EGOlink: Supporting Editors of Online Historical Sources through Automatic Link Discovery

Hatem Mousselly Sergieh
mousselly-sergieh@ukp.informatik.tu-darmstadt.de
UKP Lab, Technische Universität Darmstadt, Germany

Michael Piotrowski
Leibniz Institute of European History (IEG), Germany
piotrowski@ieg-mainz.de

Iryna Gurevych
gurevych@ukp.informatik.tu-darmstadt.de
UKP Lab, Technische Universität Darmstadt, Germany

Introduction

EGO (European History Online) is a transcultural history of Europe on the Internet published by the Leibniz Institute of European History. The success and the consequential growth of EGO is, however, a challenge for the editorial office. The interlinking of EGO articles with each other and with external resources is an important aspect of EGO's conceptual design, so each new article has to be integrated into the existing link structure. This means not only that the new article has to be linked to relevant existing articles or online sources, but the copy editors must also check whether links to the new article need to be added to existing articles.

Doing the linking manually is a tedious task. Therefore, we aim in the EGOlink project at developing methods for (semi-)automatically linking EGO articles in order to reduce the manual effort for the editorial office and to improve the navigation for readers.

As a first step, we present a tool for visualizing and analyzing the current link structure in EGO document collection. Besides detecting problems such as under-linked articles, the analysis tool provides information that is needed by a (semi-) automatic linking system. In the next step, the automatic generation of links is addressed; the two main research questions here are: (1) the automatic identification and ranking of potential link anchors in EGO articles, and (2) the discovery of suitable link targets in other EGO articles or external resources.

Background

EGO Document Collection

EGO provides a platform for publishing refereed academic articles about the history of Europe with a special focus on modern Europe from the end of the Middle Ages up to contemporary history. EGO is bilingual, i.e., it contains articles in English and German. Most articles written in one language are translated into the other language and published in both languages. Currently, EGO contains 670 articles: 436 in German and 234 in English.

Linking articles to each other as well as to external resources is a major focus of EGO. Once a new article is submitted the editorial staff put a special effort on linking the articles to already existing EGO articles as well as to other resources or multi-media contents, such as images, interactive maps, video and audio clips.

Entity Linking

As already mentioned, linking articles is at the heart of EGO. So far, this task has been done manually. We aim to optimize this process by developing a (semi-)automatic linking tool based on state-of-the-art approaches while taking the peculiarities of EGO into account.

The problem of (semi-)automatically linking EGO articles to each other or to external resources can be modeled as link discovery, or entity linking (Larson, 2010). In general, entity linking consists of two main steps (Erbs et al, 2011): 1) anchor discovery, which identifies text mentions that can function as link anchors, and 2) target discovery which identifies the best matching references for each anchor from a set of candidates. While anchor discovery can be handled independently of the document collection at hand – e.g. using NER (Named Entity Recognition) (Nadeau et al, 2007) or keyphrase extraction methods (Turney, 2000), discovering and ranking the targets are much affected by the nature of the document collection.

Currently, several state-of-the-art algorithms for entity linking have been made available (Hoffart et al, 2011) (Ceccarelli et al, 2013). However, applying such algorithms directly to perform entity linking based on EGO is challenging. In general, current algorithms rely on extracting statistics about candidate entities, e.g., popularity-based priors; form a huge corpus like Wikipedia. The priors are then used during the process of the target discovery and ranking. However, such an approach is not applicable in our case due to the tiny size of EGO compared to other similar resources like Wikipedia. Another challenge is posed by the German part of EGO. Current approaches focus on English and a special effort should be made to address entity linking for German.

Erbs et al (2011) demonstrated that approaches for target ranking that leverage the linking structure available in the document collection perform best compared to other methods that only leverage the titles of the documents or the associated texts. In light of this result, we propose a tool for analyzing the linking structure in EGO's document collection. The proposed tool helps us to understand the current status of EGO's links and to gain insight regarding the decisions that should be taken to develop the actual entity linking system. Besides identifying links among the articles, the proposed tool also extracts information from an additional resource, namely the literature referenced by the articles. Our hypothesis is that articles sharing the same reference should be linked to each other. As a result, this kind of analysis can help the (semi-) automatic linking system to identify further anchors, thus the interlinking density of the document collection can be increased.

EGO Analyzer

EGO Analyzer is a web-based analysis tool for EGO document collection. The tool can handle the English as well as the German parts of EGO jointly or separately. Since EGO's articles exist in XML format, the tool was built to work on XML data directly and efficiently using the high-performance and scalable XML Database engine BaseX. Besides providing different kinds of metadata about EGO's articles, the tool allows analyzing EGO from different perspectives: the corpus, the article and the reference perspectives (Figure 1).

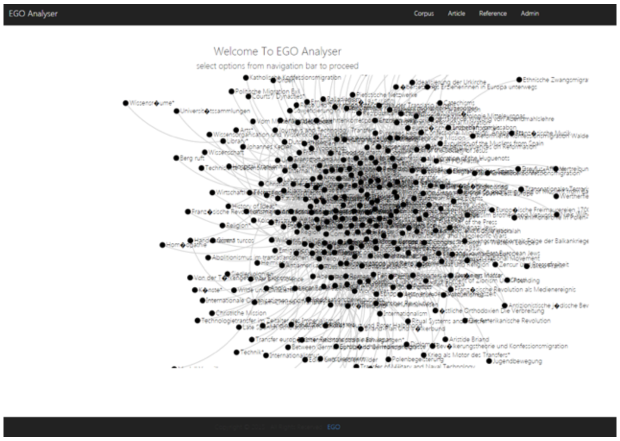
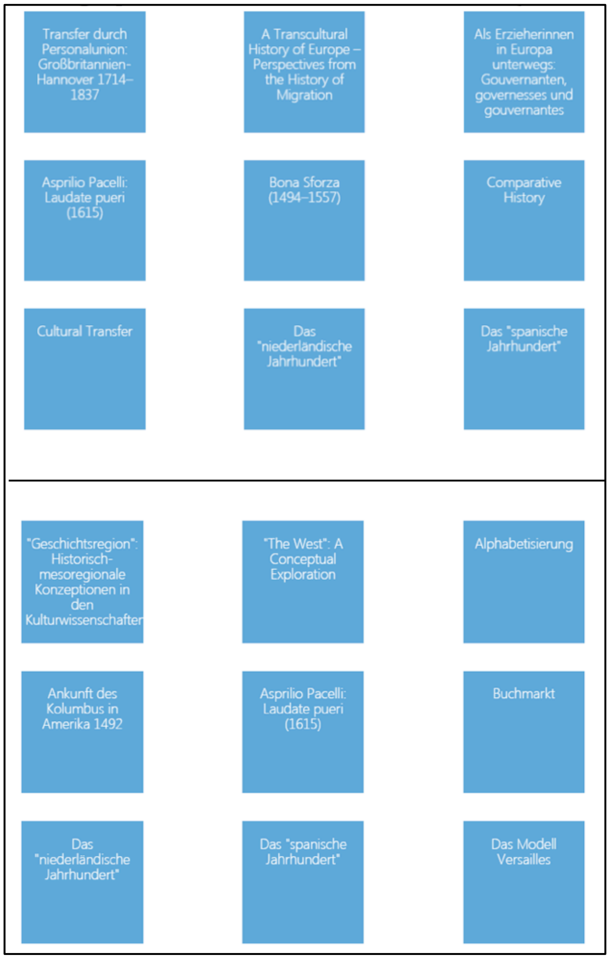


Figure 1. Screenshot: EGO Analyzer

For instance, given the article titled "Kulturtransfer" from the German collection, the tool extracts links to other articles in both directions, i.e., outgoing as well as incoming links (Figure 2). Furthermore, we can view the set of articles that reference the same bibliographical entries as the source article (Table 1).



Transfer durch Personalunion: Großbritannien-Hannover 1714–1837	A Transcultural History of Europe – Perspectives from the History of Migration	Als Erzieherinnen in Europa unterwegs: Gouvernanten, governesses und gouvernantes
Asprilio Pacelli: Laudate pueri (1615)	Bona Sforza (1494–1557)	Comparative History
Cultural Transfer	Das "niederländische Jahrhundert"	Das "spanische Jahrhundert"
"Geschichtsregion": Historisch-mesoregionale Konzeptionen in den Kulturwissenschaften	"The West": A Conceptual Exploration	Alphabetisierung
Ankunft des Kolumbus in Amerika 1492	Asprilio Pacelli: Laudate pueri (1615)	Buchmarkt
Das "niederländische Jahrhundert"	Das "spanische Jahrhundert"	Das Modell Versailles

Figure 2. Source article: “Kulturtransfer”. Top: articles referred to by the source article (outgoing links). Bottom: articles referring to the source article (incoming links)

Article	Common Bibliography
Cultural Transfer	5
Eine transkulturelle Geschichte Europas – migrationsgeschichtliche Perspektiven	2
Europäisierungen	2

Table 1. Articles sharing bibliographical entries with the source article “Kulturtransfer”

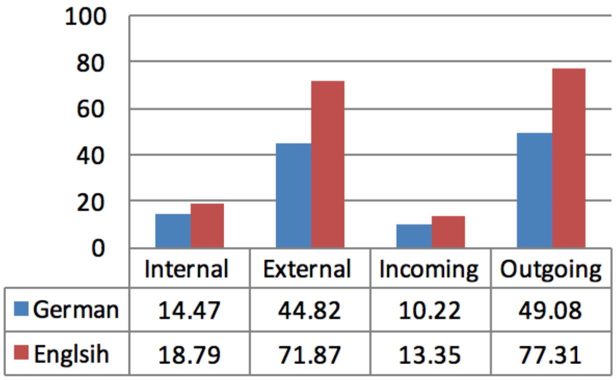
Results of the Analysis

Using EGO Analyzer we collected statistics about the linking structure in EGO. We divided the links into two categories: internal and external. Internal links point from one EGO article to another, while external links refer to external targets. For each article, we also distinguish between two types of links: incoming and outgoing links. Figure 3 shows the average number of internal/external as well as incoming/outgoing links per article.

The results show that dominance of the external links (3 times as much as internal links). This can be explained due to the relatively small size of the document collection. Accordingly, it is hard for the editors to find corresponding internal links for entities in a newly added article. Moreover, this also explains the higher number of outgoing links per article. Indeed, most of these links go to external resources which are not necessarily linked back to EGO articles. This observation demonstrates the urgent need for an automatic linking system that enables the editors to increase the interconnectivity of EGO.

Furthermore, the linking tool can also run as a background process that updates the linking structure of EGO every time a new article is added.

Regarding the shared bibliography, we identified 13,309 unique references, 737 of them are referenced by at least two articles. We will use this kind of information as auxiliary input for the linking algorithm to identify additional linking candidates.



	Internal	External	Incoming	Outgoing
🟦 German	14.47	44.82	10.22	49.08
🟥 Englsih	18.79	71.87	13.35	77.31

Figure 3: Average number of links per article and category

Conclusion

In this abstract, we presented work in progress for assisting editors of EGO to (semi-)automatically link the articles. We focused on the first step towards such a system, namely, analyzing the link structure in the target document collection. We presented a web-based tool that is able to perform such analysis. Currently, we are working on the actual linking system as well as a user-friendly interface for the editorial office. By the time of the conference, we will also provide a demonstration of the actual linking tool.

Bibliography

Ceccarelli, D., et al. (2013). Dexter: an open source framework for entity linking. Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval. 2013, pp. 17--20.

Erbs, N., Zesch, T., and Gurevych, I. (2011). Link discovery: A comprehensive analysis. Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. 2011, pp. 83--86

Hoffart, J, et al. (2011). Robust disambiguation of named entities in text. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011, pp. 782--792.

Larson, R. R. (2010). Information retrieval: Searching in the 21st century; human information retrieval. s.l. : Wiley Online Library.

Nadeau, D., and Sekine, S. (2007). A survey of named entity recognition and classification. 2007, Vol. 30, 1, pp. 3--26.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. 2000, Vol. 2, pp. 303--336.