You Shall Know People by the Company They Keep: Person Name Disambiguation for Social Network Construction

Mariona Coll Ardanuy
Göttingen University
Göttingen Centre for
Digital Humanities
mcollar@uni-goettingen.de

Maarten van den Bos
Utrecht University
Dept. of History and
Art History
M.J.A.vandenBos@uu.nl

Caroline Sporleder
Göttingen University
Göttingen Centre for
Digital Humanities
csporled@uni-goettingen.de

Abstract

The growth of digitization in the cultural heritage domain offers great possibilities to broaden the boundaries of historical research. With the ultimate aim of creating social networks of person names from news articles, we introduce a person name disambiguation method that exploits the relation between the ambiguity of a person name and the number of entities referred to by it. Modeled as a clustering problem with a strong focus on social relations, our system dynamically adapts its clustering strategy to the most suitable configuration for each name depending on how common this name is. Our method's performance is on par with the state-of-the-art reported for the CRIPCO dataset, while using less specific resources.

1 Introduction

Resolving person names across documents is an open problem of unquestionable importance in natural language processing. Person names represent 30% of the overall number of queries in the web domain (Artiles et al., 2005), and have an equally significant presence in the news domain, where people are often at the core of the events reported in articles. This is particularly interesting in historical research. As more and more historical newspapers are digitized, new potentialities arise to explore history in a way that was infeasible until recent years. People are drivers and carriers of change, and newspapers have traditionally been the platform for someone to become a public figure. High-quality entity mining, though, is at the moment difficult to achieve, partly because of the high ambiguity which is often associated with person names.

Cross-document coreference resolution (from now on CDCR) is the task of grouping mentions of the same person entities together.[1] Person names are not uniformly ambiguous. Very uncommon names (such as 'Edward Schillebeeckx') are virtually non-ambiguous, whereas very common names (such as 'John Smith') are highly ambiguous. CDCR is closely related to word sense disambiguation, from which it differs greatly in one aspect: contrary to word senses, the set of entities referred to by a person name is *a priori* unknown. The approach we propose assumes a correlation between the commonness of a name and the number of entities referred to by it. Our disambiguation strategy relies on the social circle of the query name. We bring the maxim "you shall know a word by the company it keeps" back to the social realm. Can the social network of a person be an indicator of who that person is? We intend to bring CDCR to the social dimension, with the assumption that the social circle around our target entity can be a source of evidence for disambiguation. Partially-supervised, our approach is competitive with state-of-the-art methods, without relying on a knowledge base (KB) nor other expensive resources. It is easily portable and adaptable to different datasets and different languages without the need of learning new parameters.

2 Formal definition of the task

Given a query name qn and a set of documents in which it appears $\{d_1, d_2, ..., d_j\}$, CDCR aims at grouping together documents containing references to the same entity e. The expected output for each query name is a set of clusters $\{c_1, c_2, ..., c_k\}$, each corresponding to a different

[1] Unlike traditional coreference resolution, CDCR does not usually attempt to resolve definite NPs and pronouns. Following this tradition, we focus only on linking person names.

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 63–73,
Berlin, Germany, August 11, 2016.

entity $\{e_1, e_2, ..., e_k\}$ and each containing the documents referring to it.

For clarity, we describe the terminology used in this paper, which we illustrate with an example:

(1) The character of **John Smith** expresses some of the confusion in **Alexie**'s own upbringing. He was raised in **Wellpinit**, the only town on the **Spokane Indian Reservation**.

A *person name* is any named entity expression in a text referring to a person. An *entity* is the real-world referent that is referred to by a person name. In example 1, 'John Smith' and 'Alexie' are person names, and the real persons behind these names are entities. The *query name* is the target person name to disambiguate, in this case 'John Smith', which is mentioned at least once per document. We assume all mentions of the query name to refer always to the same entity within a document, hence person name clustering amounts to grouping together the documents in which a specific person name refers to a given entity. A *mention name* is any person name that is mentioned in a document, except for the query name, i.e. 'Alexie' in our example. We call a *full name* any person name with at least two tokens (first name and last name), whereas a *namepart* is each of the tokens that form a full name. 'John Smith' is the only full name in our example, and 'John' and 'Smith' are its nameparts. Finally, by *non-person mention* we mean any named entity expression that does not refer to a person ('Wellpinit' and 'Spokane Indian Reservation' in our example).

3 Related work

The idea of using social networks to find information from historical texts is not a new one. One of the first and more influential works is Padgett and Ansell (1993), in which its authors use networks of marriages between the most eminent Florentine families in the 1430s to illustrate the dramatic political changes in the Florence of the time. There exist several recent studies advocating for the use of social networks in historical research (see Jackson (2014), Rochat et al. (2014), i. a.). Most studies relying on social networks concern pre-modern history, where sources are much more limited in number and thus the networks are created either manually or from structured data, thus avoiding one of the greatest challenges in network creation, namely person name disambiguation. One of the few fully automatic approaches is Coll Ardanuy et al. (2015), which does not so much focus on the problem of person name disambiguation, however.

Resolving and disambiguating person names across documents is an open problem in natural language processing, its difficulty stemming from the high ambiguity which is often associated with person names.[2] Sentences 2, 3, and 4 provide three examples of cases in which the same name (in this case 'John Smith') refers to three different persons: the CEO of General Motors, the Labour Party leader, and a coach.

(2) UAW President Stephen Yokich then met separately for at least an hour with chief executives Robert Eaton of Chrysler Corp., Alex Trotman of Ford Motor Co. and finally with John Smith Jr. of General Motors Corp.

(3) Blair became Labour leader after the sudden death of his successor John Smith in 1994 and since then has steadily purged the party of its high-spend and high-tax policies and its commitment to national ownership of industrial assets.

(4) Two years ago, Powell switched coaches from Randy Huntington to John Smith, who is renowned for his work with sprinters from 100 to 400 meters.

These examples are drawn from *The John Smith Corpus*, the first reference set for CDCR, which was introduced by Bagga and Baldwin (1998). The authors also proposed a new scoring algorithm, B-Cubed, in order to evaluate the task, which was modeled as a document clustering problem. To solve the problem, the authors applied the standard vector space model based on context similarity. Several subsequent studies adapted and extended the approach (Ravin and Kazi (1999), Gooi and Allan (2004)). More recent methods apply LDA and other topic models (Song et al. (2007), Kozareva and Ravi (2011)).

Yoshida et al. (2010) distinguish between weak and strong features. Weak features are the context words of the document, as opposed to strong features such as named entities, biographical information, key phrases, or temporal expres-

[2] According to the U.S. Census Bureau, only 90,000 different names are shared by up to 100 million people (Artiles et al., 2009a).

sions (see Mann and Yarowsky (2003), Niu et al. (2004), Al-Kamha and Embley (2004), Bollegala et al. (2006)). The most exploited source of evidence for clustering is named entities (Blume (2005), Chen and Martin (2007), Popescu and Magnini (2007), Kalashnikov et al. (2007)). Artiles et al. (2009a) thoroughly study the role of named entities in the task and conclude that they often increase precision at the expense of recall, even though they leave the door open to more sophisticated approaches using named entities, such as in combination with other levels of features (Yoshida et al., 2010) or in graph-based approaches (Kalashnikov et al. (2008), Jiang et al. (2009), Chen et al. (2012)). Over the last years, the trend has moved towards using resource-based approaches, such as a knowledge base (KB) (Dutta and Weikum, 2015) or Wikipedia, and the person name disambiguation task has been in most cases subsumed by entity linking. Bunescu and Pasca (2006), Cucerzan (2007) and Han and Zhao (2009) are only some of the many approaches that exploit the wide coverage of Wikipedia by linking entity mentions to the referring Wikipedia articles.

An evaluation campaign was organized in 2007 to tackle the problem of name ambiguity on the WWW and the interest of this task moved largely to the web domain (Artiles et al., 2007). However, web pages and news articles differ greatly in their form. Even though more heterogeneous, web pages tend to be more structured and provide additional features that can be exploited (url, e-mail addresses, phone numbers, etc.). In 2011 a similar evaluation campaign was proposed at EVALITA 2011 in order to evaluate CDCR in Italian in the news domain (Bentivogli et al., 2013).

Pairwise clustering has been the most popular clustering method: two documents are grouped together if their similarity is higher than a certain threshold. To date, most approaches have used a fixed similarity threshold. Very few approaches (Popescu (2009), Bentivogli et al. (2013)) have warned of the importance of determining the ambiguity degree of a person name in order to be able to estimate the number of output clusters. In Zanoli et al. (2013), a dynamic threshold similarity is introduced by estimating the ambiguity of the query name. This work, which in this aspect is the most similar to ours, differs greatly from ours with respect to the clustering strategy, since they rely on a KB, whereas we exploit only the context.

Our method aims at providing a solution for the problem of person name disambiguation in the task of automatically constructing social networks from historical newspapers. The articles that constitute our corpus are likely to be populated by many people that are absent from historical accounts and, therefore, also from KBs. We intentionally refrain from linking entities to a knowledge base to avoid the bias towards entities which are present in it. Ter Braake and Fokkens (2015) discuss the problem of biases in historiography and the importance of rescuing long-neglected individuals from the oblivion of history.

4 The model

Given the assumption that a person name always refers to the same entity in a given document,[3] person name clustering amounts to document clustering. In order to cluster documents, a similarity measure is needed. The core idea is that two documents should be clustered together if they are similar enough, i.e. if there exists enough evidence that they belong together. The evidence needed, though, may vary greatly depending on the query name. If the query name is not ambiguous at all, very low similarity between documents suffices to group them into one cluster. Conversely, if the query name is very ambiguous, a higher similarity is required to ensure that only documents that refer to the same entity are clustered together. In section 4.1, we describe how we assess person name ambiguity. Our model relies heavily on the social dimension of news, so we model document similarity based on social network similarity. Thus, for each query name we represent documents as social networks in which the nodes are the people mentioned in them. To determine network similarity (see section 4.2.1), we take two types of information into account: the amount of node overlap (for which we learn a threshold from a small manually labeled data set) and the ambiguity of the overlapping nodes (for which we manually set a penalty function). Network overlap is not always a sufficient source of information (in particular, small overlap does not mean that the documents involved should not be clustered together), and we additionally make use of further features in those cases where networks do not provide sufficient ev-

[3] This is an assumption made by previous approaches and reminiscent of the ‘one sense per discourse’ assumption in word sense disambiguation.

idence: BoW representations of the content, the dominant topic according to a topic modeling algorithm, and the overlap in other named entity expressions (see 4.2.2). These additional features model document content.

4.1 Assessing name ambiguity

Person names are usually combinations of a first name, a last name, and occasionally one or more middle names. Only with the list of all the people in the world would it be possible to assess the true ambiguity of each person name. Since this is an unavailable resource, alternative ways of approximating person name ambiguity need to be found.

4.1.1 Building the resource

Zanoli et al. (2013) use an Italian specific resource, the phonebook *Pagine Bianche*. It has wide coverage, but it could be argued that its use leads to a gender-biased calculation of name ambiguity, since only one person per household is included in its pages, usually its male head. We extract person names from a large corpus of text using a named entity recognizer. To optimize precision, we consider only names consisting of at least two tokens, since single tokens are often misidentified or misclassified by the recognizer. The identified person names are then used to build three lists — one for first names, one for last names, and one for middle names — in which each distinct name is associated with its occurrence frequency in the corpus.

4.1.2 Name ambiguity calculation

We propose an ambiguity scale that spans from 0 to 1, in which very ambiguous names would occupy the highest range and very non-ambiguous names would take the lowest range. Formally, we distinguish three types of names that we can encounter in texts: **(1) Single-token names** are the most ambiguous. In order to calculate the ambiguity of a given single-token name, we merge the first, middle, and last names lists into one and estimate the relative frequency of the target name in the resulting list. We place them within the range 0.8 (the rarest) to 1.0 (the most common). **(2) Two-token names** (usually first and last name) are the most common combination to be expected. Thus, they occupy the central and largest part of the spectrum, the range between 0.2 and 0.8; the most ambiguous name being 0.8, the least ambiguous starting from 0.2. We calculate the weighted average of the two nameparts according to our observation that first names are 15 times more ambiguous than last names. The frequency of the most common two-token name (‘Giovanni Rossi’ for Italian, ‘John Smith’ for English) is taken as the maximum value against which we calculate the ambiguity value of any other two-token name. **(3) Multiple-token names** consist of three parts or more (usually first name, middle name(s), and last name) and are given the lowest ambiguity range, from 0.0 to 0.2. The most common multiple-token combination will have an ambiguity of 0.2, while the ambiguity of the least common name will start from 0.0. Multiple-token names are weighted in the same fashion as two-part names, distributing the weight of the first and the middle names equally.

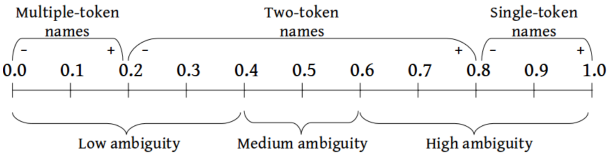

Figure 1: Person name ambiguity range.

We distinguish three degrees of ambiguity. **Low-ambiguity** consists of the multiple-token names and the least ambiguous two-token names. **High-ambiguity** consists of the single-token names and the most ambiguous two-token names. **Middle-ambiguity** contains the names that fall into the middle spectrum (see Figure 1). Table 1 shows examples of English names that fall into each range.

AmbR	Examples
0.0-0.1	Lena Mary Atkinson, Edward William Elgar
0.1-0.2	Mary Anne Smith, John Douglas Williams
0.2-0.3	Douglas Morris, Anne Atkinson
0.3-0.4	Donald Taylor, Emma White
0.4-0.5	Mary Johnson, George Williams
0.5-0.6	Thomas Jones, James Williams
0.6-0.7	John Williams, Mary Smith
0.7-0.8	John Smith, William Smith
0.8-0.9	Atkinson, Terrence
0.9-1.0	John, William

Table 1: On the left, the ambiguity range; on the right, some examples of names from each range.

4.2 Clustering scheme

Fixed similarity thresholds have been the most used for this task: two documents are clustered together if their similarity surpasses a predefined threshold. Such algorithms do not take the ambiguity of the query name into account. Ideally, ambiguous names should have high thresholds, allowing fewer documents to be clustered together, whereas non-ambiguous names should have low thresholds, therefore yielding less clusters.

Our method's first step is to represent each document containing the query name as a social network of the people mentioned in it. To identify the names we used a named entity recognizer. We perform naive within-document coreference resolution of mention names based on their matching surface forms and construct undirected social networks weighted by the coocurrence of entities within a text window. We initiate our clustering algorithm by taking the social network with the highest number of nodes, and sort the remaining networks by decreasing number of nodes overlapping with the largest network. If the similarity between both social networks is bigger than a certain threshold (this is discussed more thoroughly in the next section), we cluster both documents together and merge the two social networks, re-rank the list of overlapping networks and take again the top one from the list. We repeat this process until no partially-overlapped network is found. In this case, we repeat the whole procedure of finding the largest remaining network and finding its fully- or partially-overlapping networks. We continue until all the networks/documents have been considered. This is a greedy algorithm, and it is thus of prime importance that two documents are only clustered together if there exists enough evidence that this should be the case.

Each query name is assigned an ambiguity range, which falls into one of the three ambiguity degrees: low, medium or high. The clustering strategy varies according to the range and degree of ambiguity of each query name, so that non-ambiguous names allow low-similarity documents to be clustered together, whereas ambiguous names require high document similarity.

4.2.1 Social network similarity

The core idea behind our approach is that the social circle of people tells us who they are: it is their social context. A very naive version of our approach would consist in joining together under the same entity all documents with at least one shared person name (apart from the query name). This is obviously dangerous, as using this method in a large enough dataset would eventually cluster all documents together. In order to understand how reliable it is to cluster networks together when sharing a certain number of nodes, we decided to learn clustering probabilities from a development set. For each ambiguity range, we learn the probabilities of two documents being clustered together when they have one, two or three nodes in common. A pair of networks with no overlapping nodes gives us no information about the social context. We observed in the development set that, with more than four overlapping nodes, two documents are unequivocally clustered together.

Node overlapping quality. We have so far talked about overlapping nodes as a synonymous expression for overlapping entities, assuming that a mention name that appears in two documents refers to the same entity. This is of course not necessarily the case. Mention names can range from single tokens to multiple tokens, and correspond to names that can be both very ambiguous (such as 'John') or very unambiguous (such as 'Edward Cornelis Florentius Alfonsus Schillebeeckx'). The confidence that we are talking about the very same person varies greatly from the first case to the second case. The likelihood that two documents belong to the same cluster given a certain overlap of person names will therefore depend on the 'quality' of these overlaps. An overlapping name that provides greater evidence that we are dealing with one only entity (i.e. a low-ambiguity name) is considered of higher quality than an overlapping name that provides little evidence that it corresponds to one only entity (i.e. a high-ambiguity name).

Node ambiguity penalty. We compute the ambiguity of each mention name and assign it an ambiguity degree: high, medium, or low. A penalty function is defined to lower the learned probabilities when applied to networks with low-quality overlapping nodes:

$$penalty = \frac{Pr(n[i]) - Pr(n[i-1])}{i+1} \qquad (1)$$

where i is the number of overlapping nodes between two documents, n the set of networks sharing a certain number i of nodes, and thus $Pr(n[i])$ the probability that two networks belong together if they have i nodes in common. Table 2 shows how probabilities are recalculated.

4.2.2 Other similarity metrics

Even though the skeleton architecture of our clustering scheme is based on the social circle of peo-

ONE OVERLAPPING NODE	
$penalty = \frac{Pr(n[1])-Pr(n[0])}{2}$	
Amb	**Probability recalculated**
↑	$Pr(n[1]) - 2 \cdot penalty = Pr(n[0])$
→	$Pr(n[1]) - penalty$
↓	$Pr(n[1]) - 0 \cdot penalty = Pr(n[1])$

TWO OVERLAPPING NODES	
$penalty = \frac{Pr(n[2])-Pr(n[1])}{3}$	
Amb	**Probability recalculated**
↑↑	$Pr(n[2]) - 4 \cdot penalty$
↑→	$Pr(n[2]) - 3 \cdot penalty = Pr(n[1])$
→→	$Pr(n[2]) - 2 \cdot penalty$
↑↓	$Pr(n[2]) - 2 \cdot penalty$
→↓	$Pr(n[2]) - 1 \cdot penalty$
↓↓	$Pr(n[2]) - 0 \cdot penalty = Pr(n[2])$

THREE OVERLAPPING NODES	
$penalty = \frac{Pr(n[3])-Pr(n[2])}{4}$	
Amb	**Probability recalculated**
↑↑↑	$Pr(n[3]) - 6 \cdot penalty$
↑↑→	$Pr(n[3]) - 5 \cdot penalty$
↑↑↓	$Pr(n[3]) - 4 \cdot penalty = Pr(n[2])$
↑→→	$Pr(n[3]) - 4 \cdot penalty = Pr(n[2])$
↑→↓	$Pr(n[3]) - 3 \cdot penalty$
→→→	$Pr(n[3]) - 3 \cdot penalty$
↓→→	$Pr(n[3]) - 2 \cdot penalty$
↓↓↑	$Pr(n[3]) - 2 \cdot penalty$
↓↓→	$Pr(n[3]) - 1 \cdot penalty$
↓↓↓	$Pr(n[3]) - 0 \cdot penalty = Pr(n[3])$

Table 2: Recalculation of probabilities. The left column shows the combination of nodes according to their ambiguity degree. Each arrow represents one node: ↑ a high-ambiguity name, → a medium-ambiguity name, and ↓ a low-ambiguity name. In the right column, the probability of two networks being clustered together based on the number of nodes they share is lowered according to the quality of their overlapping nodes.

ple, the evidence social network similarity provides is limited. As discussed in Artiles et al. (2009a), approaches that focus on named entities achieve high precision at the cost of recall. Our method is especially vulnerable when two networks share zero or one overlapping nodes, since the evidence that the two networks should be clustered together is in these cases non-existent or very small. In order to address this problem, each social network stores the set of named entity expressions that were not used for the network creation (e.g. locations and organizations) and three bag-of-words representations of the document: with tf-idf weightings, with simple counts, and with non-person mentions. For each ambiguity range and for each feature, we learn the probabilities that two networks sharing one or no overlapping nodes still belong together. Finally, we applied LDA using collapsed Gibbs sampling to our datasets to produce a lower dimensional representation of our dataset, and assign the most relevant latent topic to each network.

4.3 Clustering decisions

We have so far discussed the general clustering architecture, but not how the actual decision of whether to group a pair of documents together is made. We base this decision on a set of seven features which can be extracted for each document pair: **(1)** number of person overlaps; **(2)** number of non-person mention overlaps; **(3)** probability that, given an ambiguity range (that of the query name), two networks are clustered together if they share one, two, or three nodes; **(4)** probability that, given an ambiguity range, two documents are clustered together in terms of a BoW vector representation of word counts; **(5)** probability that, given an ambiguity range, two documents are clustered together in terms of a BoW vector representation with tf-idf weightings; **(6)** probability that, given an ambiguity range, two documents are clustered together if they have a certain number of non-person mentions in common; **(7)** and the most relevant topic for the document.

Since a less ambiguous name tends to correspond to fewer entities than a more ambiguous one, the clustering decision threshold for a low-ambiguity query name should be more permeable than the threshold for an ambiguous name. Each query name is assigned an ambiguity value that corresponds to one of three ambiguity degrees: low, medium, or high. Since a **low-ambiguity** query name is likely to refer to very few entities, if any of the extracted features is true, we consider this evidence enough to cluster the two documents together. On the other side of the spectrum, **high-ambiguity** names are likely to correspond to several entities, so the amount of evidence needed in order to cluster documents is bigger. We assume that an overlap of five entities (be them person names, locations, or organizations) should be enough evidence that we are talking about the same person. The smaller the named entity overlap is, the more evidence will be required and thus the more features will have to be true. **Medium-ambiguity** names will have a middle stance between low-ambiguity and high-ambiguity names when it comes to permeability.

5 Experiments

5.1 Data

To our knowledge, no datasets are available for assessing the value of our method in historical newspaper texts. Therefore, we evaluate our model on three existing datasets from the contemporary press (with articles starting from the year 1987). The **Cross-document Italian People Coreference corpus (CRIPCO)** (Bentivogli et al., 2008) comes with a development and test set, in Italian, of 105 and 103 query names respectively, with an average of 3.45 entities per query name and a total of 20,754 documents. The **NYTAC Pseudo-name Corpus** is an artificial corpus created by conflating dissimilar person names together. With a total of 19,360 documents, this dataset consists of 100 pairs of conflated person names (i.e. 200 entities), matching in gender and 50 of which being topically similar, such as Robert Redford and Clint Eastwood (actors) or Plácido Domingo and Luciano Pavarotti (opera singers). Finally, the **John Smith Corpus** consists of only one query name, 'John Smith', the most common name of the English language. It consists of 197 documents containing at least one instance of 'John Smith', representing 35 entities. The documents are not equally distributed among the different entities: 24 entities appear mentioned only in one document, whereas one entity is mentioned in 88 documents.

In addition to the quantitative evaluation on contemporary data, we also provide a qualitative evaluation on historical data in section 6.

5.2 Baselines

We compare our method **SNcomplete** with two baseline methods: (1) **SNsimple** is the base case, the most naive representation of our method, in which two documents are grouped together if their network representations share at least one node; and (2) **TopicModel** clusters together the documents that share the most relevant topic. We also provide the state-of-the-art results for the *CRIPCO* dataset (Zanoli et al., 2013) and for the *NYTAC pseudo-name corpus* (Rao et al., 2010), who also presented results on the *John Smith Corpus*.

5.3 Settings

We use the Stanford NER (Finkel et al., 2005) and TextPro (Pianta et al., 2008) to identify NEs in English[4] and Italian[5], respectively. We make use of an unannotated Italian corpus, PAISÀ,[6] consisting of 1.5GB of raw text at the moment of download (March 2015), from which we extract person name lists to compute ambiguity ranges for Italian. The extracted list of 718,568 person names is not a census of the Italian population, but a list of people mentioned in news, webpages or blogs. For the English experiment, we used the Persondata information from the DBPedia[7] project (only available for English and German at the moment), which was built by collecting all the Wikipedia articles about people. The Persondata database had 7,889,574 entries at the moment of download (December 2014).

Our method does not require a big amount of training data, but just a representative selection spreading over the ambiguity range is enough to set the appropriate parameters. The CRIPCO corpus provides a development set of documents corresponding to 103 different query names, but a small fraction of it (15 query names, about 15% of the set) is already sufficient to set the appropriate parameters (using the whole dataset makes no significant difference in the performance). We randomly selected the query names, making sure we would, when possible, have a query name for each of the ten ambiguity ranges.[8] Our training dataset does not have a query name for all of the ambiguity ranges: we lack training examples from the range 0.1-0.2, as well as for the three upper ranges (0.7-0.8, 0.8-0.9, and 0.9-1.0). In our experiment, if a query name from the testing dataset falls into one of these ranges, it would take the probability of its immediately precedent ambiguous range. The mentioned fifteen instances from the development set have also been used to find the optimal combination of features. The learned probabilities and feature combination strategy have been applied directly, without further learning nor tuning, to the other two datasets.

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

[5] http://textpro.fbk.eu/

[6] http://www.corpusitaliano.it/

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

[8] The fifteen training instances for each range are: 'Isabella Bossi Fedrigotti' *(0.0-0.1)*; 'Marta Sala', 'Alberto Sighele', 'Roberto Baggio', 'Bruno Degasperi', 'Ombretta Colli', and 'Leonardo da Vinci' *(0.2-0.3)*; 'Luisa Costa', 'Mario Monti', and 'Andrea Barbieri' *(0.3-0.4)*; 'Antonio Conte', 'Antonio de Luca', and 'Antonio Russo' *(0.4-0.5)*; 'Paolo Rossi' *(0.5-0.6)*; and 'Giuseppe Rossi' *(0.6-0.7)*.

	cripco			*nytac_sel*			*johnsmith*		
Approach	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SNsimple	0.94	0.67	0.78	0.65	0.74	0.67	0.65	0.6	0.62
TopicModel	0.91	0.44	0.55	0.76	0.27	0.37	0.71	0.51	0.59
Zanoli et al. 2013	0.89	0.97	**0.93**	–	–	–	–	–	–
Rao et al. 2010 *[1]*	–	–	–	0.61	0.78	**0.68**	0.60	0.63	0.61
Rao et al. 2010 *[2]*	–	–	–	0.82	0.24	0.37	0.85	0.59	**0.70**
SNcomplete	0.87	0.95	**0.91**	0.63	0.75	**0.68**	0.79	0.60	0.68

Table 3: Evaluation results.

5.4 Results and discussion

Table 3 shows the results of applying our model to the three datasets. We use the evaluation metrics provided for the WePS task (Artiles et al., 2009b). While our results are slightly lower than Zanoli et al. (2013), this difference is not statistically significant (Wilcoxon test, p=0.054). An advantage of our method is that it can easily be adapted to any other dataset without requiring expensive resources, such as a knowledge base.

The only work we are aware of that has reported results for the *NYTAC pseudo-name corpus* is by the creators of the dataset (Rao et al., 2010), who also report results for the *John Smith Corpus*. The NYTAC dataset was artificially created, and some of our assumptions do not hold: in this dataset, ambiguity of the query name does not play a role because there are invariably two clusters for each query name, one for each conflated name. Besides, half the entity pairs of the dataset are very closely related (e.g. Luciano Pavarotti and Plácido Domingo, two names that very often appear mentioned in the same text). Therefore, their social networks have much less predictive power than in natural data, where we assume that two people with the exact same name have low probability to share a big portion of their social networks. That would explain why we report low precision for this dataset, and yet the results obtained are comparable to those from the best of the two models introduced by Rao et al. (2010).

The result reported for *John Smith Corpus* improves upon recent models, such as Singh et al. (2011), who obtained 0.664, but is far from the most recent approach (Rahimian et al., 2014), who obtained around 0.80. This might be well due to the fact that there was only one query name in our development set that had high ambiguity, which was, still, far from being as ambiguous as 'John Smith'. Our method works overall better than any of the two methods from Rao et al. (2010) when we average the results for both English datasets.

Using the ambiguity of the query name to dynamically decide on a clustering strategy is crucial for the success of our method. Failing to choose an adequate ambiguity range for query names can lead to considerably lower results. Our F-Score for the *John Smith Corpus* drops to 0.37 if we consider 'John Smith' a low-ambiguity name, and to 0.52 if we consider it of medium-ambiguity. The F-Score for the *CRIPCO* dataset drops to 0.77 when the ambiguity range of the query names of this dataset is randomly assigned.

6 Impact in the social sciences: a case study on Dutch religious history

To assess the impact of this approach in the social sciences, we introduce here a case study that analyzes its performance and proves its contribution. Due to lack of annotated data from the historical news domain, we can only offer a qualitative analysis. As a use case, we focus on two actors who played a pivotal role in the religious transformations of the postwar years in the Netherlands: Willem Banning and Edward Schillebeeckx. The first was a leading intellectual in the movement responsible for a major transformation within the Reformed Church; the latter was a prominent member of an international network of progressive theologians who deeply influenced discourse on the future of the Catholic Church.

Our data consist of all the articles from the newspaper collection of the Dutch National Library containing the query words 'Banning' and 'Schillebeeckx'. In order to remove obvious outliers, we applied some heuristics to disregard those articles in which the query name was preceeded by any capitalized word not coinciding with their first and middle names, their initials, or with any title. We restricted the data to the years in which we are interested, namely between 1930 and 1970 in the case of Banning, and 1950 and 1990 in the case of Schillebeeckx. We ended up with 26,984 documents for Banning (137 MB) and 2,796 documents for Schillebeeckx (8.5 MB). The name ‘Banning’ is much more common in Dutch than the name ‘Schillebeeckx’, which is probably the reason behind the large difference in the number of articles between the two. Whereas all mentions of ‘Schillebeeckx’ in the collection seem to refer to the person in which we were interested, a quick search at the beginning of the experiment revealed that there were several different persons with the name ‘Banning’ in the collection, among which at least a shopkeeper, a swimming champion, a man on trial, and an amateur fisherman.

Our method returns one network for each disambiguated entity. Figure 2 shows an example of social network created with our method.[9] As mentioned, each edge is a container of information (context words and non-person mentions weighted with tf-idf) that can be found in the articles where the two nodes connected by the edge are present. This information is encoded for each pair of nodes that can be found in the network. Each edge also stores the list of documents in which both nodes appear, in order to grant access to the original sources to the historian.

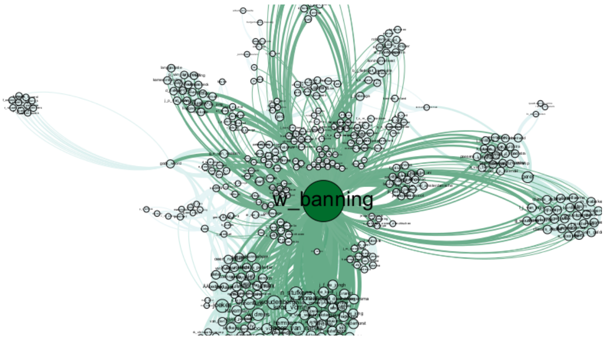

Figure 2: Fragment of the resulting social network for Willem Banning for the year 1963.

The amount of noise that can be found in the networks created from historic newspapers is clearly higher than in the standard benchmarks, mostly due to OCR. As a result, the named entity recognizer, trained on modern Dutch,[10] performs worse, but the final networks do not suffer much from this, since noisy nodes are pushed to the periphery of the networks. The historian in our team was able to find only expected names in the center of the networks, with very few exceptions. By thoroughly looking at the connections between the nodes of the networks and the context information stored in the edges, several points and episodes of the lives of the two politicians could be confirmed: the importance of Schillebeeckx as an advisor of the Dutch episcopacy and his triple heavy scrutiny by the Vatican, and a higher number of international relations than in the case of Banning. Expected information in the networks is interesting because it proves the validity of the approach. Even more interesting is the presence of unexpected results in the network, since they can lead to potential hypotheses that may challenge the dominant narratives of history. Our networks suggest, contrary to what is believed, that Schillebeeckx was a popular theologian not only because of his conflict with Rome, but also because of his theological ideas, and that Banning’s work in politics was not separated from his ideas on the role of the church in society. Given these promising findings, we intend to pursue research in this direction.

The network approach provides historians with a quick but thorough overview of the role of someone in the public eye: with whom was he or she connected, which topics were central and in which debates he or she participated. By navigating through the networks, one can explore the collection at ease, validating well-known historical reports, developing new ideas, and even rediscovering new actors who may have had a bigger role in the past than that which History granted them, always from the perspective of a certain newspaper collection. It is then the task of the historian to verify, by looking at the pieces of news selected by our method, whether there is some truth in the information yielded by the network.

7 Conclusions

We have presented a new method for constructing social networks of disambiguated person entities from news articles. Our method explores the relationship between name ambiguity and the amount of different entities that can be referred to by the same name. Our approach is partially supervised and has proved to be competitive in different languages and throughout very different collections without need to retrain it. The method outputs a set of social networks, one for each distinct entity, which can be of great assistance in the exploration of historical collections.

[9] We used Gephi (`https://gephi.org/`) for visualizing it, the size and position of the nodes depend on the weights of their edges.

[10] We use the training data from CoNLL-2002: `http://www.cnts.ua.ac.be/conll2002/ner/`

References

Reema Al-Kamha and David W. Embley. 2004. Grouping search-engine returned citations for person-name queries. In *Proceedings of the 6th ACM WIDM Workshop*, pages 96–103.

Mariona Coll Ardanuy, Maarten van den Bos, and Caroline Sporleder. 2015. Laboratories of community: How digital humanities can further new European integration history. In *Histoinformatics 2014*, pages 284–293, Barcelona.

Javier Artiles, Julio Gonzalo, and Felisa Verdejo. 2005. A testbed for people searching strategies in the WWW. In *Proceedings of SIGIR*, pages 569–570.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In *Proceedings of SemEval*, pages 64–49.

Javier Artiles, Enrique Amigó, and Julio Gonzalo. 2009a. The role of named entities in Web People Search. In *Proceedings of EMNLP'09*, pages 534–542.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2009b. WePS-2 evaluation campaign: Overview of the web people search clustering task. In *Proceedings of the 2nd WePS Workshop*.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of Coling*, pages 79–85.

Luisa Bentivogli, Alessandro Marchetti, and Emanuele Pianta. 2008. Creating a gold standard for person cross-document coreference resolution in Italian news. In *Proceedings of LREC Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, pages 19–26.

Luisa Bentivogli, Alessandro Marchetti, and Emanuele Pianta. 2013. The news people search task at EVALITA 2011: Evaluating cross-document coreference resolution of named person entities in Italian news. In *Proceedings of EVALITA 2012*, pages 126–134.

Matthias Blume. 2005. Automatic entity disambiguation: benefits to NER, relation extraction, link analysis, and inference. In *Proceedings of the Intl Conference on Intelligence Analysis*.

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2006. Extracting key phrases to disambiguate personal name queries in web search. In *Proceedings of the ACL Workshop on How Can Computational Linguistics Improve Information Retrieval?*

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, pages 9–16.

Ying Chen and James Martin. 2007. Towards robust unsupervised personal name disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 190–198.

Liwei Chen, Yansong Feng, Lei Zhou, and Dongyan Zhao. 2012. Explore person specific evidence in web person name disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 832–842.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL*, pages 708–716.

Sourav Dutta and Gerhard Weikum. 2015. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 3, pages 15–28.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of ACL*, pages 363–370.

Chung Heong Gooi and James Allan. 2004. Cross-document coreference on a large scale corpus. In *Proceedings of HLT-NAACL*, pages 9–16.

Xianpei Han and Jun Zhao. 2009. Named entity disambiguation by leveraging Wikipedia semantic knowledge. In *Proceedings of CIKM*, pages 215–224.

Cornell Jackson. 2014. Using social network analysis to reveal unseen relationships in Medieval Scotland. In *Digital Humanities Conference*, Lausanne.

Lili Jiang, Jianyong Wang, Ning An, Shengyuan Wang, Jian Zhan, and Lian Li. 2009. Grape: A graph-based framework for disambiguating people appearances in web search. In *Proceedings of IEEE International Conference on Data Mining*, pages 199–208.

Dmitri V. Kalashnikov, Stella Chen, Rabia Nuray, Sharad Mehrotra, and Naveen Ashish. 2007. Disambiguation algorithm for people search on the web. In *Proceedings of IEEE International Conference on Data Engineering*, pages 1258–1260.

Dmitri V. Kalashnikov, Rabia Nuray-Turan, and Sharad Mehrotra. 2008. Towards breaking the quality curse. a web-querying approach to web people search. In *Proceedings of SIGIR*, pages 27–34.

Zornitsa Kozareva and Sujith Ravi. 2011. Unsupervised name ambiguity resolution using a generative model. In *Proceedings of EMNLP*, pages 105–112.

Gideon S. Mann and David Yarowsky. 2003. Unsupervised personal name disambiguation. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, pages 33–40.

Cheng Niu, Wei Li, and Rohini K. Srihari. 2004. Weakly supervised learning for cross-document person name disambiguation supported by information extraction. In *Proceedings of ACL*, pages 598–605.

John F. Padgett and Christopher K. Ansell. 1993. Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology*, pages 1259–1319.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro tool suite. In *Proceedings of LREC*, pages 2603–2607.

Octavian Popescu and Bernardo Magnini. 2007. IRST-BP: Web people search using name entities. In *Proceedings of SemEval*, pages 195–198.

Octavian Popescu. 2009. Person cross document coreference with name perplexity estimates. In *Proceedings of EMNLP*, pages 997–1006.

Fatemeh Rahimian, Sarunas Girdzijauskas, and Seif Haridi. 2014. Parallel community detection for cross-document coreference. In *Proceedings of IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pages 46–53.

Delip Rao, Paul McNamee, and Mark Dredze. 2010. Streaming cross document entity coreference resolution. In *Proceedings of Coling*, pages 1050–1058.

Yael Ravin and Zunaid Kazi. 1999. Is Hillary Rodham Clinton the President? Disambiguating names across documents. In *Processinds of the Workshop on Coreference and its Applications*, pages 9–16.

Yannick Rochat, Melanie Fournier, Andrea Mazzei, and Frédéric Kaplan. 2014. A network analysis approach of the Venetian Incanto system. In *Digital Humanities Conference*, Lausanne.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of ACL-HLT*, pages 793–803.

Yang Song, Jian Huang, Isaac G. Councill, Jia Li, and C. Lee Giles. 2007. Efficient topic-based unsupervised name disambiguation. In *Proceedings of Joint Conference on Digital Libraries*, pages 342–351.

Serge ter Braake and Antske Fokkens. 2015. How to make it in History. working towards a methodology of canon research with digital methods. In *Biographical Data in a Digital World*, pages 85–93.

Minoru Yoshida, Masaki Ikeda, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. 2010. Person name disambiguation by bootstrapping. In *Proceedings of SIGIR*, pages 10–17.

Roberto Zanoli, Francesco Corcoglioniti, and Christian Girardi. 2013. Exploiting background knowledge for clustering person names. In *Proceedings of EVALITA'2012*, pages 135–145.