Uncertainty in Visual Text Analysis in the Context of the Digital Humanities

Markus John
Institute for Visualization and Interactive Systems
Universitätstraße 38
70569 Stuttgart, Germany
Markus.John@vis.uni-stuttgart.de

Steffen Koch
Institute for Visualization and Interactive Systems
Universitätstraße 38
70569 Stuttgart, Germany
Steffen.Koch@vis.uni-stuttgart.de

Thomas Ertl
Institute for Visualization and Interactive Systems
Universitätstraße 38
70569 Stuttgart, Germany
Thomas.Ertl@vis.uni-stuttgart.de

Abstract
In recent years, visual text analysis in the Digital Humanities (DH) has grown rapidly through the digitization of literary document collections. Normally, DH-projects in this context bring together scholars from the research fields of Computational Linguistics (CL), Visual Analytics (VA), and the Humanities. These research fields combine information technological methods and hermeneutic procedures to support each other in favor of an algorithmic criticism. However, hermeneutic and algorithmic approaches have different sources of uncertainty that can lead to dissatisfaction in text analysis. Therefore, it is necessary to determine these sources and to provide approaches that support users in exploring and communicating uncertainty. In this paper, we present several sources of uncertainty in the mentioned research context and propose approaches to achieve a more reliable analysis.

Author Keywords
Text Visualization, Visual Text Analytics, Digital Humanities, Uncertainty Visualization, NLP

ACM Classification Keywords
H.5.2 [Information interfaces]: User-centered design.; I.2.7 [Natural Language Processing]: Text analysis.
Introduction

Traditionally, humanities scholars are interested in analyzing literary works. Typical research questions are how authors introduce and discuss certain recurring topics or which works of literary authors are named and cited. The digitization and proliferation of literary works provide new means to support these tasks with visual abstractions that are generated by automatic methods. Generally, humanities scholars analyze literary works by using so-called close reading that focuses on the thorough interpretation of text passages [5]. On the contrary, Moretti introduced the technique Distant Reading [7], which provides visual abstractions of the text, such as graphs to show the relations between occurring persons. To provide visual abstractions for textual data, a combination of natural language processing (NLP) methods and visualization techniques is required. However, uncertainties can arise or propagate during this process. Humanities scholars are often unaware of uncertainties in their analysis. Automatic methods typically offer uncertainty information with their results. Communicating these information to the user can support them in decision making. Interactive visualization can play an important role in communicating, analyzing, and understanding uncertainty. For example, by providing visual clues to represent uncertainty or interaction methods that let users explore and adapt visualization methods according to their needs, as well as by offering different views on the text. In addition, it is essential to combine distant and close reading that humanities scholars can analyze uncertain instances in more detail. That way, humanities scholars get a better idea of the accuracy of the different methods. The aim of this contribution is to identify sources of uncertainty in this research context and to describe ways to ensure a more reliable analysis.

Sources of Uncertainty

Figure 1 represents a typically process pipeline in the DH which is subdivided into two parts; the linguistic analysis and the visualization pipeline adapted from Card et al. [1]. The linguistic analysis is based on the text corpus that is normally compiled by humanities scholars and represents the object of research. The next step is to derive and tag important information, such as persons or concepts, from the text through NLP techniques or manual annotations. The tagged corpus can then be stored in an appropriate data format. Subsequently, the visualization pipeline converts the information into interactive visual representations. The first step is to transform the tagged corpus in a suitable data format (if still needed).

Figure 1: The process pipeline represents a typically workflow consisting of a linguistic and visualization part.

The next visualization step is to map the data into visual variables such as visual glyphs. The last step integrates the visual variables into views and provides several view transformations. Finally, experts can analyze and explore the corpus by interacting with the views in order to gain insights or to generate new hypotheses. In addition, users
can interact with the different analysis steps to adapt the visualization result.

However, there are different sources of error in this workflow, as depicted in (Figure 1). These errors can lead to uncertainties in the analysis and influence other analysis steps. For example, through digitizing methods such as Optical Character Recognition (OCR) or manual transcriptions. Also, NLP methods often involve a certain degree of uncertainty, since they are normally trained on specific text corpora, such as newspaper or journal article texts, and do not provide entirely correct results for older or even historic texts. Additionally, there are very complex research questions, which need to be simplified so that automatic methods can process them. In those cases, not all instances can be determined and thus much information may be lost. Furthermore, many automatic methods are based on manual annotations. For example, as an initial set of instances on which a classifier is trained. During the manual annotation step, other errors can occur caused by mismatches in conceptualization, erroneous annotations, or missing annotations. These uncertainties can therefore affect the automatic methods and the visualization result. In addition, to the uncertainties of the data and above NLP methods, the visualization itself can contain uncertainties. There are challenges such as an adequate visual mapping and suitable design decisions. Since misleading visual metaphors can lead experts to erroneously insights, hypotheses or decisions. Whereas inappropriate design decision can produce clutter, which may prevent experts in gaining insights of the underlying data. Further, inherent uncertainties can occur when filtering or aggregating the textual data. Zuk and Carpendale [9] discuss such visualization effects in detail. The last discussed source of uncertainty is the expert, since human world knowledge and intuition is important for a correct analysis [8]. Uncertainties may arise during the analysis, such as ambiguities of a term that lead to incorrect decisions. In addition, the above mentioned examples can influence the expert’s decisions.

Increasing Uncertainty Awareness

Humanities scholars are often unaware of uncertainties in their data sources, preprocessing steps, or during the visual analysis [8]. Therefore, it is necessary to communicate uncertainty in the visualization to avoid incorrect decisions. Griethe and Schumann [3] summarize different methods to indicate uncertainty in data. These techniques include visual variables (position, color), extra objects (images, labels), animation (speed, blur motion), or other human senses (incorporation of acoustics, vibration). The visual clues can assist humanities scholars in getting a better idea about the accuracy of the respective procedures. This, in turn, supports experts in a better-informed decision making. In addition, to communicate the uncertainty with the aid of visual clues, it is essential to offer interactive methods that facilitate the understanding how the different uncertainties influence the final output. For example, interaction methods that support a seamless switching between distant and close reading, because it is clear that the humanities need to work with the text directly to investigate uncertain instances. Furthermore, by offering several views on the textual data that uncertainties are easier understandable and recognizable. VA can supports these tasks, since it facilitates iterative and explorative loops for analytical reasoning.

However, there are only a few works that implement an interactive and visual approach to achieve a more reliable analysis in visual text analysis. For example, VarifocalReader [6] supports humanities scholars by combining close and distant reading. It enables a intra-document analysis through an advanced navigation technique. The VarifocalReader integrates an active
learning algorithm to classify quotes between texts into three different categories. The approach uses color saturation to show the accuracy of the algorithm. Light colors indicate instances with a high uncertainty, which are good candidates for manual correction, while strong colors indicate instances which are assumed to be accurate. Users can correct or confirm these classifications and trigger a retraining to improve the classifier and thus a more certain result. Collins et al. [2] propose a visualization technique that depicts the uncertainty in lattices. It uses a hybrid layout along with varying transparency, color, and size to reveal the lattice structure and support experts in order to make better-informed decisions about statistically derived results. Another approach that lets experts interactively train a support vector machine for text document has been presented by Heimerl et al. [4]. It represents the uncertainty respectively the confidence value for each document by the horizontal position in the scatterplot. Furthermore, the approach provides several interaction methods and views, such as to inspect the documents in detail, which support experts in decision making. These mentioned approaches are first steps in the right direction, however more emphasis needs to be placed that uncertainties are better communicated to humanities scholars.

Conclusion

In conclusion, we identified several sources of uncertainty in visual text analysis in the context of DH and we gave a brief summary of visual clues that can be used to communicate uncertainty. Additionally, we introduced interactive visualization approaches that support users in understanding and exploring uncertainties in the data. In our opinion, it is necessary to evaluate visual clues and interaction methods to gain insights how the humanities can be better supported with a more reliable analysis in the future.

References