Some Challenges Posed to Computer Science by the eHumanities

Gerhard Heyer, Marco Büchler

Natural Language Processing Group
Institute for Mathematics and Computer Science
University of Leipzig, Germany
[heyer|mbuechler]@informatik.uni-leipzig.de

1 eHumanities

To the extent that applications of computer science have always lead to a replacement of analogue by *digital* media and processes, digital media and processing models have an increasing impact also on traditional work flows based on analogue media in the humanities and social sciences. The interdisciplinary combination of methods from computer science and traditional humanities with large amounts of digital data and advanced tools for processing these is commonly known as *eHumanities* [Ca05]. In a broad sense, eHumanities are concerned with any digitized data that are subject to investigation in the humanities and the social sciences, e.g. text, images, and objects (such as in Archeology). However, focusing on text as the main data type in the text oriented humanities helps to highlight the benefit that can be gained from the combination of digital document collections and new analysis tools derived from the area of information retrieval and text mining. Thereby all kinds of historically oriented text sciences as well as all sciences that work with historical or present day texts and documents are enabled to ask completely new questions and deal with text in a new manner. In detail, these methods concern, amongst others,

- The *qualitative improvement* of the digital sources (standardization of spelling and spelling correction, unambiguous identification of authors and sources, marking of quotes and references, temporal classification of texts, etc.);
- The *quantity and structure* of sources that can be processed (processing of very large amounts of text, structuring by time, place, authors, contents and topics, comments from colleagues and other editions, etc.);
- The kind and quality of the *analysis* (broad data driven studies, strict bottom-up approach by using text mining tools, integration of community networking approaches).

To use such computational methods, an individual researcher can proceed by employing two strategies depending on his, or her, own degree of computer literacy. On the one hand, there is the individual software approach. Given a selection of digital text data, the research question is being transferred into a set of issues and methods that can be dealt

with by a number of individual programs. This approach allows for a highly dynamic and individual development of research issues. It requires, however, a high degree of software engineering know-how. On the other hand, there is the standard software approach. For well-defined and frequently encountered tasks, a digital humanities infrastructure will offer solutions that provide the users with data and analysis tools that are well understood, have already delivered convincing results, and can be learnt without too much effort.
Both approaches are interdependent. Probably good solutions in one domain of text oriented humanities can be transferred to other domains by just using different kinds of text. A good infrastructure must be capable of making such solutions accessible as best practices.

2 Interactions with Computer Science

While eHumanities are becoming increasingly popular in the humanities, the focus of this workshop is to investigate which consequences and potentials for **computer science** have emerged in turn from the digitization of the social sciences and humanities. Computer Science and Humanities so far have acted in their working methodologies more as antipodes rather than focusing on the potential synergies. For computer science turning towards the humanities as an area of application may pose new problems that may also lead to rethinking present approaches hitherto favored by computer science and developing new solutions that help to advance computer science also in other areas of media oriented applications.

By way of example, let us consider the so-called digital classics in detail. Historically oriented classical studies (Ancient History, Classical Philology – Latin, Greek, and Byzantine Studies – Epigraphics, Papyrology) nowadays all use digital text corpora that are available in various media (digital WEB libraries, CD-ROM, DVD) and formats (Beta-Code in all varieties, UTF-8, ASCII). Thanks to substantial digitization efforts, most of the ancient greek and latin text corpora are available nowadays over the internet and represent. The public resources of ancient texts include, amongst others, Thesaurus Linguae Graecae (TLG), Perseus, Packard Humanities Institute (PHI), and Bibliotheca Teubneriana Latina (BTL). While the use of digital texts has up to now been mainly restricted to search for, e.g., specific wordings, text passages, or proper names, advancements in text mining make it possible to exploit digital text resources also as raw material for acquiring structured knowledge out of the huge amount of unstructured textual data. A good reference to illustrate the point is *eAQUA*, an interdisciplinary project set up between the departments for classical studies at the University of Leipzig, Heidelberg, and Hamburg, and the division for natural language processing at the computer science department of the University of Leipzig, explicitly addressing the issue of text mining in digital text corpora for the classical studies [HS10]. eAQUA is funded by the German Ministry for Research, Education, and Technology, and aims to explore interdependencies between computer science and the humanities and to lay the foundations for an e-science infrastructure in the humanities. The goals of the project are threefold,

i. to establish a research infrastructure for researchers in the classical studies by setting up a portal that makes specific solutions to research issues accessible to the scientific community by way of best practice applications that can easily be adapted to similar research issues;
ii. to make different digital resources of ancient text accessible to researchers from the eAQUA platform; and
iii. to apply in an experimental way advanced text mining to the ancient texts in order to gain experience which kind of issues in the classical studies lend themselves to this approach, also advancing where possible the classical studies in their detailed research issues.

By using text mining, factual and content related interdependencies can be reconstructed that otherwise could not have been derived as rapidly and exhaustively. Their graphic depiction indicates streams of meaning that visualize historical traditions as well as historical facts in particular contexts like from a bird's eye view. In the traditional way of exploring historical texts such insights on a "higher level" are very time consuming and require an intimate knowledge of and a long time experience in handling ancient text sources.

3 Impacts on Computer Science

From a computer science point of view, we see four main challenges for Computer Science as a result of such interactions, (i) software engineering issues, (ii) impacts on semantic technologies, (iii) visual analytics, and (iv) infrastructure issues.

3.1 Software Engineering Issues

As a general approach to the software engineering issue of how to transform a research issue of the classical studies into a software problem, eAQUA, for example, follows a methodology that distinguishes between Data, Algorithms, and Applications, and that has successfully been applied in the area of natural language processing. Starting with an application as the actual scientific research question, we need data on the one hand, and algorithms on the other. In eAQUA, text data are being imported by eAQUA's standardized interface to available resources. Algorithms are taken from the field of natural language processing and text mining. The main challenge then is to select the suitable algorithms and specify their combination and interaction in order to contribute to a solution of the problems set out in the application scenario. Figure 1 represents a simple instance of this methodology for the reconstruction or the correction of a papyrus or an inscription.

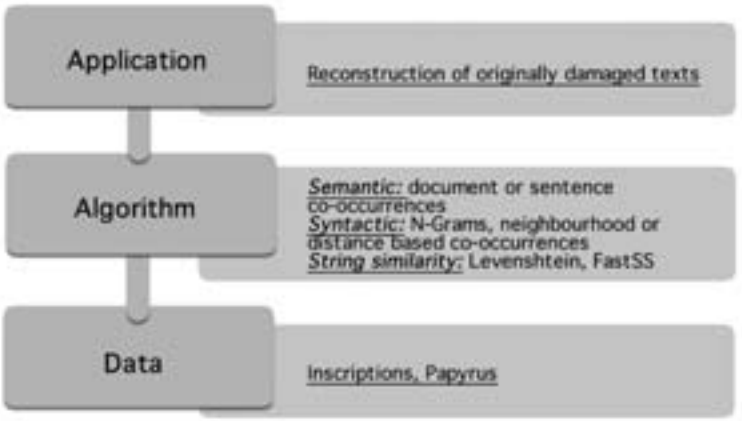

Figure 1: Distinguishing between data, algorithms, and applications for solving the problem of correcting damaged papyri

How can this approach be generalized? And how can it be implemented by using a flexible architecture that allows as much as possible reuse of individual software components, and rapid prototyping of new applications?

3.2 Impacts on Semantic Technologies

Research in Classical Studies is often data driven, i.e. researchers are looking for phrases or words, collect the results and interpret them in view of their field's background knowledge. In contrast, the preferred method of computer scientists is model driven. For dedicated texts, or fragments of text, they try to define significant features that can be used as input for algorithmic analyses of the text, such as classification, or clustering, and latent semantic analysis.

A good example of a possible trade-off between both methods again is eAQUA. On the one hand, researchers in the classical studies need access to digital libraries for searching and comparing texts, on the other hand, the power of model driven analytics can also reshape a research question from a more formal point of view. While a confirmation of what is already known from studying ancient texts always is helpful in scientific work (in particular when it is speeding up research), it is not by itself really innovative and would not justify considering it a novel method of search. The situation is different when by using, e.g., a contextual search we get hints on semantic relations that are neither obvious, nor well known, and that cannot easily be derived by using traditional or otherwise established tools and search strategies like dictionaries, concordances, indices, or dedicated search engines like Diogenes and TLG-online.

Scientific practice in the humanities clearly indicates that we may need to distinguish between knowledge concerning what is obvious and commonly accepted on the one hand, and knowledge concerning particular, or rare events, knowledge that summarizes

the exception from the rules and is distinctive of experts, but that many wish to share. For a number of reasons, including considerations of relevance and optimization, in information retrieval as well as in data and text mining the main methods today are based on statistical methods. By using such methods we can well detect and analyze statistically significant patterns, but these methods are not really suitable for dealing with rare events and their significance within a given context of research. How can this knowledge of rare events be represented and adequately be dealt with? How can we modify and apply semantic technologies in information retrieval to also deal with rare events?

3.3 Visual Analytics

A common assumption in computer science up till now is that first data need to be analyzed, and that visualization then is needed only to visualize the results. However, the necessity to deal with an increasing amount of data may require us to rethink this premise. Visual analytics aims to interactively filter the information relevant to an application and to communicate it to humans in an appropriate way. Again, a good example is eAQUA and its tools for the analysis of textual reuse [BGH10]. Similar to modern publications, classical authors also used the texts of others as sources for their own work. Hence, the analysis of textual reuse plays an important role in Classical Studies research. Leaving aside the technical point of view of semantic technologies for this application, the research of Classicists includes both an application of a *macro view* for Historians as well as one for the *micro view* of Classical Philologists. The visualization dimension of textual reuse is important since text mining approaches typically generate a huge amount of data that can't be explored manually. In effect, visualization helps to better understand the data, allows to gain new insights, and to interact effectively with the data analysis methods. The interaction permits the user to bring his expert knowledge into the data analysis process. From a methodological point of view, however, the semantic technologies and text mining tools also need to get adapted in order to meet the challenge that important characteristics of the data such as dimensionality, homogeneity, topicality, precision, and completeness need to be considered. The visual analysis techniques themselves need to visually convey the quality and relevance of the data and to secure in this way the quality of the gained knowledge.

3.4 Infrastructure Issues

Finally, building an IT infrastructure for digital classics and other digital humanities also is a major challenge to computer science, as there seem to be as many data formats and user interfaces as there are use case scenarios. Often, the tools and resources are designed and suitable just for a dedicated research question. Typically, researchers in the field of Classics are working with more than one tool or web site at once and have to frequently switch between them. Hence, we need to find a way to efficiently deal with distributed and heterogeneous data resources. Most text resources differ in format as they are stored in proprietary and non-standard formats. Only some of the available

resources support the epiDOC standard, the extension of TEI P4 for epigraphic annotations like the Leiden Convention. In addition, the integration of text and image data poses a challenge that equally opens new perspectives as it is difficult. From a research infrastructure point of view, most of the tools for searching are standalone and only allow for local applications. Very rarely text data are complemented with more advanced natural language processing (NLP) applications such as *Morpheus*, the morphological analyzer for ancient Greek and Latin, or the Latin Treebank, that are both part of the Perseus project. In summary, we notice a lack of a suitable component-based and service-oriented framework for language-based resources in the eHumanities. But this in turn also poses a challenge to computer science to develop an infrastructure for the systematic and structured acquisition, generation, processing, administration, presentation, reuse, and publication of *contents*. Content services make available the resources and programs needed for that. Public digital text and data resources may be linked together and made accessible by common standards, while new software architectures integrate digital resources and processing tools to develop new and better access to digital contents.

Literature

[BGH10] Büchler, M.; Geßner, A.; Heyer, G., Eckart, T.: *Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project.* In: *Proceedings of the Digital Humanities Conference 2010,* King's College London: London 2010.

[HS10] Heyer, G.; Schubert, C.: *Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project.* In: *Proceedings of the Digital Humanities Conference 2010,* King's College London: London 2010.

[Ca05] McCarthy, W.: *Humanities Computing*, Basingstoke (U.K.): Palgrave 2005.