From Text to Networks: Combining Entity and Segment Annotations in the Analysis of Large Text Corpora

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
University of Stuttgart, Germany

Maximilian Overbeck
maximilian.overbeck@sowi.uni-stuttgart.de
University of Stuttgart, Germany

Sandra Murr
sandra.murr@ilw.uni-stuttgart.de
University of Stuttgart, Germany

Introduction

In this half-day tutorial we will offer a full-fledged, implemented and tested workflow that has been developed in the interdisciplinary *Center for Reflected Text Analytics* (CRETA, a research center connecting both scholars from Humanities/Social Sciences and Computational Linguistics at the University of Stuttgart). Our focus is the valid and reliable identification of various kinds of entities and segments from raw, un-annotated texts and the extraction of specific relational information via network visualizations. Given the recent interest in networks for data representation and visualization (e.g., Gephi-tutorial at DH 2016), we argue that the following three-step-workflow is applicable to many research questions in the Social Sciences and Humanities:

1. Detection of entity references in texts of different genres (*e.g. references to chancellor Merkel in parliamentary debates)*,
2. Segmentation of the texts guided by research questions (*e.g. parts of a parliamentary speech dealing with the Greek financial crisis*), and
3. Creation of networks of entities that co-occur within a segment *(e.g. references to national or international organizations in a parliamentary debate dealing with the issue of wars and military interventions).*

This workflow is one example of modularizing complex research questions into concrete steps and can moreover be combined with computational methods for the semi-automatic analysis of very large text corpora. The concepts of "entity" and "segment" are sufficiently generic to allow the same set of tools to be employed in different research questions originating from different fields of research. The tutorial is therefore not aimed at a specific Humanities or Social Sciences discipline and instead open to all researchers interested in the analysis of entity relations in large amounts of textual data.

In our tutorial we will make use of the web-based annotation tool CRETAnno developed to support semi-automatic annotation. CRETAnno provides tools for annotation and continuous assessment of *inter-annotator agreement*, thereby facilitating the production of reliable and valid data. Our tool facilitates the annotation of large text corpora: After some training instances are annotated, a machine learning model can be trained to predict new instances on additional texts, which can then be corrected and used as additional training material. This way, large texts can be annotated (relatively) quickly, given systematic manual annotation and clear annotation guidelines. This 3-step approach is currently investigated within the *Center for Reflected Text Analytics* (CRETA) on four distinct text corpora, connected to diverse research questions in different disciplines. Although establishing broadly applicable workflows has its merits (Kuhn & Reiter, 2015), we believe it is important to be able to "parameterise" them to take into account the specificities of a concrete research question. Research questions should govern the definition of entities, segments and weighting criteria in the network. In the tutorial, participants will be free to bring in (and work on) their own research questions (within the time limits of the tutorial).

Entity Reference Detection

Every concept of interest within a real or fictional world can be considered as an entity. Words in a text refer to th ese entities and are therefore called *entity references*. We have established annotation guidelines that distinguish six entity classes, oriented at the re-

search questions within CRETA: *Person*, *Location*, *Organization*, *Work* (e.g., a piece of art), *Event* and *Abstract Concept* (e.g., art).

While these entities are semantically diverse, their linguistic representation in texts is similar: References are either proper nouns (*Hillary Clinton*/*EU*), pronouns (*she*/*it*) or appellative noun phrases (*an American politician* / *the international organisation*).Most of the entity references consist of a few words, but we generally opt for annotating full noun phrases (e.g., *the British people after having voted for the Brexit*). In order to be able to link entities semi-automatically, we focus on appellative noun phrases and proper nouns, and ignore pronouns (see below).

The notion of "entity reference" we are aiming for differs from what is known in **Named Entity Recognition** (NER) and **Coreference Resolution** (CR). In NER, only proper nouns are detected, while CR also aims to resolve pronouns. Our notion of entity reference detection is aiming for the middle ground. By excluding pronouns, we also exclude the most ambiguous words, whose co-reference properties typically can only be judged in context of their appearance. Appellative NPs contain enough information such that we can establish their identity with proper nouns with relatively simple lists and rules.

Text Segmentation

Researchers from Humanities or Social Sciences generally want to inquire either the interaction between entities (within certain contexts) or between entities and the contexts themselves. Text segmentation is our way of operationalising this context. The notion of segment -- again -- is a generic one, to be adapted to specific research questions and/or theoretical assumptions made within a discipline or research area. Different kinds of segmentation are distinguishable: A **segmentation according to structural units** like chapters (narratives), speeches (minutes of parliamentary debates) or acts (dramatic texts) relies on the proper detection of such segments in the original texts and is therefore highly intertwined with the concrete text format at hand. Although machine learning models can be trained to perform such tasks, they likely do not generalize well to new texts. Even in TEI-encoded dramatic texts (which are strongly structured), there are a lot of options how to encode acts. We therefore aim for making it easy for researchers from Humanities and Social Sciences to detect such segments using metadata (e.g. dates of publication of a newspaper article or a parliamentary debate), text-specific regular expressions and/or rules.

A second kind of segmentation is **segmentation according to content** criteria. Depending on text genre and research question, this can mean segmentation by topic, narrative level, plot, time, location etc. One possible application is the segmentation of newspaper content according to various topics (Kantner & Overbeck 2017, forthcoming).

Structurally, segment annotations differ from entity reference annotations by being longer and thus sparser within a text. This has consequences for the semi-automatic support, because annotating a sufficient number of training instances requires more text to be read (and analysed with respect to its segmentation) and thus takes more time. CRETAnno therefore supports a number of unsupervised segmentation algorithms that can be used directly. In addition, researchers can specify text patterns using regular expressions and simple rules and thus focus the segmentation on the specific research question they have.

Entities + Segments = Networks

Given entity reference and segment annotations, it is only a small step to extract network-like data based on co-occurrence. As the entity reference annotation does not include links between annotations referring to the same entity, we developed a small tool to mark co-reference, given the annotated entity references. Currently, this has to be done manually, but we will explore automatisation possibilities in the future. Given that we can already identify string-identical entity references automatically, it is a manageable workload.
CRETAnno offers an interface to the graph exploration software Gephi, which can be used to edit, explore, inspect and visualise the network (the tutorial covers the annotation, ex- and import, but only basic functionality of Gephi.).

Tutorial

Participants will have the opportunity to work on texts of their own choosing within the first half of the workshop. To that end, they will be asked to submit their texts before the workshop. We will supply hands-on material to participants that do not submit. The tutorial focuses on hands-on sessions and active participation.

Appendix

Tutorial Instructors

All submission authors work jointly in the Center for Reflected Text Analytics (CRETA) at Stuttgart University, Germany.

Sandra Murr

Sandra Murr , is a PhD candidate in the Department of modern German literature at the University of Stuttgart. Within CRETA, she analyzes literary works of the productive reception of J. W. v. Goethe's *Sorrows of the Young Werther*, the so-called Wertheriaden, focusing on the analysis of the central character constellation with respect to emotions.

Maximilian Overbeck

Maximilian Overbeck is a PhD candidate in Political Science at the Chair of International Relations and European Integration at the University of Stuttgart. In his PhD he analyses Western debates on religion in the context of wars and armed conflicts where he uses highly innovative computational-linguistic approaches for the valid and reliable analysis of large newspaper corpora.

Nils Reiter

Dr. Nils Reiter works at the Department of Natural Language Processing and coordinates the scientific work in CRETA. Since his PhD thesis with the title Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms (Link), he is working in and for the Digital Humanities area, with a particular focus on literary texts, annotation and the operationalisation of Humanities research questions.

Target Audience

Any student or scholar interested in qualitative and quantitative text analysis is invited. Prior knowledge in text analysis techniques is not obligatory but might be helpful. Programming skills are not necessary, but familiarity with Gephi is helpful. We welcome 20 to 30 participants.

Acknowledgements

We at CRETA are grateful to the German Federal Ministry of Education and Research (BMBF) for its generous funding in the years 2016 until 2018 (project ID: 01UG1601).

Bibliography

John, M., Lohmann, S., Koch, S., Wörner, M., Ertl, T. (2016) Visual Analytics for Narrative Text: Visualizing Characters and their Relationships as Extracted from Novels. *Proceedings of the 7th International Conference on Information Visualization Theory and Applications* (IVAPP '16). SciTePress, 2016.

Kuhn, J., and Reiter, N. (2015). A Plea for a Method-Driven Agenda in the Digital Humanities. In *Proceedings of Digital Humanities 2015*, Sydney, Australia, June 2015.

Kantner, C., Overbeck, M. (2017, forthcoming): „Die Analyse ‚weicher' Konzepte mit ‚harten' korpuslinguistischen Methoden. In: J. Behnke, A. Blaette, J.-U. Schnapp & C. Wagemann (eds.) *Big data? New Data*. Baden-Baden: Nomos Verlag.

Overbeck, M., (2015). Observers turning into participants: Shifting perspectives on religion and armed conflict in Western news coverage. The Tocqueville Review/La revue Tocqueville, 36, 95-124.

Reiter, N. (2015) Towards Annotating Narrative Segments. Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 34–38, Beijing, China, July 30, 2015.