Semantic Corpus Exploration: Introducing DBpedia Spotlight and WideNet for Digital Humanities

This tutorial provides a hands-on experience with semantic annotations for selecting text sources from large corpora. While many humanist scholars are familiar with annotation during the analysis of their selected source texts, we invite participants to discover the utility of semantic annotations to identify documents that are relevant to their research question.

Introduction

Digitization efforts by libraries and archives have greatly expanded the potential of historical and diachronic corpora to be used as research objects. Whereas paper-based research favors a selection of sources that are known *a priori* to be relevant, the digitization of sources has opened up ways to find and identify source texts that are relevant beyond the "usual suspects." Full-text search has been the primary means to retrieve and select a suitable set of sources, e.g. as a research corpus for close reading. Achieving a balanced and defensible selection of sources, however, can be highly challenging when delimiting sources with keyword-based queries. It, too often, requires scholars to handcraft elaborate queries which incorporate long synonym lists and enumerations of named entities (Huistra and Mellink, 2016).

In this tutorial, we instruct participants in the use of semantic technology to bridge the gap between the conceptually-based information needs of scholars, and the term-based indices of traditional information retrieval systems. Since we address working with corpora of a scale that defies manual annotation, we will focus on automated semantic annotation, specifically on (named) entity linking technology.

Entity Linking with Spotlight

We start the tutorial with a practical introduction to entity linking, using the open source DBpedia Spotlight software. The introduction follows several examples of entity linking in digital humanities projects, with an emphasis on search applications. Participants will be provided with a step-by-step explanation of the statistical model that forms the backbone of Spotlight and many other entity linking systems. The aim of this explanation is to provide participants with an intuition to distinguish between cases in which entity linking systems will reliably produce correct annotations, and more challenging cases which may lead to errors.

To conclude this part, participants will engage with a sequence of exercises using Spotlight through a web interface[1]. They will start off using Spotlight on text fragments that have been selected by the organizers, but may proceed to evaluate the software on documents that they have either brought to the tutorial, or can gather on the spot.

[1] Similar to https://www.dbpedia-spotlight.org/demo/, but provided by the organizers.

Exploratory Semantic Search with WideNet

The second part of the tutorial engages with exploratory semantic search from the angle of tool criticism. We explore WideNet, a tool based on entity-linking that facilitates the exploration of complex concepts in longitudinal data. Even though linking technologies such as Spotlight remain imperfect, they can be leveraged for useful research applications. WideNet has been created to assist researchers with investigating references to complex concepts in large data sets (Olieman, Beelen, and Kamps, 2017). Participants learn to apply semantically-enhanced search for the purpose of corpus building[2] (e.g. to find all documents related to the "French Revolution" or the "Dutch Golden Age") and investigate the underlying technology that enables them to perform such intricate and comprehensive queries. We discuss how entity-based search solves some of the (usability) limitations of traditional keyword search, but we also attract attention to issues that arise with semantic search.

Individual Use and Shared Infrastructure

In the final part of the tutorial, we will invite participants to follow along with an (online) Jupyter notebook that implements a lightweight entity linking and document selection workflow. It processes text with DBpedia Spotlight, and stores the resulting annotations in an embedded database. The search target, a complex concept, is resolved into concrete entities by performing SPARQL queries against the DBpedia endpoint. We will take the time to inspect the intermediate output and data structures of these steps. At the end of the notebook, the annotations on the input text are filtered and faceted by the result of the SPARQL queries, leading to a set of potentially-relevant search results.

For the remaining time, two options are provided to participants. They may choose to continue to work with the notebook, in which case we will assist to adapt the provided code to work on their own documents. This will be recommended for those who intend to apply the workflow from the notebook on locally stored documents in the future. The other participants will be encouraged to independently work with WideNet, to explore topics of their own interest in the pre-loaded corpora. The instructors will be available to answer any questions related to the usage of WideNet, and will provide information about the needed steps to load additional corpora into WideNet.

Target audience: Scholars in the fields of Historical, Heritage, Memory, and Literary Studies who wish to explore larger text corpora beyond the traditional keyword search. The tutorial also targets stakeholders who are interested in understanding humanities research practices in digital environments, such as archivists, librarians, and infrastructure providers.

Website: A tutorial website has been set up to provide materials to participants. This is a living document that will be updated in the run-up to the tutorial, and which can be used to publish the results of activity during the tutorial.

---

[2] For an impression, see our DH2017 demo <http://widenet.politicalmashup.nl/dh2017/> (older WideNet version)

Expected number of participants: 20-30

Allotted time:

- 45 minutes for Entity Linking with Spotlight
- 60 minutes for Exploratory Semantic Search with WideNet
- 75 minutes for Individual Use and Shared Infrastructure

Bibliography

Beelen, K., Olieman, A., and Kamps, J. (2017). Historical Event Search in Digital Heritage. *Joint Proceedings of SEMANTiCS 2017 Workshops: EVENTS*. ceur-ws:2063/events2.

Huistra, H. and Mellink, B. (2016). Phrasing history: Selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *49*(4): 220-229.

Olieman, A. and Beelen, K. (2017). Cast Your Net Wide: Finding Historical References in Parliamentary Data. *Digital Humanities 2017: Conference Abstracts*. Montréal: McGill University, pp. 771-72.

Olieman, A., Beelen, K., Lange, M. van, Kamps, J., and Marx, M. (2017). Good Applications for Crummy Entity Linkers?: The Case of Corpus Selection in Digital Humanities. *Proceedings of the 13th International Conference on Semantic Systems (SEMANTiCS 2017)*. Amsterdam: ACM, pp. 81-88. doi:10.1145/3132218.3132237 arXiv:1708.01162.

Olieman, A., Beelen, K., and Kamps, J. (2017). Finding Talk About the Past in the Discourse of Non-Historians. *Joint Proceedings of SEMANTiCS 2017 Workshops: Drift-a-LOD*. ceur-ws:2063/dal2. arXiv:1710.01127.

Instructor Information

Alex Olieman <olieman@uva.nl>

Alex Olieman is a PhD candidate at the Institute for Logic, Language, and Computation at the University of Amsterdam. His research focuses on the application of information extraction, specifically entity linking, to improve the accessibility of information and data. Alex is interested in working with large open datasets, and has done much of his work with government archives. In addition to his research work, he holds a position as an R&D engineer for Qollap, a communication and knowledge sharing platform. He also is an avid open source software contributor, and is an active participant in the DBpedia community.

Kaspar Beelen <KBeelen@turing.ac.uk>

Kaspar Beelen is a research associate at the Alan Turing Institute. After completing his PhD at the University of Antwerp (under the supervision of Marnix Beyen), he worked on the

Digging into Linked Parliamentary Data Project at the University of Toronto and the University of Amsterdam. In Amsterdam, he held the position of assistant professor in Digital Humanities at the Media Studies department. He focuses on computational history, more specifically on the use of text-mining for political and cultural history. His main areas of interest include: gender and politics, the history of political representation and the evolution of affective discourse.

Jaap Kamps <kamps@uva.nl>

Dr. Ir. Jaap Kamps is an associate professor of information retrieval at the University of Amsterdam, PI of a stream of large research projects on information access funded by NWO and the European Union, vice-chair of the ACM SIG-IR, organizer of evaluation efforts at TREC and CLEF, and a prolific organizer of conferences and workshops. His research interests span all facets of information storage and retrieval – from user-centric to system-centric, and from basic research to applied research. A common element is the combination of textual information with additional structure, such as document structure, Web-link structure, and/or contextual information, such as meta-data, anchors, tags, clicks, or profiles.