DocTAG: A Customizable Annotation Tool for Ground Truth Creation

Fabio Giachelle[0000−0001−5015−5498], Ornella Irrera[0000−0003−2284−5699], and Gianmaria Silvello[0000−0003−4970−4554]

Department of Information Engineering, University of Padua, Padua, Italy
{fabio.giachelle, ornella.irrera, gianmaria.silvello}@unipd.it

Abstract. *Information Retrieval* (IR) is a discipline deeply rooted on evaluation that in many cases relies on annotated data as ground truth. Manual annotation is a demanding and time-consuming task, involving human intervention for topic-document assessment. To ease and possibly speed up the work of the assessors, it is desirable to have easy-to-use, collaborative and flexible annotation tools. Despite their importance, in the IR domain no open-source fully customizable annotation tool has been proposed for topic-document annotation and assessment, so far. In this demo paper, we present DocTAG, a portable and customizable annotation tool for ground-truth creation in a web-based collaborative setting.

Keywords: Annotation tool · Passage annotation · Evaluation · Ground-truth creation

1 Motivation and Background

Ground-truth creation is an expensive and time-consuming task, involving human experts to produce richly-annotated datasets that are fundamental for training and evaluation purposes. In IR, gold standard relevance judgments are essential for the evaluation of retrieval models. The creation of experimental collections in the context of large scale evaluation campaigns (e.g., *Text Retrieval Conference* (TREC)[1] and *Cross Lingual Evaluation Forum* (CLEF)[2]) requires a huge deal of human effort to manually create high quality annotations. To this aim, evaluation campaigns usually adopt custom made annotation and assessment tools to support human assessors and ease their workload [1, 7–9, 13, 15]. Since, the relevance assessment process is usually carried out in a short time, an effective annotation tool can be of great help to speed up the overall process or at least to reduce the annotation bargain. However, in the typical IR scenario, it is common to develop a custom annotation tool for a specific evaluation task or campaign; available annotation tools are tailored for specific tasks, thus making them difficult to reuse for others without a significant overhaul.

[1] https://trec.nist.gov
[2] http://www.clef-initiative.eu

The annotation software currently available [10] can be divided into general-purpose and domain-specific tools. General-purpose ones provide a set of common features that cover most of the typical annotation scenarios and use-cases [4, 14,17] but require a great deal of customization to fit in a domain-specific setting – e.g., the typical topic-document pair is not handled by these systems. In contrast, domain-specific tools provide ad-hoc functionalities that meet the needs of very specific fields, focusing especially in the biomedical domain [2,3,5,6,11,12].

A recent exhaustive comparison of the major annotation tools [10] points out that choosing the best suitable tool is a demanding task, since each one presents specific advantages and disadvantages in terms of the functionalities provided. In addition, even the most comprehensive tool may present drawbacks such as a tricky installation procedure, no support for online use or a complex user interface. In addition, adapting existing tools not designed for a specific domain is a burdensome process requiring not naive programming skills.

For these reasons, we propose DocTAG, an annotation tool designed specifically for the typical IR annotation tasks. DocTAG provides a streamlined user interface in a collaborative web-based setting. DocTAG provides several features to support human annotators, including: (i) topic-document annotation with customized labels (binary or graded relevance judgements or other custom labels for instance for sentiment/emotion classification) or based on custom defined ontological concepts; (ii) passage-level annotation; (iii) inter-annotation agreement via majority vote; (iv) collaborative facilities (e.g., annotation sharing between assessors); (v) annotation statistics; (vi) responsive interface for long document visualization; (vii) download of ground-truths in CSV and JSON formats; (viii) customizable parsing and ingestion of document corpus, runs and topic files in several formats; (ix) annotation highlighting; (x) topic-document matching words emphasized (i.e. TF-IDF weighted highlight of the words present in the topic-document pair) and (xi) multi-lingual support – i.e. users can annotate the same topic-document pair in different languages (if provided). In case of multiple languages, the documents are grouped by language, so that users can search and filter them accordingly.

DocTAG is portable since it is provided as a Docker container, that ensures code isolation and dependencies packaging. Thus, it can be either installed as a local Webapp or deployed in a cloud container orchestration service.

The rest of the paper is organized as follows: Section 2 describes the annotation tool and the main aspect of the demo we present, and Section 3 draws some final remarks.

2 DocTAG

DocTAG is a web-based annotation tool specifically designed to support human annotators in the IR domain. The DocTAG source code is publicly available at `https://github.com/DocTAG/doctag-core`. In addition, to present the main DocTAG features, we provide a demonstration video[3] and a step-by-step "tuto-

[3] `https://bit.ly/3pqwHtF`

rial" section, included in the DocTAG web interface. DocTAG allows the users to customize several annotation aspects including the set of labels or ontological concepts used for both document-level and passage-level annotation, and the document fields to be visualized and/or annotated. The users can specify all the setting parameters via a wizard configuration procedure[4]. There is no limit to the number of labels and concepts that can be used for the annotation. Since the concepts are custom, the users can specify also concepts defined in external ontologies and terminological resources.

In addition, the configuration interface allows the users to specify (i) the document corpus to be annotated in CSV or JSON format; (ii) the topic files in CSV or JSON format and (iii) the runs (to build the pool to be annotated) in CSV, JSON or plain text.

Architecture and implementation. DocTAG architecture consists of (i) a web-based front-end interface built with *React.js*; (ii) a back-end for REST API and services built with the Python web framework Django; (iii) a PostgreSQL database to guarantee the persistence of the annotated data.

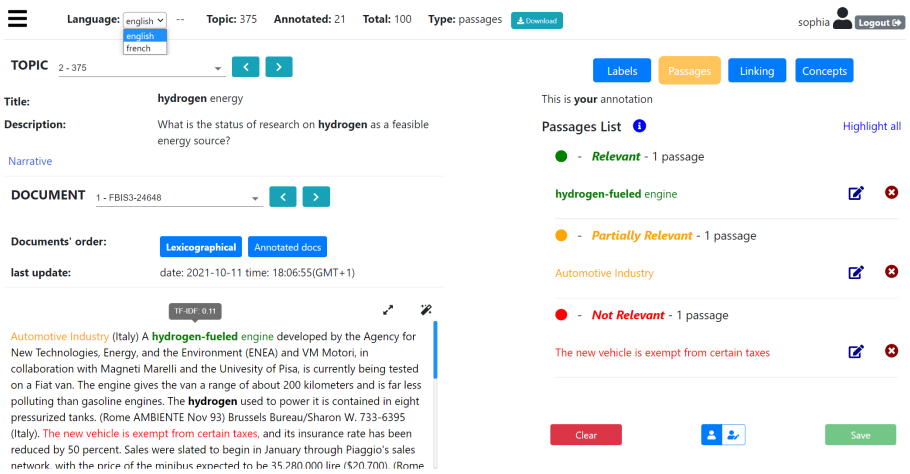

Fig. 1. DocTag interface, with the *Passages*-level annotation annotation mode (yellow button) active.

User interface and interaction. Figure 1 shows the main DocTag annotation interface. In the upper part, the header shows the current annotation statistics (i.e., the number of annotated documents for the selected topic out of the total number of documents) for each annotation type (i.e. labels, passages, concepts and linking). The header includes also the button to download

[4] `https://github.com/DocTAG/doctag-core#customize-doctag`

the ground-truth in CSV and JSON formats. On the left side of the header, the menu button allows us to access the DocTAG options and settings (e.g. configuration, inter-annotator agreement and annotation statistics). On the right side instead, the user section shows the username of the current user and the log out button. The interface body is divided in two sections: the document and the annotation sections. The first one (left-side), presents the information concerning both the current topic (e.g. title and description) and the document (e.g. document identifier and text). To switch between documents, users can use either the *next-previous* buttons or the keyboard arrows. The annotation section (right-side), shows the annotations (e.g., labels and concepts) made for the selected document. The users can visualize their own annotations and also the ones made by other annotators, by clicking on the user icons in the lower part of the right-side of the interface. In addition, assessors can import and edit (in their own profiles) the annotations made by other assessors, by clicking on the *Upload and transfer* menu option.

DocTag users can use of four annotation modes: (i) *Labels* where each topic-document pair can be associated with a label (only a single label is allowed since a document cannot be marked, for instance, as *relevant* and *not relevant* at the same time); (ii) *Passages* where document passages can be marked with labels (one label per topic-passage pair) highlighted with different colors; (iii) *Linking* where each passage can be linked to user-defined or ontological concepts (one or many) and (iv) *Concepts* where each document can be associated with several user-defined or ontological concepts. Figure 1 shows the *Passages* annotation mode with several passages annotated. For instance, *hydrogen used to power* (highlighted in green) is labelled with *Relevant* for the considered topic. All the passages marked with the same label are highlighted with a label-specific color to facilitate their recognition in the text. To quickly annotate long passages, users can click on the first passage word and on the last word, DocTAG automatically identifies the words in-between as a unique passage. By default, DocTAG provides automatic saving; nevertheless, manual saving is allowed as well, via *Save* button. Finally, to remove all the annotations made for the current annotation mode, users can click on the *Clear* button.

3 Final Remarks

In this paper, we present DocTag, a web-based annotation tool specifically designed to ease the ground-truth creation process and support human annotators, with regards to the IR domain. DocTag is an open-source, portable and customizable annotation tool that aims to be a reusable solution, for instance, in the context of IR evaluation campaigns. For the demo, we plan to showcase the annotation tool instantiated with the TIPSTER document collection along with the TREC 7 topics [16], since it is a very well-known collection in the IR domain. As future work, we plan to conduct a user study to improve the annotation interface, in terms of accessibility and inclusive design.

References

1. Biega, A.J., Diaz, F., Ekstrand, M.D., Kohlmeier, S.: Overview of the TREC 2019 fair ranking track. CoRR **abs/2003.11650** (2020)
2. Cejuela, J.M., McQuilton, P., Ponting, L., Marygold, S.J., Stefancsik, R., Millburn, G.H., Rost, B.: tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. Database J. Biol. Databases Curation **2014** (2014)
3. Dogan, R.I., Kwon, D., Kim, S., Lu, Z.: Teamtat: a collaborative text annotation tool. Nucleic Acids Res. **48**(Webserver-Issue), W5–W11 (2020)
4. Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9. Association for Computational Linguistics (June 2018)
5. Kwon, D., Kim, S., Shin, S., Chatr-aryamontri, A., Wilbur, W.J.: Assisting manual literature curation for protein-protein interactions using bioqrator. Database J. Biol. Databases Curation **2014** (2014)
6. Kwon, D., Kim, S., Wei, C., Leaman, R., Lu, Z.: eztag: tagging biomedical concepts via interactive learning. Nucleic Acids Res. **46**(Webserver-Issue), W523–W529 (2018)
7. Lin, J., Mohammed, S., Sequiera, R., Tan, L., Ghelani, N., Abualsaud, M., McCreadie, R., Milajevs, D., Voorhees, E.M.: Overview of the TREC 2017 real-time summarization track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017. NIST Special Publication, vol. 500-324. National Institute of Standards and Technology (NIST) (2017)
8. Lin, J., Roegiest, A., Tan, L., McCreadie, R., Voorhees, E.M., Diaz, F.: Overview of the TREC 2016 real-time summarization track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016. NIST Special Publication, vol. 500-321. National Institute of Standards and Technology (NIST) (2016)
9. Lin, J., Wang, Y., Efron, M., Sherman, G.: Overview of the TREC-2014 microblog track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014. NIST Special Publication, vol. 500-308. National Institute of Standards and Technology (NIST) (2014)
10. Neves, M., Ševa, J.: An extensive review of tools for manual annotation of documents. Briefings in bioinformatics **22**(1), 146–163 (2021)
11. Neves, M.L., Leser, U.: A survey on annotation tools for the biomedical literature. Briefings Bioinform. **15**(2), 327–340 (2014)
12. Salgado, D., Krallinger, M., Depaule, M., Drula, E., Tendulkar, A.V., Leitner, F., Valencia, A., Marcelle, C.: Myminer: a web application for computer-assisted biocuration and text annotation. Bioinform. **28**(17), 2285–2287 (2012)
13. Sequiera, R., Tan, L., Lin, J.: Overview of the TREC 2018 real-time summarization track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018. NIST Special Publication, vol. 500-331. National Institute of Standards and Technology (NIST) (2018)

14. Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for nlp-assisted text annotation. In: Daelemans, W., Lapata, M., Màrquez, L. (eds.) EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012. pp. 102–107. The Association for Computer Linguistics (2012)
15. Voorhees, E.M., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W.R., Lo, K., Roberts, K., Soboroff, I., Wang, L.L.: TREC-COVID: constructing a pandemic information retrieval test collection. SIGIR Forum **54**(1), 1:1–1:12 (2020)
16. Voorhees, E.M., Harman, D.K.: Overview of the seventh text retrieval conference (TREC-7) (1999)
17. Yimam, S.M., Gurevych, I., de Castilho, R.E., Biemann, C.: Webanno: A flexible, web-based and visually supported system for distributed annotations. In: 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria. pp. 1–6. The Association for Computer Linguistics (2013)