Entity Centric Historical Text Mining

A people-centric approach to modern Dutch religious history

Mariona Coll Ardanuy
Georg-August University Göttingen
mcollar@uni-goettingen.de

Maarten van den Bos
Utrecht University
m.j.a.vandenbos@uu.nl

The mass digitization of newspapers in concordance with the development of text mining techniques promises new possibilities in furthering historical research. However, dealing with large amounts of unstructured data still poses great challenges to natural language processing. In our paper, we propose the use of networks of personal entities (disambiguated person names) as a sophisticated strategy to approach historical data. As a use case, we focus on recent religious history and the important role of historical actors as drivers of change in the postwar debate on the role of religion in modern society. [1] In previous work, we have demonstrated a method to automatically create social networks from news and showed the possibilities of the method to further new European integration history. [2] Now we want to take our method one step further. Whereas in our previous paper we did not study the problem of name ambiguity, we address this problem here.

Person name disambiguation is probably the greatest challenge automatic network creation faces: one same person name can refer to several entities whereas one same entity can be referred to by many different names. In order to illustrate this problem, we will focus on two different actors who have played a pivotal role in postwar Dutch debate on the role and position of religion in the Netherlands. As both Michael Burleigh and Dominic Sandbrook have demonstrated, understanding this debate is crucial for a better understanding of societal changes in the 1960s. [3] And whereas previous studies on recent religious history were mostly focussed on larger, anonymous and long-term societal processes such as secularization, recent works emphasize the importance of ideas and the role of individual actors as important drivers of change. [4]

In our paper, we will focus on Willem Banning and Edward Schillebeeckx. Schillebeeckx was a member of the Dominican order, professor in theology at the Catholic University of Nijmegen and a prominent advisor of the Dutch bishops during, before and after the Second Vatican Council. Banning was professor in the sociology of religion at Leiden University and director of the Sociological Institute of the Dutch Reformed Church. Both played a pivotal role in the religious transformations of the postwar years. Banning was a leading intellectual in the movement responsible for a major transformation within the Reformed Church in the postwar years and Schillebeeckx was a prominent member of an international network of progressive theologians who deeply influenced discourse on

the future of the Catholic Church. [5] It would be most interesting to further investigate their public profile, using the digitized newspaper collection of the Dutch National Library, and gain a better understanding of people and topics they were connected with in contemporary press. Also, comparing the networks they were part of and topics related can provide valuable information on the course of public discussion on the future of religion in Dutch society.

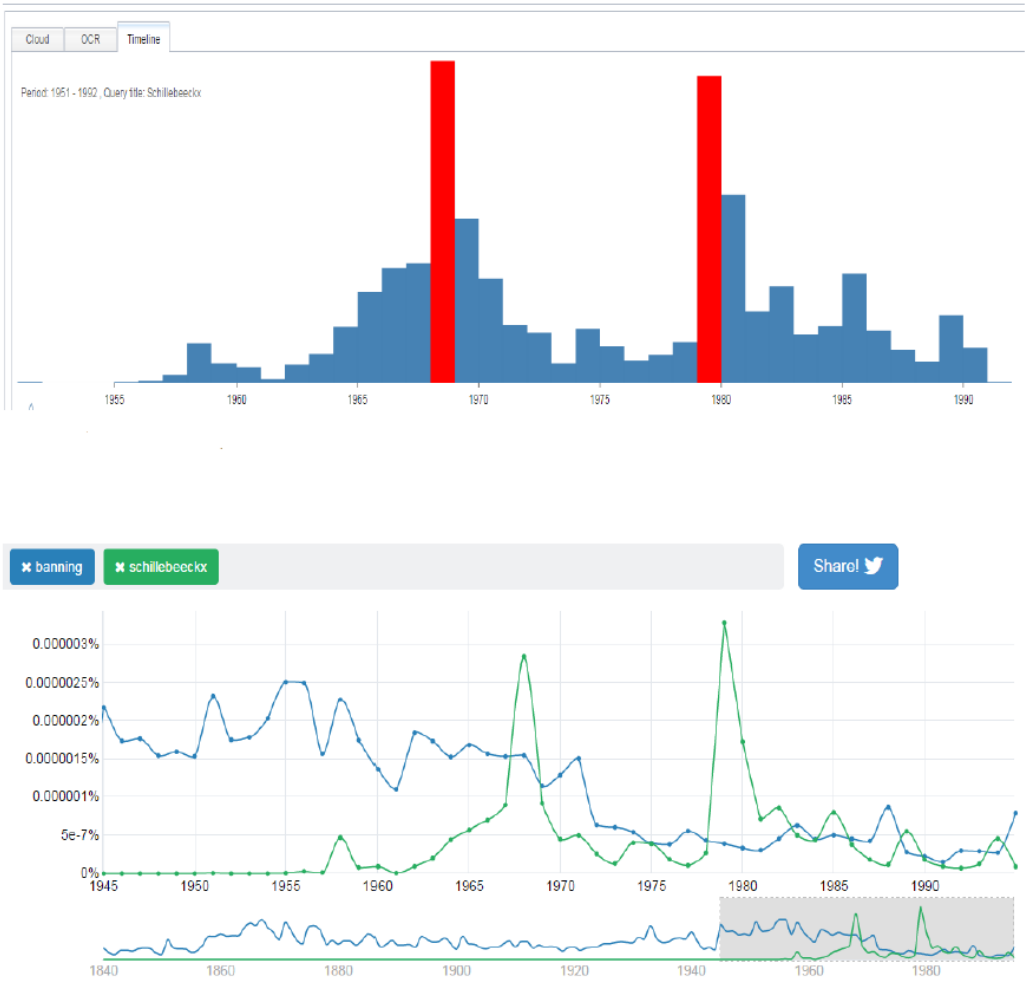

Figure 1: Timeline of articles mentioning 'Schillebeeckx' and n-gram of articles mentioning 'Banning' and 'Schillebeeckx' in the digital newspaper collection of the Dutch National Library. Using this method, we cannot be sure articles mentioning Banning all actually refer to Willem Banning. The large difference in total number of articles on Schillebeeckx (2.796) and Banning (26.984) is a first hint that results on Banning contain articles on multiple entities. [6]

Comparing Schillebeeckx and Banning is interesting also from a technical point of view. Whereas Schillebeeckx has a very distinct name, the name 'Banning' was much more common in the newspapers. In this sense, whereas articles mentioning the name 'Schillebeeckx' are very likely to refer to the person we are interested in, this is not clearly so in the case of the name 'Banning'. An artificially-created network which does not deal with this problem is likely to suffer obvious problems of either including different people with the same name or missing articles referring to Banning by last name only when queried for "Willem Banning".

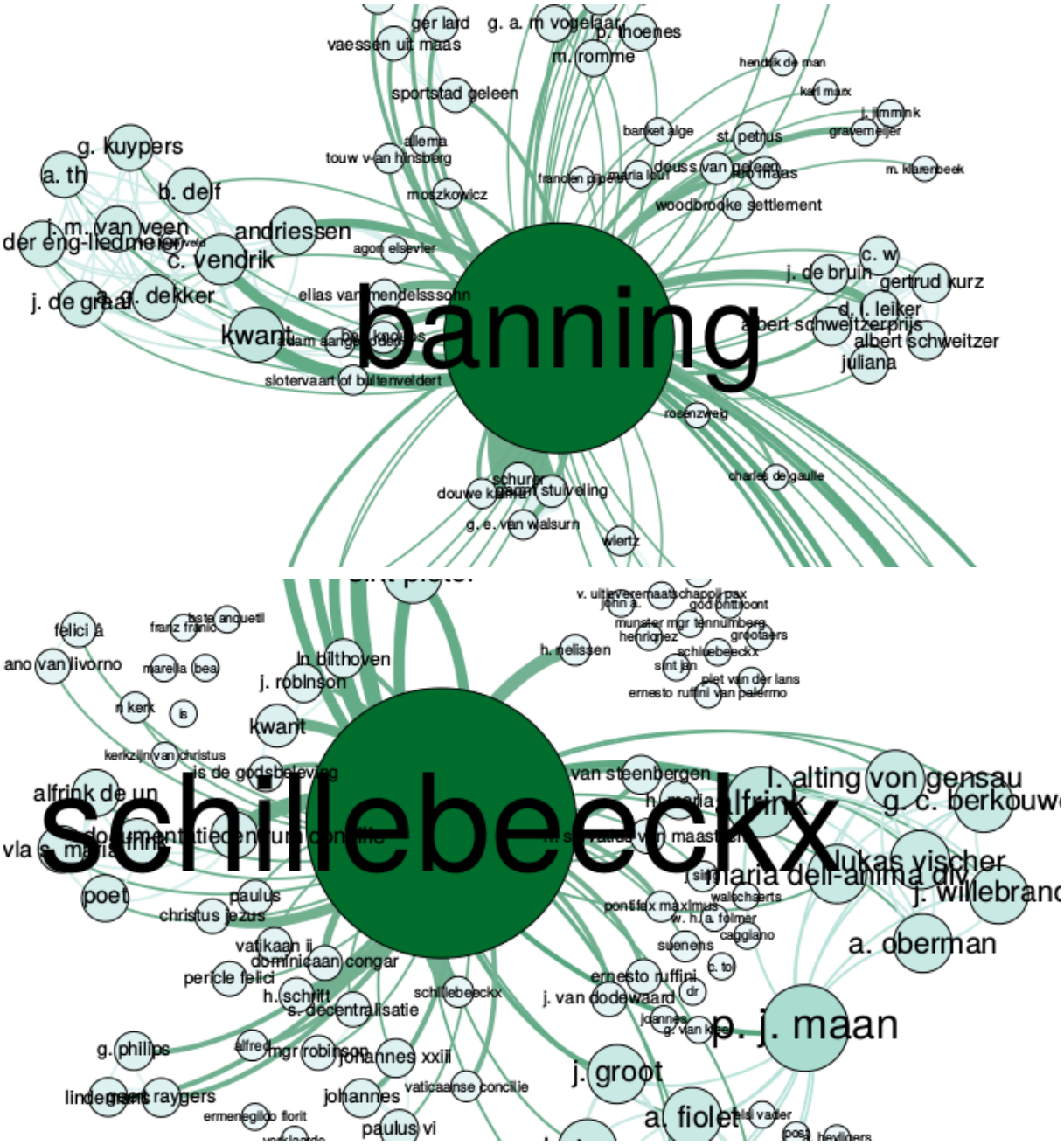

Figure 2: Fragments of the networks drawn from the articles from 1964.

In order to tackle this problem, we have developed a disambiguation module which relies on the social dimension of news. Our system exploits the relation between a person name's ambiguity (calculated from the Dbpedia Persondata [7] database) and the number of entities referred to by it. Named entities were recognized by means of Stanford NER [8] trained on modern day Dutch data. Since we assume that the same person name within an article refers always to the same entity, first step is to group all completely – or partially – matching surface forms of person names together. Then, each news article is converted into a network: the person names mentioned become the nodes and are linked together through an edge when occurring in the same paragraph. Modelled as a weakly-supervised clustering problem with a strong focus on social relations, our system dynamically adapts the strategy according to how ambiguous a name is. When the ambiguity of a name is low ('Schillebeeckx'), the probability that two documents containing this name refer to the same entity is

much higher than in case of a more ambiguous name ('Banning'). Finally, since an approach relying only on networks produces high precision at the cost of recall, we also use context similarity between documents. The output of the disambiguation module is the decision whether or not the two articles refer to the same entity. If the module concludes they do, the networks are merged. The cluster of articles which has the highest context similarity to a biographical note of each actor is taken as his social network and is manually validated.

In our paper, we will further explain our method, its potential and difficulties, and use the comparison between the networks of Banning and Schillebeeckx as a use case in order to demonstrate its value for humanities research. Since our method is both language-independent and highly adaptable to different experiments and data sources, we think it can be valuable for many different research questions.

[1]Building on earlier research: Maarten van den Bos, *Verlangen naar vernieuwing. Nederlands katholicisme, 1953-2003* (Amsterdam: Wereldbibliotheek 2012); Idem, 'Een nieuwe bijdrage aan de receptiegeschiedenis van het Tweede Vaticaans Concilie', *Religie en Samenleving* 9, 3 (2014) 211-233.

[2]Mariona Coll Ardanuy, Maarten van den Bos, and Caroline Sporleder, 'Laboratories of Community. How Ditigal Humanities Can Further New European Integration History', *Social Informatics – SocInfo 2014 International Workshops, Barcelona, Spain, November 11 2014, Revised Selected Papers* (Cham: Springer 2015) 284-293.

[3]Michael Burleigh, *Sacred Causes. The Clash of Religion and Politics, from the Great War to the War on Terror* (New York: Harper 2006) 348; Dominic Sandbrook, *White Heat. A History of Britain in the Swinging Sixties* (London: Abacus 2006) 458. For the Dutch case, see also: Piet de Rooy, *A Tiny Spot on the Earth. The Political Culture of the Netherlands in the Nineteenth and Twentieth Century* (Amsterdam: Amsterdam University Press 2015) 244-245.

[4]Hugh McLeod, 'Why were the 1960s so religiously explosive?', *Nederlands Theologisch Tijschrift* 60, 2 (2006) 109-130.

[5] Maarten van den Bos and Stephan van Erp, *A Happy Theologian. A Hundred Years of Edward Schillebeeckx* (Nijmegen: Valkhof Pers 2015); R. Hartmans, H. Noordegraaf and R. van de Woude, 'Willem Banning. Opvoeder van het volk' in: P. Werkman and R. van der Woude, *Bevlogen Theologen. Geëngageerde predikanten in de negentiende en twintigste eeuw* (Hilversum: Verloren 2012) 287-316.

[6] Searches are conducted and analysed using Texcavator (http://texcavator.surfsaralabs.nl). Cf. J. van Eijnatten, T. Pieters and J. Verheul, 'TS Tools: Using Texcavator to map public discourse', *Tijdschrift voor Tijdschriftstudies* 35 (2014) 59-65; The n-Gram was generated by the n-Gram viewer for Dutch National Library Newspaper Collection (http://kbkranten.politicalmashup.nl).

[7] http://wiki.dbpedia.org/downloads

[8] http://nlp.stanford.edu/software/CRf-NER-shtml