Find Problems before They Find You with AnnotatorPro's Monitoring Functionalities

Mohammed R. H. Qwaider, Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini
Fondazione Bruno Kessler, Trento, Italy
{qwaider,minard,manspera,magnini}@fbk.eu

Abstract

English. We present a tool for annotation of linguistic data. ANNOTATORPRO offers both complete monitoring functionalities (e.g. inter-annotator agreement and agreement with respect to a gold standard) and highly flexible task design (e.g. token and document level annotation, adjudication and reconciliation procedures). We teste ANNOTATORPRO in several industrial annotation scenarios, coupled with Active Learning techniques.

Italiano. *Presentiamo uno strumento per l'annotazione di dati linguistici. AnnotatorPro offre sia complete funzionalità di monitoraggio (es. accordo tra annotatori, accordo rispetto ad un gold standard), sia la alta flessibilità nel definire task di annotazione (per esempio, annotazione per parole o per documento, procedure di aggiudicamento e re-conciliazione). AnnotatorPro è stato sperimentato in diversi scenari di annotazione industriali, accoppiato con tecniche di Active Learning.*

1 Introduction

Driven by the popularity of machine learning approaches, there has been in the last years an increasing need to produce human annotated data for a large number of linguistic tasks (e.g. named entity recognition, semantic role labeling, sentiment analysis, word sense disambiguation, and discourse relations, just to mention a few). Datasets (development, training and test data) are being developed for different languages and different domains, both for research and industrial purposes.

A relevant consequence of this is the increasing demand for annotated datasets, both in terms of quantity and quality. This in turn calls for tools with a rich apparatus of functionalities (e.g. annotation, visualization, monitoring and reporting), able to support and monitor a large variety of annotators (i.e. from linguists to mechanical turkers), flexible enough to serve a large spectrum of annotation scenarios (e.g. crowdsourcing and paid professional annotators), and open to the integration of NLP tools (e.g. for automatic preannotation and for instance selection based on Active Learning).

Although there is a large supply of annotation tools, such as *brat* (Stenetorp et al., 2012), *GATE* (Cunningham et al., 2011), *CAT* (Bartalesi Lenzi et al., 2012), and *WebAnno* (Yimam et al., 2013), and several functions are included in common crowdsourcing platforms (e.g. *CrowdFlower*[1]), we believe that none of the available tool possesses the full range of functionalities for a real and intensive industrial use. As an example, none of the afore mentioned tools allows one to implement adjudication rules (i.e. under what condition an item annotated by more than one annotator is assigned to a certain category) or to visualize items with disagreement among annotators.

This paper introduces ANNOTATORPRO, a new annotation tool which was mainly conceived to fulfill the above-mentioned needs. We highlight two main aspects of the tool: (i) a high level of flexibility to design the annotation task, including the possibility to define adjudication and reconciliation procedures; (ii) the rich set of functionalities allowing for constant monitoring of the quality of the data being annotated.

The paper is organized as follows. In Section 2 we compare ANNOTATORPRO with some state-of-the-art annotation tools. Section 3 provides a general description of the tool. Sections 4 and 5 focus on the task design and on the monitoring functionalities, while Section 6 provides a brief overview of the tool's application and future extensions.

[1] https://www.crowdflower.com

2 Related Work

Many annotation tools are available to the community. However, some of them are limited by license, e.g. *CAT* (Bartalesi Lenzi et al., 2012) and *GATE* (Cunningham et al., 2011) are available for research use only, while some others have open licenses, e.g. *brat* (Stenetorp et al., 2012), but offer limited features.

The *brat rapid annotation tool* (*brat*) is an open license annotation tool that supports different annotation levels, in particular annotation at the token level and annotation of relations between marked tokens. It supports multiple annotators, in the sense that many annotators can collaborate on annotating the same corpus, but needs an in-house installation. Despite all these advantages, *brat* does not support either annotation monitoring or annotator/task reports.

Other tools (e.g. *CAT*) provide advanced functionalities to perform annotation at different levels (e.g. token and relation level) through a user-friendly interface, although they do not support annotation monitoring.

CrowdFlower is an outsourcing annotation service that provides a platform for annotation (focusing on annotation at the document level) employing non expert contributors. It uses gold standard tests to evaluate the annotators and supports automatic adjudication features, but no inter-annotator agreement metrics are available. In addition an important issue which could limit the use of outsourcing is the non in-house storage of the data, in particular when sensitive data covered by privacy regulations are concerned.

GATE is a powerful tool that implements most of the features to facilitate the annotation production in all its phases (e.g. task creation, annotator assignment, annotation monitoring and multi-layer annotation of the same corpus). However, visualization of disagreement is not available and no automatic adjudication is available.

3 Overall Description

ANNOTATORPRO is a web-based annotation tool built on top of the open source tool MT-EQUAL (Machine Translation Error Quality Alignment), a toolkit for the manual assessment of Machine Translation output that implements three different tasks in an integrated environment: annotation of translation errors, translation quality rating (e.g. adequacy and fluency, relative ranking of alternative translations), and word alignment (Girardi et al., 2014).

ANNOTATORPRO inherits from MT-EQUAL the capability of scaling over big data in an optimized platform that is able to save annotation in real-time. It also makes use of the MT-EQUAL web-based interface which is a multi-user and user-friendly interface.

It performs simple tokenization based on spaces, punctuation, and other language-dependent rules, but the user can also upload directly tokenized files.

We designed new functionalities to fulfill the requirements of high quality corpus annotation performed by multiple annotators. ANNOTATORPRO's main novel features are:

- The interface includes different options to design the annotation task (Section 4.1), which are set by the project manager.
- The tool enables annotation at two levels (Section 4.2): annotation at the token level (e.g. part-of-speech tagging and named entity recognition) and annotation at the document level (e.g. sentiment analysis).
- ANNOTATORPRO's interface offers functionalities for annotation monitoring (Section 5), which include inter-annotator agreement (IAA) monitoring and quality monitoring.

ANNOTATORPRO has been implemented in PHP and JavaScript, and uses MySQL to manage a database. It takes as input several UTF-8 encoded formats: TXT (raw text), IOB2[2] and TSV (tab separated values). It also accepts ZIP archives containing the source files.

As regards data storage, document's annotations are saved in a MySQL database in real time (i.e. while data being annotated). The annotated data can be exported in the following formats: IOB2 and TSV.

4 Annotation Task Design

ANNOTATORPRO distinguishes two types of users, i.e. managers and annotators. Managers

[2]The IOB2 tagging format is a common format for text chunking. B- is used to tag the beginning of a chunk, I- to tag tokens inside the chunk and O to indicate tokens not belonging to a chunk.

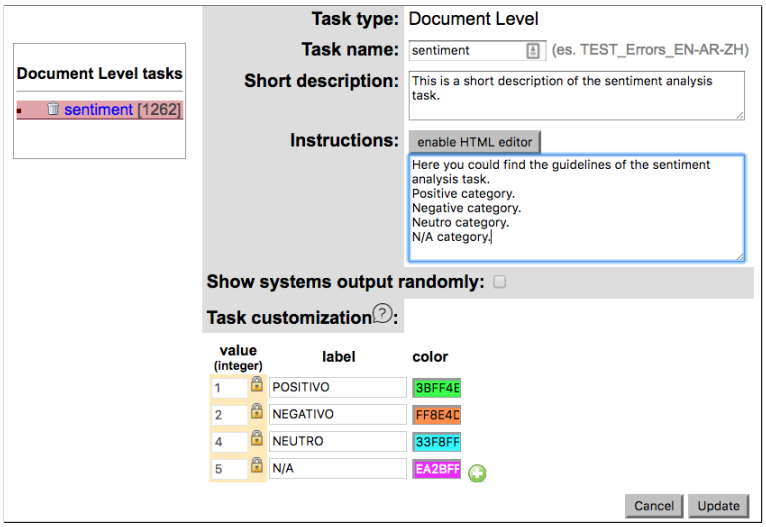

Figure 1: Annotator's task definition: annotation level, task's name, task description, and annotation categories.

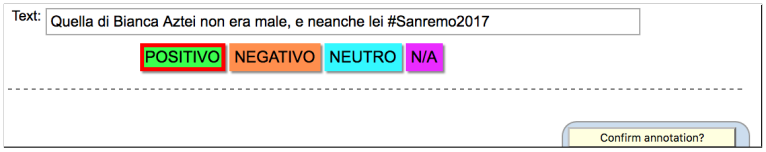

Figure 2: An example annotation interface: sentiment annotation of tweets.

take care of designing the annotation task at hand; in particular, they (i) define the annotation procedure, which depends on the number of annotators, their level of expertise (for example, non-expert annotators might not be allowed to see/modify each other's work) and the use that the dataset is intended for (e.g. evaluation, training, etc.), and (ii) the annotator's task, which includes selecting the most appropriate annotation level and creating the annotation categories/labels (Figure 1). As opposed to managers, annotators are basic users, who only have access to a limited number of (annotation) functionalities (Figure 2).

4.1 Annotation Procedure

One of the main tasks of the manager is to define the annotation procedure, which consists mainly of:

- Defining the number of annotators (one or more) who can collaborate on annotating the same corpus.
- In case of multiple annotators, defining the type of collaboration among them, i.e. whether data are to be annotated only by one or more of them (document level only).
- Defining the automatic adjudication rules in the case where multiple annotations of the same data are collected (document level only). The two basic options are:
  - considering an annotation as solved if the majority of annotators agreed on a certain annotation;
  - considering an annotation as solved if a minimum number of concordant annotations is reached.
- Deciding whether to make the metadata of the documents (e.g. document id, document title) visible to the annotators during the annotation phase.
- Deciding whether to allow for a revision phase after the annotation has been concluded, i.e. give the annotators the possibility to modify their annotations, for example after a reconciliation step has taken place. By default, document metadata will be visible during the revision phase to facilitate the work.
- Decide the modality for the selection of data to be presented to the annotators:

- propose to the annotator preselected ordered documents (default option);
- randomly select documents from a large dataset;
- select documents from a large dataset through an Active Learning process.[3]

4.2 Annotator's Task

ANNOTATORPRO supports two different annotation levels, i.e one where annotation is performed at the document level and one where we have smaller units, typically tokens, being annotated. It is the manager's task to select the most appropriate annotation level for the task at hand; for example, named entity recognition needs data annotated at the token level, whereas for sentiment analysis a corpus is generally annotated at the document level.

Finally, the task manager defines the set of categories or the set of labels to be used by the annotator respectively to classify the documents (in the case of document level annotation) or to mark portion of text.

5 Annotation Monitoring

In ANNOTATORPRO we have implemented several monitoring functionalities aimed at guaranteeing high quality annotation as described below.

5.1 Progress Monitoring

From the manager interface two tabs display information about the annotations already performed. The **Annotation** tab presents the progress of the annotation task, i.e. the annotations done by each annotator. This is real-time information, which means that the manager can follow the progress of the work underway. Moreover the manager can visualize the annotations of each user in read-only mode.

The **Overall stats** panel displays a table which summarizes the overall statistics about the annotation. The following information is given: total number of annotated documents; number of non-annotated documents; number of partially annotated documents (i.e. documents not yet annotated by the required number of annotators); number of completely annotated documents (i.e. documents annotated by the required number of annotators, independently of whether annotators did or did not reach an agreement).

[3]The Active Learning process is not provided in the distribution of ANNOTATORPRO, but the tool can select the data to be annotated if they are associated with a confidence value (in this case the tool can either select those with the highest score or those with the lowest score).

5.2 Inter-Annotator Agreement Monitoring

IAA monitoring, which measures the level of agreement between the annotators at regular intervals, is activated every time two or more annotators annotate the same data.

IAA agreement is computed in terms of Dice coefficient (Lin, 1998) and Cohen's Kappa (Viera and Garrett, 2005); the latter represents the agreement as a continuous value from -1 to 1, where -1 means total disagreement and 1 means total agreement.

The project manager has access to different types of information to constantly monitor the level of agreement between annotators, focusing both on a single annotator and overall:

- the level of agreement each annotator obtains with every other annotator and the average of the IAA values obtained by each annotator;
- the overall average IAA.

ANNOTATORPRO also provides a visualization of the annotations made by each annotator for each document, where a different color is used to present each tag from the tagset (see Figure 3). This enables the manager to have quick and easy access to the cases of disagreement and, if needed, to give feedback to the annotators.

5.3 Quality Monitoring

Quality monitoring makes use of a gold standard dataset previously annotated by an expert. Each annotator is asked to provide an annotation for those samples. The annotators do not know if they are annotating a golden sample or not, which ensures a non-biased evaluation. This enables the project manager to assess the quality of the annotations of each annotator by comparing them against a dataset considered correct. The same quantitative information and visualization as those for IAA monitoring (see Section 5.2) are available.

6 Applications and Further Extensions

We used ANNOTATORPRO for multiple projects, on different tasks, including named entity recognition (Minard et al., 2016a), event detection (Minard et al., 2016b) and sentiment analysis. The

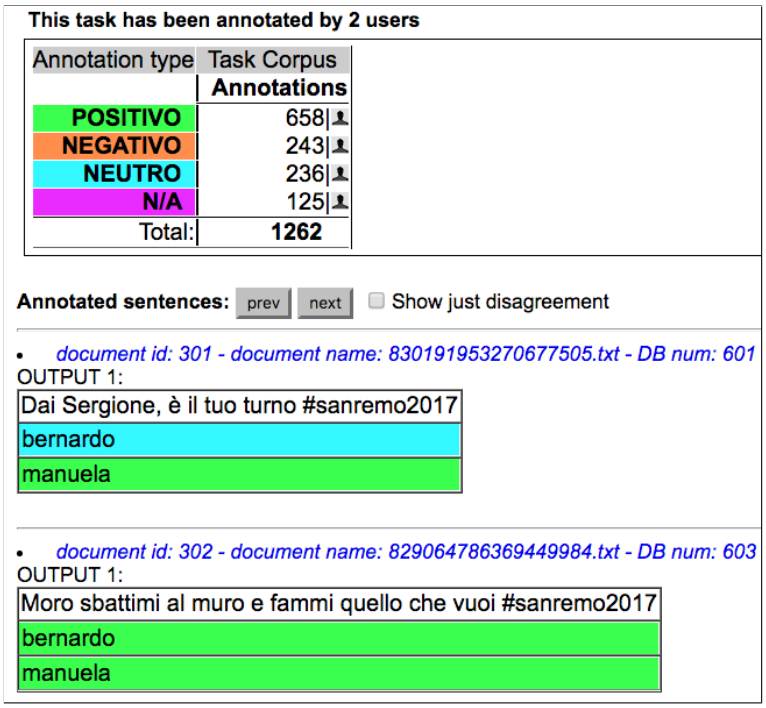

Figure 3: Visualization of the annotations made for two documents. The first example is a case of disagreement and the second a case of agreement. At the top of the page is given the number of annotations for each tag.

tool has been successfully exploited both in situations with few experienced annotators as well as with more than 20 non-expert annotators (i.e. high school students) working in parallel. ANNOTATORPRO has been fully integrated within an Active Learning platform (Magnini et al., 2016) and successfully employed in two industrial projects, resulting in high quality data.

As for our next steps, we are working to extend ANNOTATORPRO to include relations among annotated entities, such as the relation between a verb and its argument/s in semantic role labeling.

ANNOTATORPRO is distributed as open source software under the terms of Apache License 2.0.[4] from the web page: `http://hlt-nlp.fbk.eu/technologies/annotatorpro`.

Acknowledgments

This work has been partially funded by the EuclipRes project, under the program *Bando Innovazione 2016* of the autonomous Province of Bolzano.

[4] `https://www.apache.org/licenses/LICENSE-2.0`

References

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT annotation tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 333–338, Istanbul, Turkey, May 23-25, 2012.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*. University of Sheffield Department of Computer Science.

Christian Girardi, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico. 2014. MT-EQuAl: A toolkit for human assessment of machine translation output. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 120–123, Dublin, Ireland, August 23-29, 2014. ACL.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, Madison, Wisconsin, USA. Morgan Kaufmann Publishers Inc.

Bernardo Magnini, Anne-Lyse Minard, Mohammed R. H. Qwaider, and Manuela Speranza. 2016.

TextPro-AL: An active learning platform for flexible and efficient production of training data for NLP tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 131–135, Osaka, Japan, December.

Anne-Lyse Minard, Mohammed R. H. Qwaider, and Bernardo Magnini. 2016a. FBK-NLP at NEEL-IT: Active learning for domain adaptation. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, volume 1749, Napoli, Italy, December 5-7, 2016.

Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini, and Mohammed R. H. Qwaider. 2016b. Semantic interpretation of events in live soccer commentaries. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. 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*, EACL '12, pages 102–107, Avignon, France. Association for Computational Linguistics.

Anthony J. Viera and Joanne M. Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363, 5.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. 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*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.