Time will tell a different story

Mining perspectives in historical texts with the help of NLP

Applicants

Supervisor Name	Department/Group	Faculty
1. Serge ter Braake, Post Doc	History/Web & Media	LET/FEW
2. Antske Fokkens, Post Doc	Linguistics/Web & Media	LET/FEW

Project description

Provide a brief description of the project (295 out of 300 words)

'*Everything we hear is an opinion, not a fact. Everything we see is a perspective, not the truth*.' Marcus Aurelius (Roman Emperor, 121-180 AD)

Historic research has gone through significant changes over time, in an effort to produce the most objective presentations of the past. Facts may remain the same, but perspectives in the way historians describe people and events change. Historians used to write about divine interventions, kings and battles, but they now also concentrate on topics like slaves, the history of medicine and the development of farming. The angle of analysis evolves hand in hand with the change of time, society and public opinion. It is not uncommon for example, that a Dutch governor could be considered a hero in the nineteenth century and now is considered to be a war criminal.

In this project, two student assistants will investigate how these changes can be traced in historic text. They will combine expertise in historiography (research on how historians write their work) and computational linguistics to see what patterns can be observed comparing texts from different epochs. They will address the following research questions:

1. How is the historian's perspective reflected in historic text?
2. What changes in perspectives can be observed in texts from different epochs?

These questions will be addressed using data from the Biography Portal of the Netherlands, a heterogeneous dataset describing approximately 76,000 people in 125,000 biographies. In particular, there are two large resources from the late 19th and early 20th century where we find examples that illustrate increased awareness on the lot of local inhabitants of former Dutch colonies throughout time. Differences in topics, vocabulary and how people are presented will be used to illustrate change in perspective. This project uses data and tools from the BiographyNet project.[1]

[1] http://www.biographynet.nl

Project Organization

Each proposal requests two Academy Assistants from different disciplines. Describe their roles and describe the skills and expertise required from them. (300 out of 300 words)

We aim to hire two Master Students: one with expertise in history (notably historiography) and one with expertise in computational linguistics or computer science.

The historian will manually approach two comparable use cases from different angles: one in collaboration with the computer scientist/computational linguist and one for evaluation purposes.

The computer scientist/computational linguist will attempt to emulate the manual results from the historian using existing tools developed at the CLTL Lab[2] or provided through AmCAT.[3] If necessary, the student will adjust tools to the corpus and task.

The students will jointly evaluate the outcome of the analyses and analyze the possibilities as well as the pitfalls of using NLP analyses for historiographic research.

All work will be carried out under supervision of Serge ter Braake and Antske Fokkens. The supervisors will be primarily responsible for writing the papers related to the project. The students will (at least) be co-authors of the papers.

The ideal historian for this project would have: a BA in history; demonstrable interest in and/or knowledge of digital humanities research, theory of history, historiography, historical methodology; a willingness to acquire basic knowledge in computer science and computational linguistics so that the historical methodology can be described in a useful way for computer scientists; an interest in the use of digital tools for historical research.

The ideal computer scientist/computational linguist would have: either a strong background in linguistics and affinity with technology (programming skills are a plus) or a strong technological background and an interest in language technology; the capability to run experiments with existing NLP tools and understanding of their working; the ability to explain his/her work to a historian; an interest in historical questions.

Both assistants should have excellent communication skills and be good team players. Interest in an academic career after graduation is an advantage.

Collaboration

Describe how your research improves collaboration and cross-pollination between the disciplines involved (299 out of 300 words)

Research on automatically identifying perspectives in text requires knowledge of computational linguistics and domain expertise. Though basic technology to investigate perspectives exists (topic analysis, information extraction, opinion mining), research on representing different perspectives in academic text through automatic text analysis is still in its initial stages. Domain expertise is required in order to determine what information is needed in order to identify perspectives in text. History students with a BA degree should have sufficient knowledge to provide this input given that historiography forms an important part of their curriculum. A good result can only be obtained when the assistants closely work together gaining insight in the methodologies the other applies.

The computational linguist will be forced to think beyond high precision and recall on standard NLP tasks. Together with the historian, he or she must figure out which conclusions may be drawn from the imperfect output of NLP tools. This question adds an additional layer to evaluation in NLP.

---

[2] http://wordpress.let.vupr.nl/software/
[3] http://amcat.vu.nl/

The historian learns of the possibilities of automated support in historical research and is forced to look at his or her own methodology in a digital humanities way. He or she must learn to translate the historian's workflow for people outside the humanities domain.

Contributions to interdisciplinary research go beyond the lessons learned by the students participating in this project. First, insight into what makes up a perspective (selection of information, word choice, directly expressed opinions) and how this can be identified by NLP analysis can be applied to other fields of humanities. Second, translating the historian's workflow to specific automated steps contributes to developing methodologies for digital history. Third, evaluating both accuracy of tools and their value for a digital humanities question is not commonly done, but a necessity for high quality research when using NLP in digital humanities.

Deliverables

Enumerate intended project results: papers, research proposals or otherwise. (151 out of 200 words)

We will aim to produce at least **two papers**: one workshop or conference paper in NLP describing the task of identifying perspectives and one in history describing the insights gained through NLP analysis. Depending on the outcome, a journal paper on digital humanities will follow at the end of the project.

The NLP pipeline created by the computational linguist/computer scientist can be used for all BiographyNet data. It will be integrated in the BiographyNet user interface, so that it can also make use of visualizations present in the interface. This will allow historians to make use of the application built in this project directly.

We are currently working with partners from Austria and Germany to write **a proposal for international research on biographical data**. The results of this research form a direct contribution to this proposal, since they provide a stepping-stone towards investigating the presence of nationalistic perspectives in biographical dictionaries.

Planning

Provide a breakdown of the project into phases with tentative timing (148 out of 150 words)

Phase 1: highly interactive, at least 50% of time together

October: Initial exploration of the data. Examining text and output of NLP tools. The students learn each other's language.
November-December: Students work together on a first use case (proposed: governors of Suriname). What can the tools do? Should sentiment vocabulary be adjusted for the domain?

Phase 2: moderately interactive, weekly meetings

January-February The computational linguist will set up the NLP pipeline and see whether tools need to be adjusted. The historian rounds up the first use case and starts on the second (proposed: governors of Dutch Indies)
April-March Adjusting NLP tools, finishing domain specific sentiment analysis and the second use case.

Phase 3: highly interactive

May-July The students evaluate performance of the NLP pipeline on the second use case. Results will be taken up in research papers. The pipeline is prepared for integration in the BiographyNet user interface.