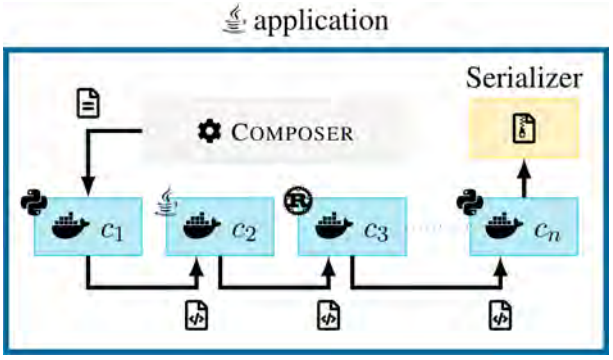
DUUI: A Toolbox for the Construction of a new Kind of Natural Language Processing

Abrami, Giuseppe
abrami@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland
ORCID: 0000-0002-7084-4909

Baumartz, Daniel
baumartz@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland
ORCID: 0009-0001-7105-5020

Mehler, Alexander
mehler@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland
ORCID: 0000-0003-2567-7539

Today, the heterogeneity of NLP tools in relation to existing methods and the constantly growing availability of models (see Hugging Face[1]) confronts various disciplines with major challenges in the daily handling of natural language processing. The spectrum of disciplines, although this is not exhaustive, ranges from biodiversity (e.g. (Lücking et al., 2021; Folk et al., 2024)), medicine (e.g. Poon et al. (2017); Redondo et al. (2019)), linguistics (e.g. Abdurakhmonova et al. (2022); Lücking et al. (2024)), all the way to the digital humanities (e.g. Brooke et al. (2015); Tasovac et al. (2023)). In parallel, the amount of available and usable corpora is also growing regularly in various areas, including, among others, corpora such as the "Collosal Clean Crawled Corpus" (C4 - (Raffel et al., 2020)), parliamentary protocols (e.g. Rauh and Schwalbach (2020); Abrami et al. (2022, 2024)), newspaper corpora (e.g. Süddeutscher Verlag (2014); New York Times (2019)), social media corpora (e.g. Dimitrov et al. (2020); Kratzke (2023)), COW (Schäfer, 2015) as well as Wikipedia (Pasternack and Roth, 2008). These are golden times for all scientific fields, as different models can be applied to the respective corpora; although in the short term this leads to non-trivial challenges in terms of a) analysis time, b) heterogeneity of (corpora) formats, c) processing input and output formats as well as d) analyzeability. These many *construction* phases show the need for a reliable working tool that can be used without intensive training, which is available in the form of **Docker Unified UIMA Interface** (DUUI)[2].



A Composer is defined within a Java application, which starts a set of **Components** that are each available as Docker images that encapsulate tools implemented in other programming languages (e.g. Python, Java or Rust). The sequential annotation enrichment between the individual **Components**, which are executed as Docker images, is done using Lua (Ierusalimschy et al., 2007), which performs the UIMA (de)serialization. At the end of each annotation, the individual documents are serialized, whereby various serializers (e.g. CoNLL, TFC, XMI *and various database backends*) are available (Abrami et al., 2024).

DUUI (Leonhardt et al., 2023) is designed as a platform-independent annotation framework for the horizontal and vertical distribution of heterogeneous NLP processes towards microservice-oriented homogenization using web services for a unified processing and reuse of unstructured data. Using microservices such as Docker also allows programming language-different as well as version-different NLP tools such as *spaCy* (Honnibal et al., 2020), *Heideltime* (Strötgen and Gertz, 2015) or *GNFinder* (Mozzherin et al., 2024) to be used in a common aggregated pipeline without causing dependency problems between the individual tools, since each instance (in Docker Swarm also in a cluster mode) is running on its own and can be utilized via a REST web service. Unification and implicit reusability of NLP processing is ensured by using the *Unstructured Information Management Applications* (UIMA – Ferrucci et al. (2009)) approach as the basis for annotation and serialization on document level. How DUUI can be used as a tool, especially for the digital humanities (c.f. Abrami and Mehler (2024)), will be presented in various hands-on demonstrations, thereby initiating discussions within the community and establishing a forum for exchange on the homogenized usability of heterogeneous annotation tools within one framework.

Acknowledgements

We gratefully acknowledge the financial support provided by the German Research Foundation (DFG) for the project "Critical Online Reasoning in Higher Education" (FOR 5404, project number 462702138) and for the project "Ausbau und Konsolidierung des Fachinformationsdienstes Biodiversitätsforschung[3] (BIOfid)" (DFG: 326061700).

Fußnoten

1. https://huggingface.co/
2. Available via GitHub under the AGPL license.
3. Expansion and consolidation of the specialized information service for biodiversity research

Bibliographie

Abdurakhmonova, Nilufar. Z., Alisher S. Ismailov, and Davlatyor Mengliev (2022). Developing NLP Tool for Linguistic Analysis of Turkic Languages. In 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), pp. 1790–1793. 10.1109/SIBIRCON56155.2022.10017049.

Abrami, Giuseppe, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler (2022, June). German Parliamentary Corpus (GerParCor). In Proceedings of the Language Resources and Evaluation Conference, Marseille, France, pp. 1900–1906. European Language Resources Association. https://aclanthology.org/2022.lrec-1.202.

Abrami, Giuseppe, Mevlüt Bagci, and Alexander Mehler (2024). German Parliamentary Corpus (GerParCor) Reloaded. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, pp. 7707–7716. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.681.

Abrami, Giuseppe and Alexander Mehler (2024, 08). Efficient, uniform and scalable parallel NLP preprocessing with DUUI: Perspectives and best practice for the digital humanities. In J. Karajgikar, A. Janco, and J. Otis (Eds.), Digital Humanities Conference 2024 - Book of Abstracts (DH 2024), DH, pp. 15–18. Zenodo. https://doi.org/10.5281/zenodo.13761079.

Brooke, Julian, Adam Hammond, and Graeme Hirst (2015, June). GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. In A. Feldman, A. Kazantseva, S. Szpakowicz, and C. Koolen (Eds.), Proceedings of the Fourth Workshop on Computational Linguistics for Literature, Denver, Colorado, USA, pp. 42–47. Association for Computational Linguistics. h ttps://aclanthology.org/W15-0705.

Dimitrov, Dimitar, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze (2020). TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, New York, NY, USA, pp. 2991–2998. Association for Computing Machinery. 10.1145/3340531.3412765.

Ferrucci, David, Adam Lally, Karin Verspoor, and Eric Nyberg (2009). Unstructured Information Management Architecture (UIMA) Version 1.0. OASIS Standard. https://docs.oasis-open.org/uima/v1.0/uima-v1.0.html.

Folk, Ryan A., Robert P. Guralnick, and Raphael T. LaFrance (2024). FloraTraiter: Automated parsing of traits from descriptive biodiversity literature. Applications in Plant Sciences 12(1). 10.1002/aps3.11563.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020). spaCy: Industrial-strength Natural Language Processing in Python. 10.5281/zenodo.1212303.

Ierusalimschy, Roberto, Luiz H. de Figueiredo, and Waldemar Celes (2007). The Evolution of Lua.

Kratzke, Nane (2023). Monthly Samples of German Tweets (2023). 10.5281/zenodo.7708787.

Leonhardt, Alexander, Giuseppe Abrami, Daniel Baumartz, and Alexander Mehler (2023). Unlocking the Heterogeneous Landscape of Big Data NLP with DUUI. In H. Bouamor, J. Pino, and K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, pp. 385–399. Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.29.

Lücking, Andy, Giuseppe Abrami, Leon Hammerla, Marc Rahn, Daniel Baumartz, Steffen Eger, and Alexander Mehler (2024, may). Dependencies over Times and Tools (DoTT). In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, pp. 4641–4653. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.415.

Lücking, Andy, Christine Driller, Manuel Stoeckel, Giuseppe Abrami, Adrian Pachzelt, and Alexander Mehler (2021). Multiple Annotation for Biodiversity: Developing an annotation framework among biology, linguistics and text technology. Language Resources and Evaluation. 10.1007/s10579-021-09553-5.

Mozzherin, Dmitry, Alexander Myltsev, and Harsh Zalavadiya (2024, June). gnames/gnfinder: v1.1.6. 10.5281/zenodo.11584025.

New York Times (2019). New York Times. https://developer.nytimes.com/apis. Accessed: 2019; Data provided by The New York Times.

Pasternack, Jeff and Dan Roth (2008). The Wikipedia Corpus. Technical report.

Poon, Hoifung, Chris Quirk, Kristina Toutanova, and Wen-tau Yih (2017, July). NLP for Precision Medicine. In M. Popović, and J. Boyd-Graber (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Vancouver, Canada, pp. 1–2. Association for Computational Linguistics. https://aclanthology.org/P17-5001.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. 21(1).

Rauh, Christian and Jan Schwalbach (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. 10.7910/DVN/L4OAKN.

Redondo, Teófilo, Julia Díaz, Antonio M. Sandoval, and Leonardo C. Llanos (2019, 03/2019). Biomedical Term Extraction: NLP Techniques in Computational Medicine. International Journal of Interactive Multimedia and Artificial Intelligence 5(4), 51–59. 10.9781/ijimai.2018.04.001.

Schäfer, Roland (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, and A. Witt (Eds.), Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), Lancaster, 20 July 2015, Mannheim, pp. 28–34. Institut für Deutsche Sprache. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-38367.

Strötgen, Jannik and Michael Gertz (2015, September). A Baseline Temporal Tagger for all Languages. In L. Màrquez, C. Callison-Burch, and J. Su (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 541–547. Association for Computational Linguistics. https://aclanthology.org/D15-1063.

Süddeutscher Verlag (2014). Süddeutsche Zeitung. Süddeutscher Verlag.

Tasovac, Toma, Natalia Ermolaev, Andrew Janco, David Lassner, and Nick Budak (2023). Humanistic NLP: Bridging the Gap Between Digital Humanities and Natural Language Processing. 10.5281/ZENODO.8107554.