Wiki Workshop (11th edition) – June 20, 2024

Aspect-Driven Structuring of Historical Dutch Newspaper Archives using Wikimedia Data

Hermann Kroll,
Bill Matthias Thang,
Wolf-Tilo Balke
TU Braunschweig, Germany

Christin Katharina Kreutz
TH Köln, Germany

Mirjam Cuper
KB, National
Library of the Netherlands

Abstract

Historical figures and their roles are particularly interesting cognitive access points in historical research. Structuring news articles would allow more sophisticated access for users to explore a digital library's content. In the past, we introduced an aspect-driven approach that structures news articles on historical persons at the showcase of the National Library of the Netherlands. However, real-world limitations such as the lack of training data, licensing restrictions, and non-English text are typical challenges for a digital library when implementing such a system. We tackled the lack of training data by building upon freely available Wikipedia data. This work summarizes our previous research and focuses on its relevance for the Wikimedia community (data used, benefits, problems, and open challenges).

Keywords: Wikipedia, Wikidata, Dutch, Historical News Archives, Digital Libraries

Introduction

This work is an extended abstract of our previous research (Kroll et al., 2023). Users of digital libraries featuring historical news articles conduct a variety of information interactions such as task planning or searching for and working with information objects. In historical research, historical figures and especially their roles are particularly interesting cognitive access points. Related digital library projects have been proposed in the past, e.g., ANNO (Müller, 2004), or Cuper's work (Cuper, 2021). However, those systems usually rely either on manual curation (Cuper, 2021) or at least domain-specific training examples for every implemented step.

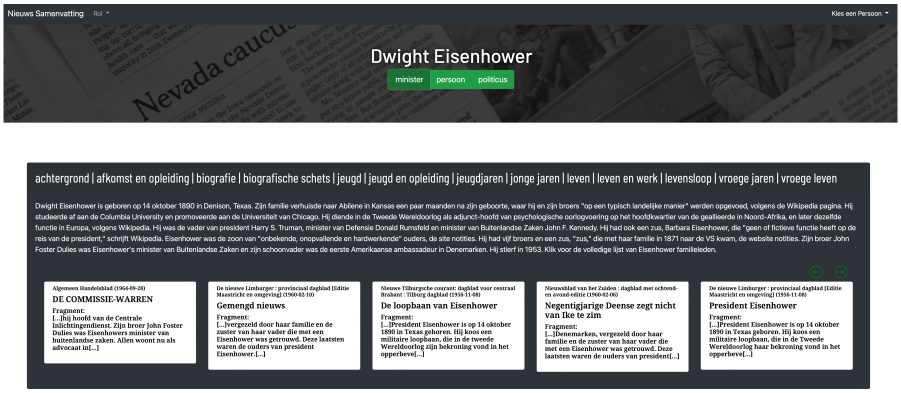
In close cooperation with an actual digital library, namely the National Library of the Netherlands, Koninklijke Bibliotheek (KB) (`https://www.kb.nl`), we bypassed a manual curation and the collection of domain-specific training data by utilizing data from Wikipedia (texts, sections, info boxes, and categories). Our approach automatically structured historical news articles on persons and provided an aspect-driven interface to explore the library's content. The central idea is that a person has different roles (e.g., *writer*, *politician*, *military person*), and each role has different aspects related to it (e.g., *early life, political career, actions*). Our approach identified and classified relevant news articles into sub-corpora, corresponding to a person's role and aspect to support research on historical persons. To implement a prototype, we used a subset of the KB's data since the KB collected news articles from the 17th century to the recent past. We selected articles on nine famous persons with various roles in the Second World War because the KB identified the topic as one users were interested in. Figure 1 shows a screenshot of our prototype. We also recorded a video[1] to introduce our prototype.

Wikimedia Data

Infoboxes and Categories. We used the Wikipedia info boxes to derive possible roles for people. The information was linked to Wikipedia categories, which are organized in a taxonomy, e.g., a *British politician* is a specialization of a *politician*. In our context, we understood a person's occupation as their role. We crawled the Dutch *occupation* categories and derived a list of occupations (in sum 30k distinct ones). Then, we iterated through the Dutch Wikipedia XML dumps (March 2023), parsed the info boxes, and checked whether a property of the info box was linked to one of those occupations. If so, we extracted the corresponding page's summary (introduction), sections, and all occupations. In sum, we derived 259k person pages, from which we observed many short ones, e.g., including a brief summary or a single section. We applied filters (had a summary length < 150 characters or < 3 sections) to retrieve representative or well-rounded person pages. Note that we disregarded sections with less than 100 characters and sections that only contained references/literature by using a hand-crafted list. This filtering reduced the number of person pages to 61k. With that, we obtained thousands of Wikipedia pages per role.

Articles. Wikipedia sections should, at best, describe one unit of information belonging to a particular aspect of a person. However, Wikipedia is crafted collaboratively through human editing. Section titles are thus usually not canonicalized. For instance, *life*, *background*, and *curriculum vitae/resume* describe the same, or at least a very

[1] `https://www.youtube.com/watch?v=0GzIydjts2E`



Dwight Eisenhower

achtergrond | afkomst en opleiding | biografie | biografische schets | jeugd | jeugd en opleiding | jeugdjaren | jonge jaren | leven | leven en werk | levensloop | vroege jaren | vroege leven

Dwight Eisenhower is geboren op 14 oktober 1890 in Denison, Texas. Zijn familie verhuisde naar Abilene in Kansas een paar maanden na zijn geboorte, waar hij en zijn broers "op een typisch landelijke manier" werden opgevoed, volgens de Wikipedia pagina. Hij studeerde af aan de Columbia University en promoveerde aan de Universiteit van Chicago. Hij diende in de Tweede Wereldoorlog als adjunct-hoofd van psychologische oorlogvoering op het hoofdkwartier van de geallieerde in Noord-Afrika, en later dezelfde functie in Europa, volgens Wikipedia. Hij was de vader van president Harry S. Truman, minister van Defensie Donald Rumsfeld en minister van Buitenlandse Zaken John F. Kennedy. Hij had ook een zus, Barbara Eisenhower, die "geen of fictieve functie heeft op de reis van de president," schrijft Wikipedia. Eisenhower was de zoon van "onbekende, onopvallende en hardwerkende" ouders, de site notities. Hij had vijf broers en een zus, "zus," die met haar familie in 1871 naar de VS kwam, de website notities. Zijn broer John Foster Dulies was Eisenhower's minister van Buitenlandse Zaken en zijn schoonvader was de eerste Amerikaanse ambassadeur in Denemarken. Hij stierf in 1953. Klik voor de volledige lijst van Eisenhower familieleden.

Algemeen Handelsblad (1964-09-28)
DE COMMISSIE-WARREN
Fragment:
[...]hij hoofd van de Centrale Inlichtingendienst. Zijn broer John Foster Dulies was Eisenhowers minister van buitenlandse zaken. Allen woont nu als advocaat in[...]

De nieuwe Limburger : provinciaal dagblad [Editie Maastricht en omgeving] (1960-02-10)
Gemengd nieuws
Fragment:
[...]vergezeld door haar familie en de zuster van haar vader die met een Eisenhower was getrouwd. Deze laatsten waren de ouders van president Eisenhower.[...]

Nieuwe Tilburgsche courant: dagblad voor centraal Brabant : Tilburg dagblad (1956-11-08)
De loopbaan van Eisenhower
Fragment:
[...]President Eisenhower is op 14 oktober 1890 in Texas geboren. Hij koos een militaire loopbaan, die in de tweede Wereldoorlog zijn bekroning vond in het opperbeve[...]

Nieuwsblad van het Zuiden : dagblad met ochtend- en avond-editie (1960-02-06)
Negentigjarige Deense zegt nicht van Ike te zim
Fragment:
[...]Denemarken, vergezeld door haar familie en de zuster van haar vader die met een Eisenhower was getrouwd. Deze laatsten waren de ouders van president[...]

De nieuwe Limburger : provinciaal dagblad [Editie Maastricht en omgeving] (1956-11-08)
President Eisenhower
Fragment:
[...]President Eisenhower is op 14 oktober 1890 in Texas geboren. Hij koos een militaire loopbaan, die in de Tweede Wereldoorlog haar bekroning vond in het opperbeve[...]

Figure 1: User interface of our system (taken from our research (Kroll et al., 2023)).

similar, aspect of a person. We designed a canonicalization step based on a pre-trained sentence transformer model (BERT-base-dutch-cased) to cluster semantically similar sections. In this way, we merged noisy, human-crafted section titles into canonicalized aspects.

In brief, we mined frequent role aspects by counting how often the aspect was used across all persons of a role (e.g., *writer*s). Given a particular person's role, we trained a classifier to predict whether some text belongs to one of the role's aspects. That means we headed for a multi-class classification scenario, e.g., a classifier for role r_1 with aspects a_1, a_2, a_3 must predict one of the aspects, or the negative class (not belonging to the role) for each news article of a historical person. Here, we retrieved Wikipedia section texts for each aspect and used them to fine-tune the Dutch model RobBERT-2022 (Delobelle et al., 2022) for text classification. We trained a classifier for each role (occupation category of Wikipedia) that had (1) at least three frequent aspects and (2) belongs to the first two category levels in Wikipedia (to select more general roles like *writer* instead of *British writer*).

Finally, we processed the Dutch news articles, derived a person's role from their Wikipedia info box, selected the pre-trained language models for their role(s), and performed the article to aspect classification. Code is available at GitHub[2] and Software Heritage[3].

Lessons Learned

In conclusion, Wikipedia allowed us to build a reliable prototype for an actual digital library by utilizing its info boxes, categories, and article data. However, we faced some issues when working with Wikimedia data: (1) Filtering, (2) Retrieving, and (3) Wikidata. (1) Wikipedia pages differ in their quality and level of detail. Crawling data and applying custom filter criteria requires extra effort. It would have simplified our approach if Wikipedia offered additional dumps (e.g., articles of a certain quality/verification level). (2) Of course, downloading the whole Wikipedia dump and filtering for personal articles is possible but exhausting. We would have loved to crawl a specific subset of Wikipedia, but we were not aware of how to identify the set of Wikipedia articles about persons. (3) In a perfect world, we would have used Wikidata to derive a person's page, roles, and aspects. However, links to the Dutch Wikipedia were incomplete and our trials revealed that the Wikipedia info boxes contained more information, and working with them (and their corresponding category links) was easier for us. Understanding which information in Wikidata could reflect a person's role and then querying them was challenging.

References

[Cuper2021] Mirjam Cuper. 2021. Researching pandemics through time: A covid-19 inspired data-driven approach to explore historical newspapers. In *25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021*, pages 227–231. Springer.

[Delobelle et al.2022] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2022. Robbert-2022: Updating a dutch language model to account for evolving language use. *CoRR*, abs/2211.08192.

[Kroll et al.2023] Hermann Kroll, Christin Katharina Kreutz, Mirjam Cuper, Bill Matthias Thang, and Wolf-Tilo Balke. 2023. Aspect-driven structuring of historical dutch newspaper archives. In *27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023*, pages 31–46. Springer.

[Müller2004] Christa Müller. 2004. *ANNO-AUSTRIAN NEWSPAPERS ONLINE: Historische österreichische Zeitungen und Zeitschriften online. Eine Digitalisierungsinitiative der Österreichischen Nationalbibliothek (http ://anno.onb.ac.at/).* K. G. Saur.

[2] https://github.com/HermannKroll/AspectDrivenNewsStructuring

[3] https://archive.softwareheritage.org/swh:1:dir:13457c154ed7ad1f571e353c1edf2f87db61b0ae

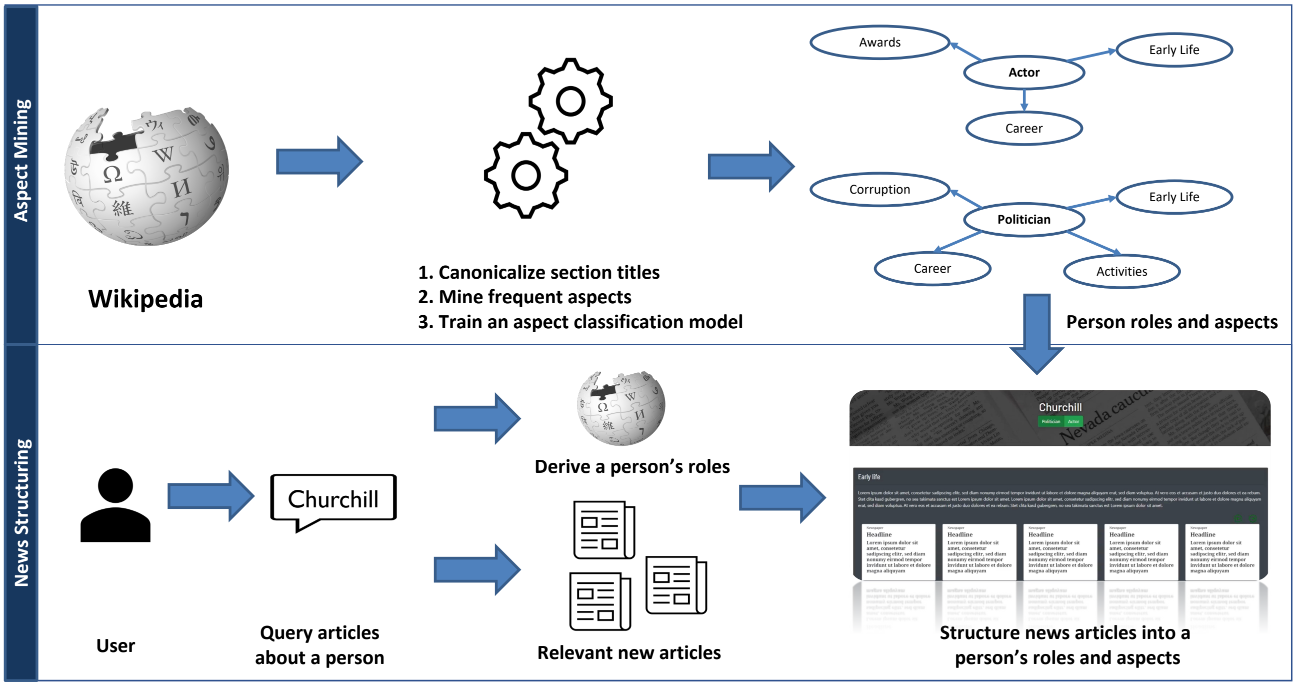


Figure 2: Systematic overview: We used Wikipedia info boxes to derive possible roles (e.g., *writer*, *politician*). Then, we processed Wikipedia article data to derive frequent aspects (e.g., *early career*, *political career*, *background*) for each role. We applied clustering to canonicalize Wikipedia section titles. Finally, the obtained and categorized texts were used to train a language model for text classification. This model then helped us to classify the KB's news articles.