Efficient, uniform and scalable parallel NLP pre-processing with DUUI: Perspectives and Best Practice for the Digital Humanities

Giuseppe Abrami **Alexander Mehler**
Goethe University Frankfurt
Robert-Mayer-Strasse 10
60325 Frankfurt am Main
{abrami,mehler}@em.uni-frankfurt.de

There is a growing need for annotating corpora in the digital humanities. This includes a large number of annotation tasks that require manual (e.g. Zhang et al. (2010); Abrami et al. (2021)) or automatic annotation. The number of these tasks is increasing because of the growing number of available tools, models (e.g. huggingface[1]) and corpora (e.g. Lücking et al. (2021); Abrami et al. (2024)).

Currently, annotation systems are implemented project-specifically, using a) heterogeneous tools (e.g. spaCy (Honnibal et al., 2020), StanfordNLP (Manning et al., 2014)) that b) run separately and require individual scaling, and c) whose input/output formats have to be adapted to each other. As this takes a lot of time and resources, a solution has been developed called DOCKER UNIFIED UIMA INTERFACE (DUUI). DUUI (Leonhardt et al., 2023) is a UIMA-based framework that provides a unified representation of NLP processes and enables scaled, parallel and platform-independent processing of large corpora, even within a cluster. *UIMA* (Ferrucci et al., 2009) enables valid, flexible, reusable, extensible and document-based annotation using a schema (*UIMA TypeSystem*) for any annotation task. DUUI was designed to unify the heterogeneity of individual NLP tools and provide a user-friendly interface for non-computer scientists.

Using DUUI[2] requires knowledge of three technologies: UIMA, Lua and Docker[3]. UIMA provides a basis for annotating texts, *Lua* (Ierusalimschy et al., 2007) allows the platform-independent use of UIMA outside of Java and ensures a minimalist selection of required annotations, while Docker encapsulates NLP analyses. The processing is done using UIMA compliant documents and is performed sequentially by defined processes – encapsulated in Docker containers – with data read and written using Lua (Figure 1). Since documents are processed in UIMA, documents in other formats (e.g. plain text, CoNLL, TCF) must be converted accordingly. For the Docker and Lua components, users can take responsibility for their own development. Note that each DUUI COMPONENT consists of at least three elements, while its execution is divided into two phases, as shown in Figure 2.

In a few steps, DUUI enables the integration of NLP tools as lightweight COMPONENTS within a pipeline, which can then be reused as a Docker image. In this way, DUUI can be used as a flexible, easily expandable tool for the scalable and uniform use of NLP routines in the digital humanities.

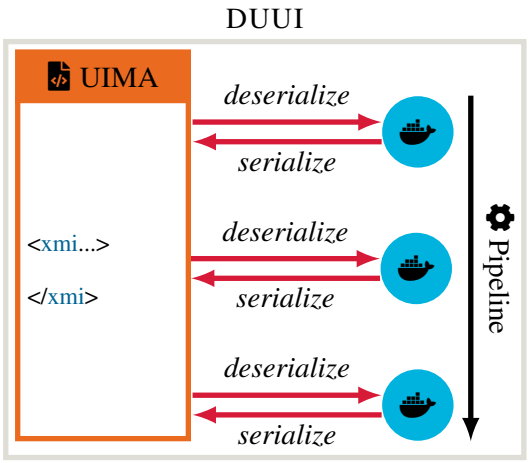

Figure 1: The way of DUUI: A UIMA document ( ) is processed sequentially through NLP processes encapsulated in Docker containers ( ), using Lua to (de)serialize the UIMA annotations.

References

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German parliamentary corpus (GerParCor) Reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguis-*

[1] https://huggingface.co/
[2] https://github.com/texttechnologylab/DockerUnifiedUIMAInterface
[3] https://www.docker.com/

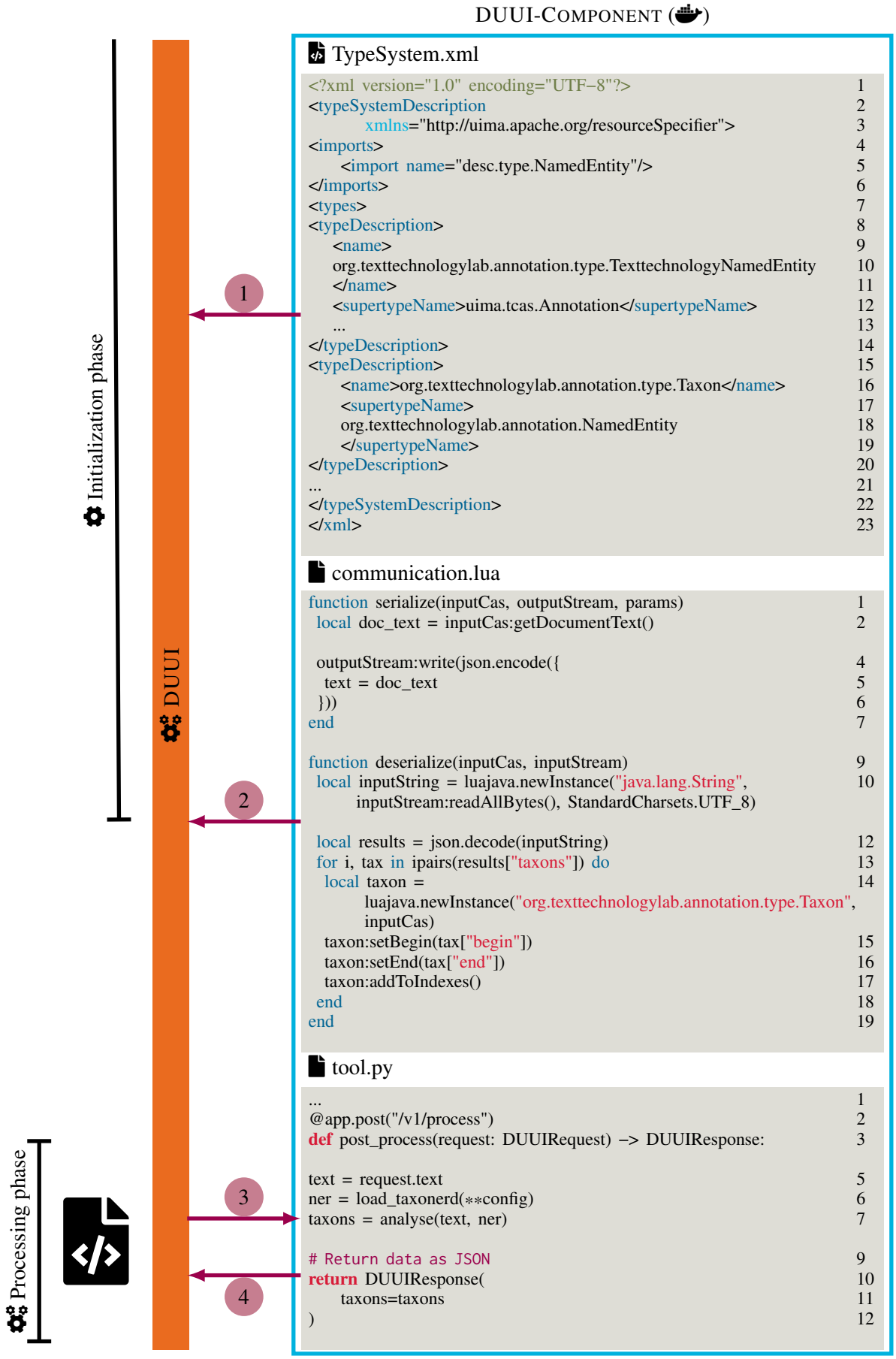

Figure 2: Encapsulation within DUUI. Communication between DUUI and the COMPONENTs can be carried out via REST, amongst other methods.

tics, Language Resources and Evaluation (LREC-COLING 2024), pages 7707–7716, Torino, Italy. ELRA and ICCL.

Giuseppe Abrami, Alexander Henlein, Andy Lücking, Attila Kett, Pascal Adeberg, and Alexander Mehler. 2021. Unleashing annotations with TextAnnotator: Multimedia, multi-perspective document views for ubiquitous annotation. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 65–75, Groningen, The Netherlands (online). Association for Computational Linguistics.

David Ferrucci, Adam Lally, Karin Verspoor, and Eric Nyberg. 2009. Unstructured Information Management Architecture (UIMA) Version 1.0. OASIS Standard.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Roberto Ierusalimschy, Luiz Henrique de Figueiredo, and Waldemar Celes. 2007. The evolution of lua.

Alexander Leonhardt, Giuseppe Abrami, Daniel Baumartz, and Alexander Mehler. 2023. Unlocking the heterogeneous landscape of big data NLP with DUUI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 385–399, Singapore. Association for Computational Linguistics.

Andy Lücking, 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*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Ziqi Zhang, Sam Chapman, and Fabio Ciravegna. 2010. A methodology towards effective and efficient manual document annotation: Addressing annotator discrepancy and annotation quality. In *Knowledge Engineering and Management by the Masses*, pages 301–315, Berlin, Heidelberg. Springer Berlin Heidelberg.