

Introduction to the Second Workshop on Humanities-Centred Artificial Intelligence

Sylvia Melzer^{1,2}, Hagen Peukert¹ and Stefan Thiemann¹

¹Universität Hamburg

²Universität zu Lübeck

In 2022, this year's workshop on Humanities-Centred Artificial Intelligence (CHAI) presents a selection of five papers that are supposed to reveal a variety of projects in the field of the Humanities, in which artificial intelligence (AI) methods are engaged to generate outcomes with higher rates of efficiency than with traditional methods. It seems that the focus on efficiency is the next logical step in this series of workshops intending to provide a circular view on all aspects of a commitment to Artificial Intelligence in the Humanities. While in 2021 the first workshop [1] prioritized all those projects that showed a deep impact on finding phenomena which the human mind is unable to think of in the first place, it bears a lot of plausibility to continue the workshop series with topics on how to best process, prepare and extract the needed information. In addition, we like to maintain the idea of presenting a very diverse array of projects and applications promoting the essence of the Humanities – a most diverse field of academic disciplines.

Admittedly, the focus on texts is prevalent throughout, even in disciplines like art history, musicology, or archaeology. Yet shifting towards new technologies in all fields is also undeniable. As an illustration, nowadays historians are increasingly using technologies to evaluate texts which are stored in a structured and machine-readable format such as Text Encoding Initiative (TEI) [2] or EpiDoc [3]. And if data are not available in appropriate formats, they will use approaches of optical character recognition maybe together with databasing on demand to automatically transform all data of interest to e.g. text encoded material and further into a structured machine-readable code that finally can be saved to a database [4, 5].

Moreover, Humanities' scholars engage computational pattern analysis (see paper 2), social network analysis (see paper 3), or Natural Language Processing (NLP) (see paper 5) to analyze the content or context of written artefacts such as manuscripts or, more specifically, inscriptions on bronze statues (see paper 5). Thus, networks of scribes are identified, the artefact itself is

Humanities-Centred AI (CHAI), Workshop at the 45th German Conference on Artificial Intelligence, September 19, 2022, Trier, Germany


✉ sylvia.melzer@uni-hamburg.de (S. Melzer); hagen.peukert@uni-hamburg.de (H. Peukert); stefan.thiemann@uni-hamburg.de (S. Thiemann)

🌐 <https://www.csmc.uni-hamburg.de/about/people/melzer.html> (S. Melzer)

🆔 0000-0002-0144-5429 (S. Melzer); 0000-0002-3228-316X (H. Peukert); 0000-0001-8300-2519 (S. Thiemann)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

correctly dated and assigned to a place of manufacture. NLP and other AI methods are used to detect patterns. However, generally, these methods often use training data from contemporary rather than historical data. This is problematic when the use of the method generated a bias in the historical record, risking incorrect conclusions about historical events, dates, or places. E.g. in contribution [6] it is shown, if a poem written in Tamil between the 1st century BCE and the 2nd century CE is translated into English using e.g. the Google Translator, the correct translation is not guaranteed. One of the reasons for incorrect translation is that the structure of a language from the past is different from that of today.

The very same phenomenon is addressed in the contribution on affix identification in Middle English (see paper 1), in which the semantic function of a bound affix may change over time. Yet this is only one side of the coin; the other side is the form of the affix that usually changes more drastically and leads to high degrees of variability hardly to be recognized either by humans or machines. So collecting representational quantitative data on the frequency of lexical affixes throughout 700 hundred years of English language use has proven to be challenging. While type frequencies of all suffixes and prefixes were determined with relative ease, the identification of token frequencies from larger text corpora turns out to be calling for AI approaches. Extracting all representations of one affix type and its exact quantities required taking into account all kinds of variability in form and usage. Exact quantities are required to make the more interesting statements on affix productivity and identification as well as interrelations with other factors of influence in the system of language, i.e. a correlation to word order or predictions of likely future changes. Again, because of the small quantities of available text training material, automated AI approaches have long been ignored as possible candidates for a viable solution. Indeed, this is comprehensible for Neural Network approaches, but as the contribution reveals in describing different stages of adjusting and exchanging methodological set ups, the correct combination of methods to solve the problem satisfactorily is finally achieved, i.e. a given (and long standing) problem in Diachronic Linguistics exemplifies how the existing inventory of AI-methods is typically applied. There are hardly any straight imperatives of proceedings that could be followed here. In fact, it cannot be plausibly predicted with a higher or lower probability as to which a certain AI method fits better than the other. Of course, it is possible to make a reasonable selection from the method inventory – that is, exclude neural networks because the data does not fulfill its very basic requirements – but it still leaves the researcher with too many alternatives from which it is impossible to estimate a success rate. What seems to be an trial-and-error approach from the outside, is a kind of systematic polling from the inside perspective. In the concrete case described in the contribution, one could learn from the history of implemented tools that, on the one hand, the right combination from a semi-automatic method (1st generation) enriched with a smart algorithm (2nd generation) would only be efficient if extended with a quality resource (4th generation). On the other hand, none of the components can be missed out, however, as the third generation showed, not all methods are equally optimal.

As further explicated in paper 4, that the algorithms of an information retrieval process produce results that frequently cannot be understood by the end users. Therefore, an information retrieval approach was presented, which explained information retrieval results in an explainable way.

To conclude, AI methods used in the Humanities should be further investigated considering

the many influential variables as in any other subject such as biases, objectivity, representativeness, validity and the like. During the CHAI 2022 workshop, the challenges in applying AI methods in the field of the Humanities and first solutions will be highlighted. In the contributions at hand, new algorithms and requirements are presented as well as one approach to fulfill the user needs during an information retrieval process through the supporting use of a Pepper robot.

The existing algorithms were developed to solve one problem and not all problems. To solve domain-specific problems, a knowledge base is needed that can be applied in the application of algorithms. But there is no algorithm that will work for all domains. There are only small parts which have to be combined effectively so that only the relevant knowledge has to be considered when selecting algorithms. [7] To gain knowledge from a variety of humanities projects and to be able to take them into account during implementation can be achieved through the interaction between humanities and computer science. This interaction space is created by the workshop Humanities-Centred Artificial Intelligence (CHAI).

References

- [1] S. Melzer, J. Gippert, S. Thiemann, H. Peukert, Proceedings of the Workshop on Humanities-Centred Artificial Intelligence (CHAI 2021), CEUR Workshop Proceedings 3093 (2022) 1–44. <https://ceur-ws.org/Vol-3093/>.
- [2] Text Encoding Initiative, P5: Guidelines for Electronic Text Encoding and Interchange, Version 4.0.0., Last updated on 13th February 2020, revision ccd19b0ba, <https://tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/>, 2020. Accessed 27 November 2022.
- [3] T. Elliott, G. Bodard, E. Mylonas, S. Stoyanova, C. Tupman, S. Vanderbilt, et al., EpiDoc Guidelines: Ancient documents in TEI XML (Version 9), Available: <https://epidoc.stoa.org/gl/latest/>, (2007-2022). Accessed January 22, 2022.
- [4] S. Schiff, S. Melzer, E. Wilden, R. Möller, TEI-based Interactive Critical Editions, in: 15th IAPR International Workshop on Document Analysis Systems, Lecture Notes in Computer Science (LNCS), Springer, 2022, pp. 230–244.
- [5] S. Melzer, S. Schiff, F. Weise, K. Harter, R. Möller, Databasing on demand for research data repositories explained with a large epidoc dataset, CENTERIS (2022).
- [6] S. Schiff, F. Kuhr, S. Melzer, R. Möller, AI-based Companion Services for Humanities, in: AI methods for digital heritage, Workshop at 43rd German Conference on Artificial Intelligence, 2020, pp. 1–3.
- [7] E. Rich, Artificial intelligence and the humanities, *Computers and the Humanities* 19 (1985) 117–122. URL: <http://www.jstor.org/stable/30204398>.