RESEARCH

Customized automatic corpus annotations using AlvisNLP/ML

Robert Bossy*, Mouhamadou Ba and Claire Nédellec

*Correspondence:
Robert.Bossy@inra.fr
MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France
Full list of author information is available at the end of the article

Abstract

We propose to develop a versatile automatic corpus annotation tool based on the AlvisNLP/ML engine [1] following the PubAnnotation API. AlvisNLP/ML allows to process corpora with custom workflows. It has been used for named-entity recognition, relation extraction, and entity semantic categorization. The service we propose would not perform a single annotation task, but give access to a library of annotation tools. In this proposal we first describe AlvisNLP/ML, then we detail the proposal and the expected benefits. Finally we provide technical details in order to clarify the current development status.

Keywords: automatic corpus processing; annotation workflow; configurable service

Description of AlvisNLP/ML

AlvisNLP/ML is a highly configurable automatic corpus annotation engine. It has been developed for ten years by our team to support our experiments in natural language processing, information extraction, information retrieval and semi-automatic acquisition of semantic resources (ontologies and terminologies) for biology. Its main design goals are customization in order to support a wide range of computational linguistics and semantic annotation experiments, and reproducibility of these experiments.

AlvisNLP/ML has been used in applications with diverse objectives (IE, IR, DM) and domains of biology (microbiology, plant biology, molecular biology, biodiversity studies). For instance AlvisNLP/ML was the main tool that allowed the construction of annotated corpora and supporting resources for the BioNLP-ST 2016 [**?**] tasks Bacteria Biotopes [2] and Plant Seed Development [3]. It was also the annotation framework for building semantic search engines [**?**], and has been integrated as a semantic annotator in AgroPortal [4].

AlvisNLP/ML embeds an extensive library of processing modules that users can combine in order to tackle their own needs. This library of modules includes:

- *Computational linguistic tools*: tokenizers, POS-taggers, chunkers, parsers. For each task, several alternative tools are available (for instance both treetagger [5] and Genia Tagger[6] are available for POS-tagging).
- *Machine Learning Algorithms*: Weka [7], Wapiti [8]. AlvisNLP/ML supports both training and prediction.
- *Low-level tools*: lexicon projection, pattern matching, regular expressions, etc. These modules can be combined to build specialized annotation tools.

- *Input and Output*: the corpus and the data can be read and written into a wide range of formats, including text, PDF, XML, HTML, CSV, XML, RDF, RIS.

Users specify the sequence of modules, the input data, and the parameters through a plan file written in XML. The syntax of the plan file targets users with natural language processing background and very little training in software development.

Proposal and expected benefits

AlvisNLP/ML has allowed to build different annotation tools. We propose to develop a Web Service for automatic corpus annotation based on AlvisNLP/ML that conforms to Linked Data and PubAnnotation specifications. This would allow the community to access the processing workflows developed with AlvisNLP/ML, some of which provide the state-of-the-art information extraction performance.

For the BLAH3 event, we will provide a couple of plans for testing and production purposes, that will be usable immediately. In the long term, this service will give access to a growing library of workflows. Moreover AlvisNLP/ML will be made interoperable with other frameworks from the OpenMinTeD project [9] which includes GATE [10], DKPro [11], and Argo [12]. This project will greatly increase the pool of available components and workflows.

Technical Information

AlvisNLP/ML is written in Java and uses Maven for dependency management. It is distributed with the Apache License version 2 [13]. It integrates external tools with wrappers that adapt the input, the output, and the configuration parameters to a shared data model. AlvisNLP allows thus a uniform access to all its modules through the plan file.

The plans can be enacted through a command-line interface, as well as a REST interface. Both interfaces allow to execute the plans, as well as to read the documentation embedded to the AlvisNLP/ML distribution.

The REST interface is in its earlier version composed of RESTful services and is developed with Jersey so it should be deployable on any servlet container. The RESTful interface uses its own data exchange protocol, this proposal aims at the adaptation to the PubAnnotation specification.

Author's contributions

RB is the main developer of AlvisNLP/ML. MB has RESTified AlvisNLP/ML. CN is the scientific advisor for AlvisNLP/ML uses and directions.

References

1. Ba, M., Bossy, R.: Interoperability of corpus processing work-flow engines: the case of alvisnlp/ml in openminted. In: Eckart de Castilho, R., Ananiadou, S., Margoni, T., Peters, W., Piperidis, S. (eds.) Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016, pp. 15–18. European Language Resources Association (ELRA), Portorož, Slovenia (2016)
2. Deléger, L., Bossy, R., Chaix, E., Ba, M., Ferré, A., Bessières, P., Nédellec, C.: Overview of the bacteria biotope task at bionlp shared task 2016. In: Nédellec, C., Bossy, R., Kim, J.-D. (eds.) Proceedings of the 4th BioNLP Shared Task Workshop, pp. 12–22. The Association for Computational Linguistics, Berlin, Germany (2016)
3. Chaix, E., Dubreucq, B., Fatihi, A., Valsamou, D., Bossy, R., Ba, M., Deléger, L., Zweigenbaum, P., Bessieres, P., Lepiniec, L., Nédellec, C.: Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016. In: Nédellec, C., Bossy, R., Kim, J.-D. (eds.) Proceedings of the 4th BioNLP Shared Task Workshop, pp. 1–11. The Association for Computational Linguistics, Berlin, Germany (2016)
4. Jonquet, C., Dzalé-Yeumo, E., Arnaud, E., Larmande, P.: Agroportal: a proposition for ontology-based services in the agronomic domain. In: 3ème Atelier INtégration de Sources/masses de Données Hétérogènes et Ontologies, dans Le Domaine des Sciences du VIVant et de l'Environnement, IN-OVIVE'15, p. 5 (2015)

5. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Jones, D., Somers, H. (eds.) New Methods in Language Processing. Studies in Computational Linguistics, pp. 154–164. UCL Press, London, GB (1997)
6. Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. In: Panhellenic Conference on Informatics, pp. 382–392 (2005). Springer Berlin Heidelberg
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009). doi:10.1145/1656274.1656278
8. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 504–513. Association for Computational Linguistics, ??? (2010). http://www.aclweb.org/anthology/P10-1052
9. OpenMinTeD – Open Mining Infrastructure for Text and Data. http://openminted.eu/
10. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6), (2011). http://tinyurl.com/gatebook
11. Eckart de Castilho, R., Gurevych, I.: A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In: Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT, pp. 1–11. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (2014). http://www.aclweb.org/anthology/W14-5201
12. Rak, R., Rowley, A., Black, W.J., Ananiadou, S.: Argo: an integrative, interactive, text mining-based workbench supporting curation. Database: The Journal of Biological Databases and Curation **2012** (2012)
13. Apache License Version 2.0. https://www.apache.org/licenses/LICENSE-2.0