision of experts [6]. For this reason manual annotation is still a central task in this domain. In recent years, several annotation tools have been developed to support human annotators in the creation of annotated corpora [7, 8, 9, 10, 11, 12, 13, 14, 15]. A recent survey [16] about both general-purpose and biomedical annotation tools, pointed out that several systems lack of important features such as the support for collaborative annotations and easy and user-friendly installation and configuration procedures.

To address these problems, we present MedTAG, a web-based, collaborative, customizable annotation tool for biomedical literature. MedTAG key features are the following. (i) *Inter Annotator Agreement (IAA)* based on majority vote; (ii) *multilingual support*: the same document can be uploaded and annotated in different languages; (iii) *support for user roles*: roles are based on the level of expertise of the users who share MedTAG; (iv) *support for document-level annotations*; (v) *support for mention-level annotations*; (vi) *user annotation statistics*; (vii) *documents' annotation statistics*: for each biomedical document an overview of its annotations is provided; (viii) *support for automatic annotations*: we rely on *Semantic knowledge Extractor Tool (SKET)*[1] for the automatic annotation process which is currently limited to three cancer use-cases: lung, colon, cervix cancer; automatically created annotations can be manually edited by the annotators; (ix) *PubMed integration*; (x) *collaborative facilities*: users who share the same instance of MedTAG can visualize the annotation of one team mate of their choice; (xi) *support for ontologies*; (xii) *support for multiple file formats*: annotations can be downloaded in several file formats (JSON, CSV, BioC/XML, BioC/JSON).
The rest of the paper is organised as follows: in Section 2 we present MedTAG and we describe the main aspects of the tool; in Section 3 we compare MedTAG to other biomedical annotation tools; in Section 4 we draw the conclusions.

2. MedTAG

MedTAG is a web-based annotation tool distributed as a Docker container. Docker ensures portability, code isolation, dependencies packaging and a fast installation procedure: MedTAG installation can be launched with a single command and takes a few minutes to complete.
The source code and the documentation are publicly available at: https://github.com/MedTAG/medtag-core.
We refer to the Github repository cited above and to [1] for a more exhaustive discussion.

Architecture. MedTAG architecture relies on: (i) a PostgreSQL relational database where annotated data are stored, (ii) a back-end realized with *Django* Python framework for the orchestration of the requests, (iii) a front-end realized with *React.js*, *HTML5* and *CSS3*.

[1] https://github.com/ExaNLP/sket

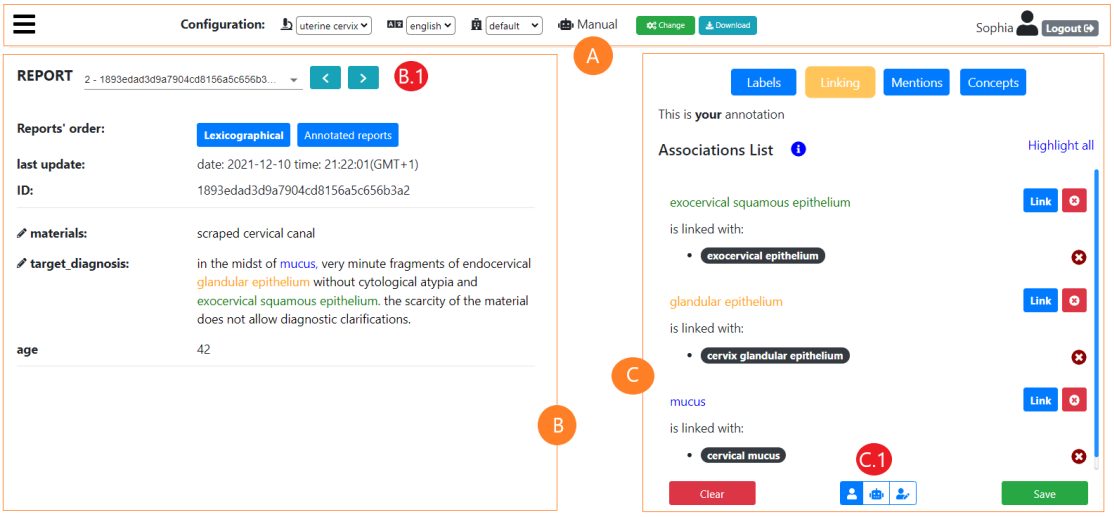

Figure 1: MedTAG user interface. It is reported an example of linking annotation. Each mention is characterized by a different color in order to be easily detected both in the original textual report (on the left) and in the list of annotations (on the right). The section **A** (*Settings and Download*) allows users to download their annotations and filter reports according to four criteria (use-case, language, institute, annotation mode). The section **B** (*Report*) identifies the textual report while the section **C** (*Annotation*) identifies the current annotation. **B.1** (*Report navigation panel*) and **C.1** (*Annotation panel*) allow users to navigate between reports and check other members' annotations respectively.

Configuration. MedTAG was designed to be easily configurable via a dedicated web-interface. At the moment of the configuration, the user is asked to provide the collection of clinical documents (or *reports*) to be annotated, the labels needed to perform document-level annotation and the concepts belonging to one or more ontologies. All the files must be uploaded in CSV format. In addition, it is possible to specify the parts of the uploaded reports where annotation is allowed. The file containing clinical reports can be replaced by (or provided with) a CSV containing a list of PubMed articles' identifiers. MedTAG automatically extracts the identifiers and downloads the title, the abstract and some additional information of each article. Annotation is allowed on *title* and *abstract* sections. MedTAG supports the upload of new batches of reports, new sets of labels and concepts at will, without losing the existing annotations. In order to make the configuration easier and faster, MedTAG offers downloadable CSV templates and provides the user with information about what changes need to be applied to the uploaded CSVs to make them comply with the required format.

User interface. In Figure 1, the MedTAG user interface is illustrated. The user interface we propose, its layout and components have been discussed with physicians and experts in the digital pathology domain.

We organized the MedTAG user interface into three main sections: *Settings and Download* (**A** in Figure 1), *Report* (**B** in Figure 1) and *Annotation* (**C** in Figure 1).

The *Settings and Download* section, placed at the top of the web page, allows the user to change the current parameters' setting; it is possible to change in this order: (i) the clinical case, (ii) the language of the clinical reports, (iii) the institute (or hospital) where the reports have been produced and (iv) the annotation mode (it can be set either to *Manual* or *Automatic* if the

reports have been automatically annotated). The button *Download* allows the user to download the annotations in JSON, CSV, BioC/XML and BioC/JSON file formats. The leftmost button allows the user to navigate to other web-pages. The rightmost button instead, allows the user to logout.

The *Report* section contains the textual report to annotate and some other relevant information such as the date of the last annotation performed by the user, the report's identifier and the report's translations (available if the report has been uploaded in different languages). The *Reports navigation panel* (**B.1** in Figure 1) allows the users to navigate between reports. Clicking on the *next* and *previous* buttons, or using the keyboard arrows, it is possible to move to the next or previous reports. To skip to a specific report, the drop-down list allows the user not only to search for the desired report from its identifier, but also to check what reports have not been annotated yet (they are marked in boldface).

The *Annotation* section contains the annotation created by the user for the clinical report examined. MedTAG offers four annotation types: *Labels, Concepts, Mentions* and *Linking*. The first two are *document-level* annotations: in this case the annotation refers to the entire textual content. The latter two are *mention-level* annotations: in this case the annotation refers to one or more text spans (mentions) in the original textual content. *Labels annotation* consists in associating one or more labels to the textual report; a label is a property that describes the report. *Concepts annotation* consists in associating one or more ontological concepts to the report. *Mentions annotation* consists in identifying entity mentions. In order to identify a mention, the user can either click on the first and on the last words composing the mention, or click on every single adjacent word of the mention in the textual report on the left. The identified mentions are characterized by different colors so to be immediately detected. In Figure 1 the annotated mentions are highlighted in blue, yellow and green respectively. *Linking* consists in associating one or more ontological concepts to a mention. Ontological concepts in *Linking* and *Concepts* annotation can be added via a drop-down menu with auto-completion facilities, this allows the user to type the desired concept without examining all the ontological concepts. In Figure 1, each mention has been linked to its corresponding concept.
Each annotation can be either saved or deleted by clicking on *Save* and *Clear* button respectively. Nevertheless, the annotation is automatically saved when the report or the annotation type change. Annotators can benefit from annotations created by other users they share MedTAG with. The *Annotation Panel* (**C.1** in Figure 1) provides the annotation created by the logged in user (left-side button), the automatically created one (button in the middle, if the automatic annotation is not available this button is disabled) and the one performed by another user of choice (right-side button). The automatic annotation of the report and the one of the team mate are read-only.

3. Evaluation

Firstly, we evaluated MedTAG from a *qualitative* perspective, which allowed us to compare its functionalities with those of other seven annotation tools exploited in the biomedical domain. Then, we evaluated MedTAG from a *quantitative* perspective which allowed us to study the performances of MedTAG compared to those of other four biomedical annotation tools.

In the qualitative analysis we compared MedTAG to: BioQRator [7], ezTag [8], TeamTat

[9], MyMiner [10], tag-tog [11], brat [12] and INCEpTION [13]. We compared the eight tools according to 22 (over the 26 proposed) criteria described in [16], a review concerning general-purpose and biomedical annotation tools. The selected criteria can be grouped according to three categories: (i) *Data*: it concerns the configuration of the tool and the format of input and output files, (ii) *Functionality*: it concerns the functionalities provided by the tools such as the support for multiple annotation types, the integration with PubMed or the support for IAA, (iii) *Technical*: it concerns technical details such as the ease of installation or the availability of the source-code.
MedTAG turned out to satisfy almost all the selected criteria (20 criteria over 22): the annotation of overlapping mentions and the relationships annotation are the only unsatisfied criteria.

In the quantitative analysis instead, we compared MedTAG performances to those of other four web-based publicly available annotation tools: MyMiner [10], tag-tog [11], ezTag [8], TeamTat [9]. To this aim, we considered the publicly available instance of MedTAG available at: http://examode.dei.unipd.it/exatag/. We compared the five tools on two annotation types, *Labels* and *Mentions* annotation, and we studied the performances in terms of (i) the elapsed time and the (ii) the number of clicks required to annotate 100 clinical reports randomly taken from a real set of reports about colon cancer. From the experimental results[2], it turned out that MedTAG achieved high performances in terms of elapsed time: to perform label annotation of 100 documents MedTAG has been 4.5 times faster than tag-tog (MedTAG took 46 seconds while tag-tog 206). For what concerns the number of clicks instead, MedTAG turned out to be less efficient than the other tools, especially in *Mentions* annotation (MedTAG required 519 clicks while TeamTat only 307). The reported results are the average among 40 runs. The five tools were tested relying on automatic agents developed using the *Selenium* Python library[3].

4. Final remarks

The lack of large annotated corpora hinders the development of NER+L techniques based on ML. Creating annotated corpora is an expensive demanding task performed by experts in the biomedical domain. In order to ease and speed up the annotation process we presented MedTAG, a dockerized, web-based biomedical annotation tool.

Thanks to its Docker distribution, MedTAG is portable and it is easy to be installed. MedTAG is customizable with documents, labels and concepts, not necessarily tied to a single ontology. The PubMed integration allows the user to upload a list of PubMed articles' identifiers whose main information are automatically downloaded. MedTAG provides four different annotation types: two of them are document-level annotation types, the others are mention-level annotation types. The annotations of each type can be downloaded in four different file formats: CSV, JSON, BioC/XML and BioC/JSON. MedTAG supports collaborative annotations: users can visualize other members' annotations and copy them in their own profile.

In order to improve MedTAG, we plan to add the support for relationships annotation and overlapping mentions. Moreover, we plan to extend the use-cases available for the automatic annotation.

[2] https://github.com/MedTAG/medtag-core#medtag-benchmark
[3] https://www.selenium.dev/

Acknowledgments

This work was supported by the ExaMode Project, as a part of the European Union Horizon 2020 Program under grant 825292.
The authors wish to thank Stefano Marchesin for the work on SKET which has been integrated in this work.

References

[1] F. Giachelle, O. Irrera, G. Silvello, Medtag: A portable and customizable annotation tool for biomedical documents, BMC Medical Informatics and Decision Making (2021) in print.
[2] T. B. Murdoch, A. S. Detsky, The inevitable application of big data to health care, Jama 309 (2013) 1351–1352.
[3] G. Gorrell, X. Song, A. Roberts, Bio-yodie: A named entity linking system for biomedical text, arXiv preprint arXiv:1811.04860 (2018).
[4] D. M. Trifiletti, T. N. Showalter, Big data and comparative effectiveness research in radiation oncology: synergy and accelerated discovery, Frontiers in oncology 5 (2015) 274.
[5] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, et al., Overview of biocreative ii gene normalization, Genome biology 9 (2008) 1–19.
[6] J. Jovanović, E. Bagheri, Semantic annotation in biomedicine: the current landscape, Journal of biomedical semantics 8 (2017) 1–18.
[7] D. Kwon, S. Kim, S.-Y. Shin, W. J. Wilbur, Bioqrator: a web-based interactive biomedical literature curating system, in: Proceedings of the Fourth BioCreative Challenge Evaluation Workshop, volume 1, 2013, pp. 241–246.
[8] D. Kwon, S. Kim, C.-H. Wei, R. Leaman, Z. Lu, eztag: tagging biomedical concepts via interactive learning, Nucleic acids research 46 (2018) W523–W529.
[9] R. Islamaj, D. Kwon, S. Kim, Z. Lu, Teamtat: a collaborative text annotation tool, Nucleic acids research 48 (2020) W5–W11.
[10] D. Salgado, M. Krallinger, M. Depaule, E. Drula, A. V. Tendulkar, F. Leitner, A. Valencia, C. Marcelle, Myminer: a web application for computer-assisted biocuration and text annotation, Bioinformatics 28 (2012) 2285–2287.
[11] J. M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik, G. H. Millburn, B. Rost, F. Consortium, et al., tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles, Database 2014 (2014).
[12] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.
[13] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, I. Gurevych, The inception platform: Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, 2018, pp. 5–9.

[14] S. M. Yimam, I. Gurevych, R. E. de Castilho, C. Biemann, Webanno: A flexible, web-based and visually supported system for distributed annotations, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2013, pp. 1–6.
[15] S. Dobbie, H. Strafford, W. O. Pickrell, B. Fonferko-Shadrach, C. Jones, A. Akbari, S. Thompson, A. Lacey, Markup: A web-based annotation tool powered by active learning, Frontiers in Digital Health 3 (2021).
[16] M. Neves, J. Ševa, An extensive review of tools for manual annotation of documents, Briefings in bioinformatics 22 (2021) 146–163.