Linking Historical Entities to the Linked Open Data Cloud

Marx, M.

Publication date
2014
Document Version
Final published version
Published in
ERCIM News

Linking Historical Entities to the Linked Open Data Cloud

by Maarten Marx

We investigate the coverage of Wikipedia for historical public figures. Unsurprisingly, the probability of a figure having a Wikipedia entry declines with time since the person was active. Nevertheless, two thirds of the Dutch members of parliament that have been active in the last 140 years have a Wikipedia page. The need to link historical figures to existing knowledge bases like Wikipedia/DBpedia comes from current large scale efforts to digitize primary data sources, including proceedings of parliament and historical newspapers. Linking entries to knowledge bases can provide values of key background variables, such as gender, age, and (party) affiliation.

The term "wikification" [1] refers to the process of automatically creating links from words or phrases in free text to their appropriate Wikipedia page. The common motivation for wikification is that a reader may want to consult additional (background) information about the phrase while reading the text. Typical candidates for wikification are named entities and rare terms. Another motivation for wikification comes from information retrieval: linking named entities in texts to external knowledge bases can improve both precision (by disambiguating names) and recall (by including spelling variants obtained from the knowledge page). A third motivation comes from the new fields of Computational Humanities and Computational Social Science [2] and is driven by current large-scale efforts to digitize primary historical sources and archives. Primary sources typically do not contain common background information about the entities (persons, organizations) mentioned in the sources. For instance, for each word spoken in the proceedings of the British parliament (Hansards), the name of the speaker and the speaker's constituency are recorded. But key variables such as age, gender, and even party affiliation are not recorded.

Wikification of Historical Texts: Research Challenges

Clearly, if wikification is desirable for texts from our own age, it is even more so for digitized historical archives. Performing wikification on historical scanned documents has several challenges that are not present in modern digital material: named entity recognition (NER) must deal with OCR-errors, spelling reforms and mismatches in language use in modern text material (on which NER taggers are trained) and historical texts. Disambiguating named entities (a key part in linking to external knowledge bases) may be harder because less background information is available. Finally, it may be that there is simply no data to link to. This latter aspect is addressed in greater detail below.

Coverage of Historical Persons in Wikipedia

We investigated Wikipedia's coverage of historical public figures using the

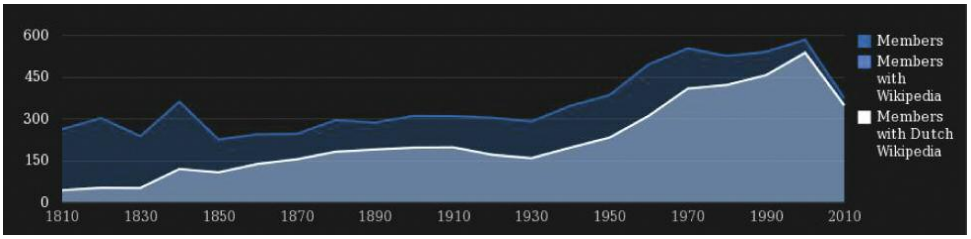

Figure 1: Coverage of Wikipedia for Dutch politicians in the period 1810-2013.

Politician	Party	# languages
Drs. M. (Mark) Rutte	(VVD)	40
G. (Geert) Wilders	(VVD)	30
Drs. A. (Ayaan) Hirsi Ali	(VVD)	23
Prof.Mr. J.G. (Jaap) de Hoop Scheffer	(CDA)	21
Dr. W. (Willem) Drees	(SDAP)	15
Drs. N. (Neelie) Kroes	(VVD)	15
Dr. A. (Abraham) Kuyper	(antirevolutionair)	14
Dr. J. (Jelle) Zijlstra	(ARP)	14
Mr. W. de Sitter	(liberaal)	13
Prof. Dr. Mr. F. (Frits) Bolkestein	(VVD)	13

Table 1: Number of Wikipedia pages in different languages per politician. Top 10 of the Dutch politicians.

Dutch parliamentary proceedings, which are available as scanned and OCRed PDF files from 1814 (http://www.statengeneraaldigitaal. nl/). Speakers in the proceedings are linked to a biographical database (http://www.parlement.com), which has an entry for each MP in this period. We linked the full names of the MPs taken from this database to the Dutch Wikipedia using the linking methodology of [3]. This software yields a ranked list of Wikipedia page candidates for each input. We filtered this list using a "political biography page" filter, which effectively removed false positives. Automatic checking of correctness of the links was facilitated by the fact that the majority of political biography pages on Wikipedia linked back to the database we used. Manual inspection and search on Wikipedia showed that with this automatic process we discovered virtually all existing biographical Wikipages.

Figure 1 presents the results. We grouped MPs in ten year periods. The top line in the graph shows the number of MPs that were active for at least one day during each period. The line below shows the number of them having a Wikipedia page. We have a coverage of over 90% for the period after 2000, at least 66% for the period 1850-2000 and a minimum of 16% for the period 1810-1820. The dip in 2010 is caused by the fact that this period consists only of three years.

The Most International Dutch Politicians

To illustrate the benefits of having the links we show the top 10 Dutch politicians with Wikipedia pages in most languages, an indicator of how well-known these people are internationally. This top 10 is made up of a rather varied club of politicians. First place is occupied by the current prime minister; second and third by leading anti-Muslim politicians. Fourth is the former secretary general of NATO. Place five shows the first person who is no longer alive: the Dutch PM in the period 1948-1958. Positions seven, eight and nine also have historical rather than modern politicians.

Conclusion

Linking entities occurring in historical material to present day knowledge bases like Wikipedia is an exciting and rewarding research field with great potential benefits. Wikipedia is rich enough to cover at least two thirds of the Dutch MPs active in the last 140 years. The linking process has the additional benefit of showing gaps in existing knowledge bases.

References:

[1] R. Mihalcea, A. Csomai: "Wikify!: Linking documents to encyclopedic knowledge", in proc. of CIKM '07, pp 233-242, 2007, http://dx.doi.org/10.1145/1321440.1321475

[2] D. Lazer, et al.: "Computational social science", Science, 323(5915):721-723, 2009, http://dx.doi.org/10.1126/science.1167742

[3] E. Meij, W. Weerkamp, M. de Rijke: "Adding semantics to microblog posts", in proc. of WSDM 2012, http://dx.doi.org/10.1145/2124295.2124364

Please contact:

Maarten Marx
University of Amsterdam, The Netherlands
E-mail: maartenmarx@uva.nl